VIVA Institute of Technology
9thNational Conference on Role of Engineers in Nation Building – 2021 (NCRENB-2021)

VIVA-TECH IJRI
ANNUAL RESEARCH JOURNAL

VIVA-TECH INTERNATIONAL JOURNAL FOR RESEARCH AND INNOVATION
ANNUAL RESEARCH JOURNAL
ISSN(ONLINE): 2581-7280

Data Mining Techniques in Smart Agriculture

Jash V Oza[1], Prof .Pradnya Mhatre[2]
[1]*Department of Computer Application , University of Mumbai*
Viva School of MCA ,Shirgaon , Virar (East).
Email: jashoza10@gmail.com
[2]*Department of Computer Application , University of Mumbai*
Viva School of MCA ,Shirgaon , Virar (East).
Email: pradnyamhatre@vivamca.org

Abstract :*Agriculture is an important sector in many countries, especially in the rural sector. It introduces a major source of food for people worldwide. However, it faces the great challenge of producing more and better quality while increasing sustainability through proper use of natural resources, reducing environmental degradation and adapting to climate change.Therefore, it is very important to move from traditional farming methods to new modern agriculture. Smart Agriculture is one of the solutions to address the growing demand for essential food products while meeting the needs of sustainability. In Smart Advanced Agriculture, the role of knowledge is growing day by day. Information on weather conditions, soil, diseases, pests, seeds, fertilizers, etc. It contributes significantly to the economic development and sustainability of the sector. Smart and Advanced Management consists of transferring, collecting, analyzing and selecting data.As the value of agricultural data increases exponentially, robust analytical techniques that are able to process and analyze large amounts of data to obtain accurate data and more accurate predictions are essential. Data Mining is expected to play a key role in Smart / Modish Agriculture managing real-time and big data analysis. The purpose of this paper is to review further studies and research on Advance and smart agriculture using the latest Data Mining practice, to solve various agricultural problems and scenerios.*

Keywords -*Data mining, IoT, Precision agriculture , Smart agriculture, WSN.*

I. INTRODUCTION

Agriculture is one of the world's largest jobs. It is a very traditional practice in all productive activities and has undergone many technological changes and transformations over time with the aim of producing better and better.

However, the sector now faces significant challenges. To reduce the negative effects of productive but dynamic agriculture, it is urgent to transform agricultural production processes in a more sustainable way, by allocating resources more efficiently and by using alternative methods of Smart Agriculture. Smart Agriculture is seen as one of the way to achieve these goals.

The Smart and Precision Agriculture programs should play an important role in improving agricultural activities. It is a combination of IoT and information technology. Its purpose is to gather information from a variety of sources so that they can better understand, predict and organize agricultural work.Smart Agriculture is based on the use of different automation technologies, data capture, data transfer, data processing and decision making. This plant is susceptible to several diseases as it grows. Their discovery is the goal of much research. Based on a combination of Data Mining techniques and image processing to overcome the lack of public viewing

Exploitation of standard digital image processing techniques integrated with Data Mining in the agricultural sector to detect, classify and measure plant diseases. The CNN Algorithm is used to diagnose several plant diseases.The paper examined various plant diseases. Depending on the results of testing the proposed method may provide in part specific information for various diseases.

II. FRAMEWORK

2.1 Convolutional NeuralNetwork(CNN)

Convolutional Neural Network (CNN) is a special type of Neural Networks, which has shown exemplary performance in several competitions related to the Computer Vision and Image Processing. Some of the exciting features of the CNN app include image classification and segmentation, object detection, video processing, natural language processing, and speech recognition. CNN's powerful reading ability is largely due to the use of multi-featured extraction categories that can automatically read presentations from the data. The discovery of a large amount of data and advances in hardware technology has accelerated research on CNN, and recently the latest CNN architecture has been reported.

