Machine Learning Assisted Plant Breeding–A Key Step to Overcome Agricultural Production Challenges

Dr. Ashima Gakhar

Associate Professor, Department of Botany, KVA DAV College for Women, Karnal – 132001, Haryana, India

Abstract - With the global population expected to mount nearly 10 billion by 2050 and climate changes shifting the growing conditions of crops, plant breeders have to accelerate their research for finding optimum traits of crops with high yield and resistance to climatic mayhems. The heavy responsibility on the breeders to find optimum traits always poses greater challenges when confronted with data obtained fromgenomics and phenomics studies, biotic and abiotic stress analysis, genetic diversity assessment,yield component analysis,yield stability approaches,and environmental interactions studies. Although traditional statistical methods continue to play important roles in data handling, cutting-edge technologies with progress in computational sciences are leading a revolution in plant breeding.Precise predictions can only be possible from the massive data, when integrated into appropriate models developed for the investigation of the complex interactions of these crucial factors. In this context,machine learning (ML) imparts a vital role in data-mining and analysis, offeringsignificant information for decision-making towards accomplishing breeding targets.Various machine learning tools like convolutional neural networks (CNNs), Artificial neural networks (ANNs), random forest (RF), support vector machines (SVMs) reproducing kernel Hilbert space (RKHS) and deep neural networks (DNNs),can be implemented as a tool to execute and manage data automatically via algorithms for data categorization, cataloguing and predictions.Modern plant breeding methods can also be precisely planned and predicted with various machine learning tools. The accuracy and the precision offered by the machine learning assisted breeding can be a breakthrough for the production of superior quality crops with maximum yield.

Key Words : Plant breeding, Machine learning, Traits, convolutional neural networks (CNNs), Artificial neural networks (ANNs).

I.INTRODUCTION

Climatic changes and the swellingworld population are the two major current global concerns which requires to be addressed with urgency,in order to avoid the escalation of the shortage in food production. The rise in temperature and frequent heat waves over the globe, increase in the frequency of serious precipitation events,swollen greenhouse emissions (up to 400%),intense atmosphericCO_2concentration(by100%), the unpredictability of rain, frequent incidences of natural disasters, water and soil pollution are the key contributors of lower agricultural productivity.[1]Contemporaneously, The world's population, currently 5300 million, is increasing by about 250 000 people per day, while the number of people living in developing countries will expand to over 900 million (United Nations Population Division, 1989,UNFPA, 1989) by 2050. A 70% increase in world agricultural food production is necessary to satisfy the food demands of the predicted population. (Figure.1). But world has lost a third of its cultivable land due to erosion and urbanization with in the past 40 years, and this peremptorilyskewed farmlands becominginsufficient to feed the projected population globally. Improving the soil, crops, water management and stress-tolerant varieties are the some of the choices to overcome the unfavourable impacts of climate changeson food production,food safety.[2] Development of stress resistant breeds of crops with maximum production is the most crucial segment of this process.

Plant breeding is an important dynamic technique of agricultural science, widely used to improve the crops and ensuring secured food production.Itbegun with straightforward selection of notable plants with superior characteristics. Later, many statistics tools were widely used in classical plant breeding. With the advancements in genetics and biotechnologicalmethodologies, modern plant breeding techniques are also emerged, which made the research and implementation more predictive.Advanced phenotyping platforms, Next-generation sequencing (NGS) technologies, genomics and phenomics in genome-wide association studies (GWAS) are some of the leading approaches for the new revolution in plant breeding[3]. Most of these plant breeding techniques are “multipledependent variables versus multiple-independent variables.” Under these circumstances, one regression model is required for each output and this will lead to the generation of complex big datasets.[4]

Drawing the conclusion with normal statistical tools are worryingly troublesome and many times proven misleading. Scientists require weighty data mining tools to successfully handle and to predict / explain complex data obtained from these breeding techniques. This review discuss the various applications of machine learning in both traditional and modern plant breeding and how, its acceptability can be a remedy for the world's food production crisis.

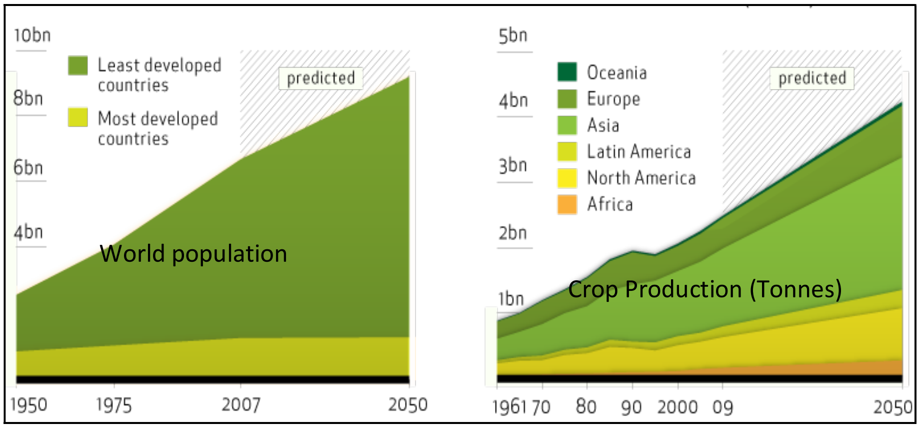

Figure1.The Predicted World Population and Crop Production by 2050
Source: (United Nations Population Division, 1989,UNFPA, 1989)