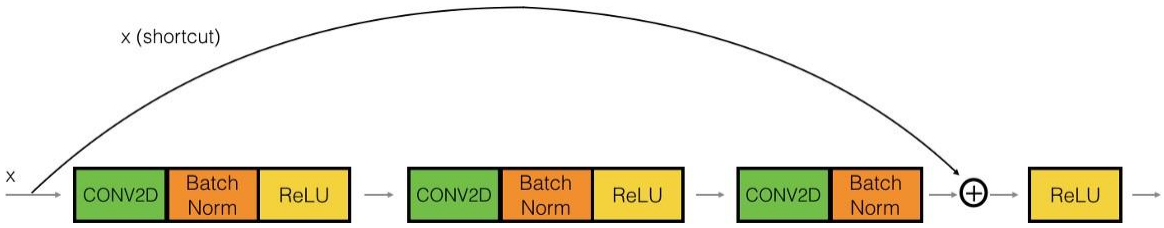


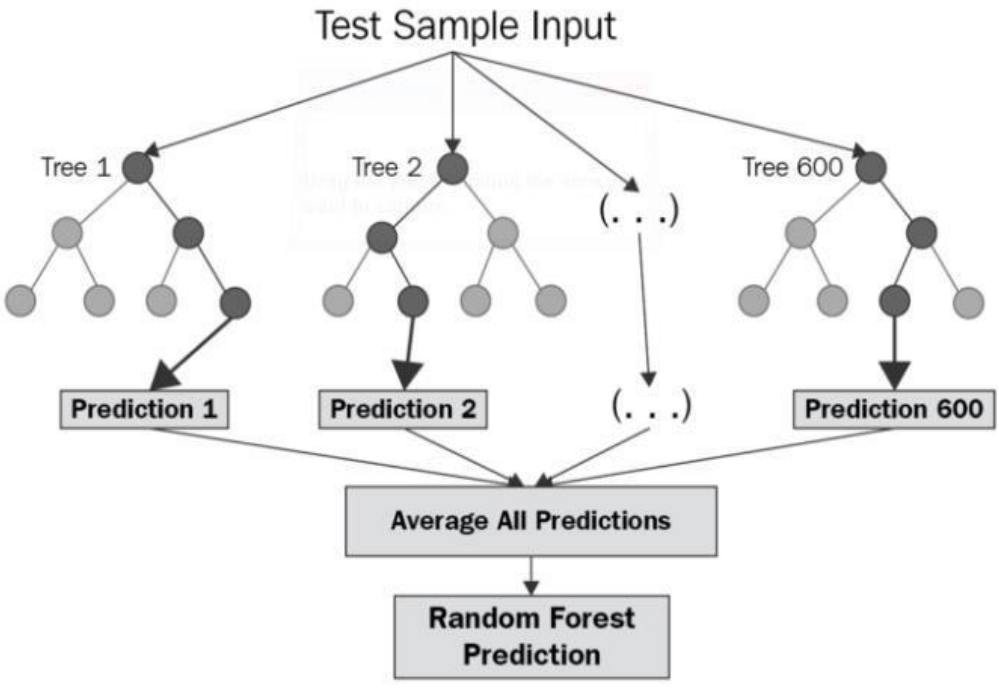
Fig 1 . ResNet9 Architecture

ResNet is a Convolutional Neural Network (CNN) platform designed to provide great power or thousands of conversion layers. ResNet installs a map of ownership, layers that are initially idle, and skips them, and re-creates transactions from previous layers. Skipping initially compresses the network into only a few layers, which makes learning faster. Then when the network trains again, all layers are expanded and the "remaining" parts of the network explore the growing space of the source image more and more.

2.2 Random Forest Regression

The Random Forest is an integrated approach that is capable of performing retrospective and subdivision operations using multiple decision-making trees and a process called Bootstrap and Aggregation, more commonly known as bagging. The basic idea of this is to combine multiple decision trees in determining the end result rather than relying on individual trees decided.

The Random Forest has a lot of decision trees as basic learning models. We randomly create a line sample and import the sample from the database that creates data samples for all models. This section is called Bootstrap



Random Forest Structure

Fig 2. Random Forest Structure

We need to look at the process of deforestation in a random forest like any other machine learning process
• Create a specific question or data and find a source to determine the required data.
• Make sure the data is in an accessible format otherwise convert it to the required format.
• Clarify all visible defects and missing data points that may be required to complete the required data.
• Create a machine learning model
• Set the basic model you want to achieve
• Train machine data reading model.
• Provide model understanding with test details.
• Now compare performance metrics for both test data and predicted data from the model.
• If it does not meet expectations, you can try to improve your model appropriately or fall in love with your data or use another data modeling process.
• You are currently translating the information you have received and reporting accordingly.