II. MACHINE LEARNING IN PLANT BREEDING

Machine learning (ML) is a category of artificial intelligence (AI) that permits software applications to become more accurate at predicting outcomes. Machine learning algorithms use historical data as input to predict new output values which can be widely applied in both classical and in vitro-based plant breeding studies to decode the flow of information about plants from the DNA sequence to the observed phenotypes.[5]
Machine learning practices, can be categorized as supervised and unsupervised models, linear and nonlinear algorithms, and shallow and deep learning models (Figure1).[5,6] Convolutional neural networks (CNNs), Artificial neural networks (ANNs), random forest (RF), support vector machines (SVMs) reproducing Kernel Hilbert space (RKHS) and deep neural networks (DNNs), are the most popular ones for processing nonlinear data in plant studies.[7,8]Supervised learning optimizes an extrapolative model by fitting its parameters on a identified training data, comprising of inputs and subsequent known outputs. The resulting models can then be used to predict new, unseen test data. Unsupervised methodologies do not consider outputs, usually by avoiding "training data''. The patterns found by unsupervised approaches can widely be used to interpret large data and allow the researchers toprepare data for more successful supervised learning, by focussing on relevant patterns.

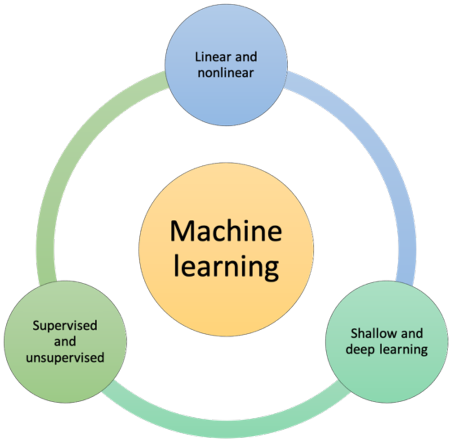

Figure 2. Types of Machine Learning Approaches.

Deep learning approach uses networks containing ''neurons'', interconnected in such a way that signals can be transferred through the network. Typically, neurons are grouped into layers, and deep learning commonly works with several of such layers. A vast number of artificial neural networks are available, each adapted to specific attributes of input data. The successful adoption of ML in plant breeding is highly depend on the availability of well quantified, correctly labelled "trained data" and the choice of the suitable model.[9]

III.APPLICATION OF MACHINE LEARNING IN PLANT BREEDING

Even though older statistical genetics methods are widely used by the researchers, for the past few years, a paradigm shift towards broad usage of machine learning algorithms in biological studies for prediction and discovery is been observed. With the escalating availability of more and different types of omics big- data, the application of machine learning methods, especially deep learning tactics, has become more apparent.The figure.3 shows the various areas of plant breeding where machine learning can be applied.

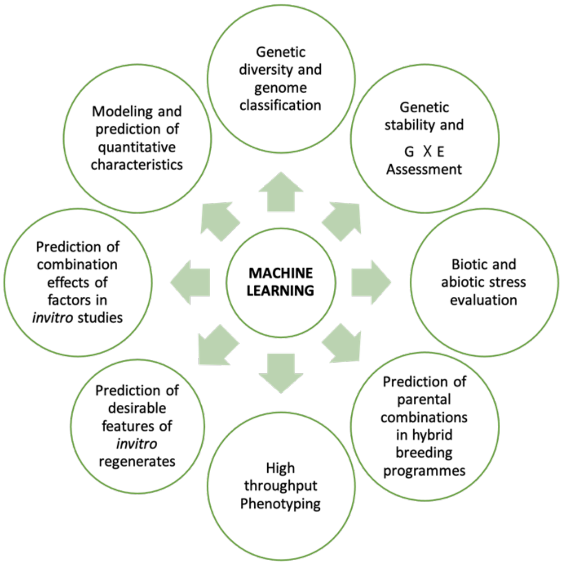
Genetic diversity and genome classification
Genetic stability and G X E Assessment
Biotic and abiotic stress evaluation
Prediction of parental combinations in hybrid breeding programmes
High throughput Phenotyping
Prediction of desirable features of *invitro* regenerates
Prediction of combination effects of factors in *invitro* studies
Modeling and prediction of quantitative characteristics
MACHINE LEARNING

Figure 3. Applications of Machine Learning in Traditional and Modern Plant Breeding