III. LITERATURE REVIEW

3.1 Convolutional Neural Network (CNN)

The typical CNN format usually consists of other layers of solution and integration followed by one or more layers that are fully connected at the end. In some cases, a fully integrated layer is replaced by a global water integration layer. In addition to different mapping functions, different control units such as batch normalization and dropout are also included to enhance CNN performance (Bouvrie 2006). CNN structural design plays an important role in building new structures and thus achieving improved performance. This section briefly discusses the role of these factors in the construction of CNN architecture.

3.1.1 Convolutional layer

The convolutional layer is composed of a set of convolutional kernels where each neuron acts as a kernel. However, if the kernel is symmetric, the convolution operation becomes a correlation operation (Ian Goodfellow et al. 2017). Convolutional kernel works by dividing the image into small slices, commonly known as receptive fields. The division of an image in to small blocks helps in extracting the feature motifs. Kernel convolves with the images using a specific set of weights by multiplying its elements with the corresponding elements of the receptive field (Bouvrie 2006). Convolution operation can be expressed as follows:

$$f_l^k(p,q) = \sum_c \sum_{x,y} i_c(x,y).e_l^k(u,v)$$

where, $ic(x, y)$ is an element of the input image tensor Ic , which is element wise multiplied by $ek(u, v)$ index of the kth convolutional kernel kl of the lth layer. Whereas output feature-map of the kth convolutional operation can be expressed

$$\mathbf{F}_l^k = \left[f_l^k(1,1),...,f_l^k(p,q),...,f_l^k(P,Q) \right]$$

Due to weight sharing ability of convolutional operation, different sets of features within an image can be extracted by sliding kernel with the same set of weights on the image and thus makes CNN parameter efficient as compared to the fully connected networks. Convolution operation may further be categorized into different types based on the type size of filters, type of padding, and the direction of convolution (LeCun et al. 2015)

3.1.2 Pooling layer

Feature motifs, which lead to the dissolution of the function of convolution, can occur in various places in the image. Once the features have been removed, its exact location becomes irrelevant as long as its relative relative position is maintained. Pooling or sampling the floor is an interesting local operation. It summarizes the same details in the available field area and elicits a dramatic response from this local region (Lee et al. 2016).

Demonstrates merge function where Zk represents the combined map of the kth layout feature of the

$$\mathbf{Z}_l^k = g_p(\mathbf{F}_l^k)$$

input map, and gp (.) Defines the type of merge function. The use of cohesiveness helps to produce a combination of features, which are not subject to change in translation with minimal disruption (Ranzato et al. 2007; Scherer et al. 2010). Reducing the size of the feature map to the fixed consistency feature not only

controls network complexity but also helps increase overall performance by reducing overload. Various types of assembly shapes such as max, scale, L2, spacing, placement of a local pyramid, etc. Used on CNN (Boureau 2009; Wang et al. 2012; He et al. 2015b).

3.1.3 Activation function

Activation serves as a decision-making function and facilitates the learning of complex patterns. Choosing the right exercise function can speed up the learning process. The feature mapping function used is described in the equation

$$\mathbf{T}_l^k = g_a(\mathbf{F}_l^k)$$

In the above equation, Fk is the result of a convolution, given the task of activating (.) Add a non-linear and return the modified effect Tk of the lth layer. In the literature, various activation functions such as sigmoid, tanh, maxout, SWISH, ReLU, and ReLU variants, such as ReLU leak, ELU, and PReLU are used to stabilize a combination of offline signals (LeCun 2007; Wang et al. 2012; Xu et al. 2015a; Ramachandran et al. 2017; Gu et al. 2018). However, ReLU and its variants are preferred as they help to overcome the problem of gradient disappearance (Hochreiter 1998; Nwankpa et al. 2018). One of the most recently proposed renewal projects is MISH, which has shown better performance than ReLU on most of the newly proposed deep network data on bench press (Misra 2019).

3.1.4 Batch normalization

Batch editing is used to address issues related to internal covariance changes within feature maps. The internal change of covariance is a change in the distribution of hidden units of numbers, which reduces the interaction (forcing the learning rate to a minimum) and requires careful initiation of parameters. The normality of a map of the modified feature shown shown is shown in the equation

$$\mathbf{N}_l^k = \frac{\mathbf{F}_l^k - \mu_B}{\sqrt{\sigma_B^2 + \varepsilon}}$$

In the equation, Nk represents the standard feature map, Fk input map, and $\sigma 2$ indicate the shape and variation of a small group map feature respectively. To avoid zero division, E is added to price stability. Batch editing involves the distribution of embedded map values by placing them at zero mean and unit variations (Ioffe and Szegedy 2015). In addition, it adjusts the gradient flow and becomes a controlling factor, which helps improve network performance.

3.1.5 Dropout

Dropout introduces performance within the network, which ultimately improves performance by periodically skipping certain units or intermittent connections. In NNs, many connections that learn non-linear relationships are sometimes interchangeable, which creates an overdose (Hinton et al. 2012b). Random collapse of connections or units produces several reduced network structures, and ultimately, a single representative network is selected with minimal weights. This selected structure is then considered as the ratio of all proposed networks (Srivastava et al. 2014).

3.1.6 Fully connected layer

Fully connected layer is mostly used at the end of the network for classification. Unlike pooling and convolution, it is a global operation. It takes input from feature extraction stages and globally analyses the output of all the preceding layers (Lin et al. 2013). Consequently, it makes a non-linear combination of selected features, which are used for the classification of data (Rawat and Wang 2016).

3.2 Random ForestRegression

Random forestsregressionareanensemblelearning method for classification,regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individualtrees.

3.2.1Ensemble Learning

An Ensemble method is a technique that combines the predictions from multiple machine learning algorithms together to make more accurate predictions than any individual model. A model comprised of many models is called an Ensemble model.

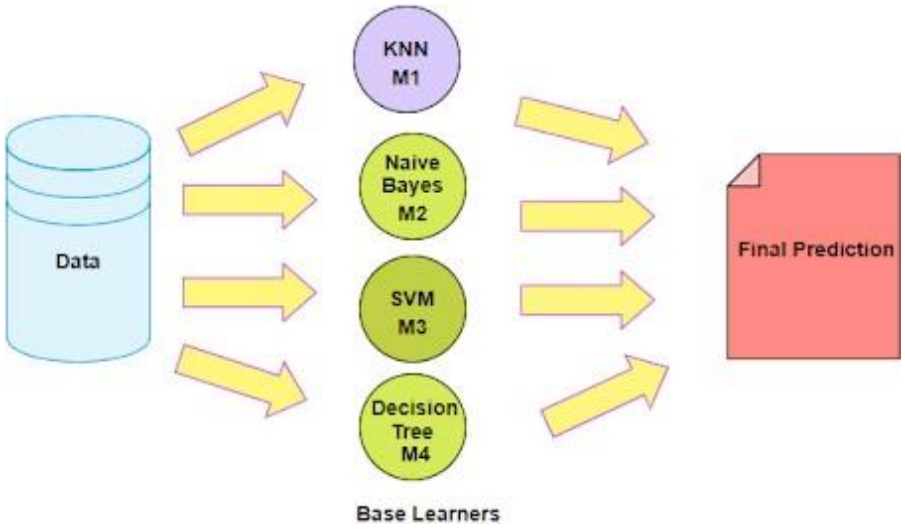


FIG 3. BASE LEARNERS

IV. TYPES OF ENSEMBLE LEARNING

1. Boosting.
2. Bootstrap Aggregation(Bagging).

3.2.1.1 Boosting

Boosting refers to a group of algorithms that use weighted scales to make weak students into stronger students. Promoting everything is about "collaboration". Each running model, dictates what aspects the next model will focus on.

3.2.1.2 In developing as the name suggests, one learns something that enhances learning.Bootstrap Aggregation(Bagging)

Bootstrap refers to a random sample for replacement. Bootstrap allows us to better understand bias and diversity of databases. Bootstrap includes random sampling of a small subset of data from the database.

It is a standard procedure that can be used to reduce the variability of those algorithms with high variability, usually tree resolutions. Bagging makes each model work independently and aggregates the results in the end without choosing any model.

V. CONCLUSION

In this sense, we have been able to build a ResNet9 model using a convolutional neural network that can detect images with 91% accuracy using Pytorch. We have achieved this accuracy by pre-processing the images to make the model more common, sorting the data into multiple collections and finally building and training the model.

We have learned about the different types of ensemble learning algorithms and how these algorithms help make Random Forest work. We found a crop prediction using Random Forest in addition to other machine learning algorithms.

Acknowledgements

I would like to express my deepest appreciation to all those who provided me the possibility to complete this report. I take this opportunity to express my profound gratitude and deep regards to my guide Prof. PradnyaMhatre for her exemplary guidance, monitoring and constant encouragement throughout the course. The blessings, help and guidance given by her time to time shall carry me a long way in the journey of life on which I am about to embark.

Furthermore, I would also like to acknowledge the staff of Post Graduate Department of MCA, who gave their valuable support. I would like to express my gratitude towards my parents for their kind cooperation and encouragement which help me in completion of this report.

REFERENCES

[1] Arivazhagan, S., Shebiah, R.N., Ananthi, S., Varthini, S.V., 2013. Detection of unhealthy region of plant leaves and classification of plant leaf diseases using texture features. Agric. Eng. Int. CIGR J. 15,211–217[1].

[2] Boniecki, P., Koszela[2]., K., Piekarska-Boniecka, H., Weres, J., Zaborowicz, M., Kujawa, S., Majewski, A., Raba, B., 2015. Neural identification of selected apple pests. Comput. Electron. Agric. 110, 9–16[2]..

[3] Brahimi, M., Boukhalfa[3]., K., Moussaoui, A., 2017. Deep learning for tomato diseases: classification and symptoms visualization. Appl. Artif. Intell. 31, 299–315

[4] Breiman, L., Friedman, J., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. CRC[4]..

[5]Aggarwal, C.C., 2014. Data Classification: Algorithms and Applications, First. ed. Chapman and Hall/CRC. [6]Agrawal, R., Gehrke, J., Gunopulos, D., Raghavan, P., 1998. Automatic subspace clustering of high dimensional data for data mining applications. ACM 27, 94–105.

[7]Agrawal, R., Srikant, R., Others, 1994. Fast algorithms for mining association rules. In: 20th Int. Conf. Very Large Data Bases, VLDB.

[8] Alipio, M.I., Cruz, Dela, A.E.M., Doria, J.D.A., Fruto, R.M.S., 2017. A smart hydroponics farming system using exact inference in bayesian network. In: 2017 IEEE 6th Global Conference on Consumer Electronics (GCCE), pp. 1–5.

[9]Amara, J., Bouaziz, B., Algergawy, A., others, 2017. A deep learning-based approach for banana leaf diseases classification. In: BTW (Workshops), pp. 79–88.

[10] Amin, M., AmanUllah, M., Akbar, A., 2014. Time series modeling for forecasting wheat production of Pakistan. J. Anim. Plant Sci. 24, 1444–1451.

[11]Anand, J., Perinbam, J.R.P., 2014. Automatic irrigation system using fuzzy logic. AEIJMR 2, 1–9.

[12]Ankerst, M., Breunig, M.M., Kriegel, H.-P., Sander, J., 1999. OPTICS: ordering points to identify the clustering structure. In: ACM Sigmod Record, pp. 49–60.

Links:

[13]https://www.kaggle.com/omnarayansharmalohar/plant-diseases-classification
[14]https://www.kaggle.com/prasadkevin/crops-prediction-indian-dataset
[15]https://www.kaggle.com/tusharagg/agriculture-data-analysis?select=produce.csv