3.1. Biotic and abiotic stress assessments

In their life cycle, plants are liable to a range of biotic and abiotic stresses. Completely different approaches will be used to assess the tolerance and resistance of plant genotypes to those stresses and to spot superior genotypes. Harmonic mean (HARM), yield stability index (YSI), yield index (YI), relative drought index (RDI), mean productivity (MP), tolerance (TOL), stress tolerance index (STI), geometric mean productivity (GMP), stress susceptibility index (SSI), and modified stress tolerance index (K1STI and K2STI) are the parameters used for studying the plants' tolerance to drought.[5,10-17]These classic approaches are based on morphological data. But a wide range of biochemical and physiological pathways involved as the plants' response to environmental stresses can also be targeted for these studies. So combining phenomics data with genomic and metabolomic data can be acompetent strategy to proceed with biotic and abiotic stresses assessments.
The usual multivariate statistical practises are not efficient enough to handle these large quantity of data. Machine learning techniques, in combination with imaging techniques could be used to simulate and envisage genotypes' responses to stressful conditions.[18] By using phenomics and omics (metabolomic and genomic) data, It can also be applied to predict the more resistant variant to stress and nonstress conditions.Traditional machine learning methodsalong with deep CNN areused for identification / categorization of crop stress and various plant diseases with 95-100% accuracy.

3.2. Classification and assessment of genetic diversity

Genetic diversity is the most sort after quality required for plant breeding programmes. It is thoroughly investigated by analysing physiological, biochemical and morphological markers. Genetic and molecular markers are also be widely used for diversity measurements. Principal component analysis (PCA), discriminant function analysis, K- nearest neighbour, support vector machine, cluster analysisare some of the traditionalmultivariable analytical tool used for these studies. Prediction of genetic diversity with these tools are extremely time consuming and often demands feature extraction.[19]But with the application of ML algorithms like convolutional neural networks (CNNs), Artificial neural networks (ANNs) implementation of object detection in genetic diversity assessments can be a big success with high accuracy.

3.3. Indirect Prediction And Yield Component Analysis

A higher and improved output yield is the prime objective for most of the plant breeding programmes. Even though the yield attribute is controlled by powerful gene expressions, The low inheritability of this is always been a result of the strong influence by the environmental factors. The complexity of this trait is one of the major challenges faced bythe plant breeders. For improving the yield of the crop researchers prefer to select highly correlated traits to have a better influence on the outcome. Multiple regression analysis, Principal component analysis (PCA),correlation coefficient analysis and path analysis are the usually followed statistical approaches. But these tools are based on linear relationship assessments of dependant variables as a function of multiple independent variables, which many times can lead to wrong predictions. Application of nonlinear ML algorithms, can ensure a better yield projections and analysis of nonlinear relationships of yield components. Artificial neural networks (ANNs) are the most efficient ML tool for these predictions and have shown exemplary results when applied to study the effect of atmosphere pressure, Precipitation, temperature, crop disease, snowfall, moisture content, humidity and evaporation rate on the yield predictions of onions, apples, pears, and chives. [20]Multilayer perceptron -ANN(MLP- ANN) is another tool applied by the plant breeders to predict the oil and protein content in sesame and wheat.[21]

3.4. Cross/ Hybrid breeding programmes

To predict the heritability of a cross breeding programme, extensive evaluation of the nature of gene action involved with phenological, morphological, and yield component characteristics is mandatory. A better prediction of parental combinations is very crucial for the choice of superior hybrid varieties. ML assisted ANN algorithms can be applied to understand the parental blends to accomplish the best breeds.

3.5. Yield stability and Genotype X Environmental interactions (GEI)

Environmental fluctuations and genotype X environmental interaction (GEI) are the elements that trigger year to year variations in the yield and phenotypic trait of a specific genotype. GEI Impairs the precise selection of genotype for a targeted trait. Yearly variations can be analysed by studying relative performance of genotype over environmental conditions using stability analysis. Compilation of yield stability,mostly relay on univariate approaches like coefficient of variance, coefficient of regression (S^2_{di}),Wilkenson's regression analysis, and Wrick's ecovalence (Wi). However, multivariate stability studies are more powerful and precise than univariate methodologies when deals with multiple independent and dependant variables. Artificial neural networks (ANNs) drastically decrease the required analyses yet predict faster with higher accuracy.[22]

3.6. Machine learning in modern plant breeding

With the advancement of biotechnology, agricultural science also adopted many of its invitro techniques to improve the productivity. In vitro regeneration is one of those sophisticated techniques used in plant breeding.[19]This approach is useful for effective mass propagation, production of bioactive compounds and germplasm conservation.A wide variety of factors are crucial for the success of this process. These factors are classified as,

1) Initial triggers environmental& physical stimuli
2) epigenetic transcriptional cellular responses
3) Molecules responsible for the formation of progress of stem cells

Apart from these process factors, experimental factors like plant genotype selected, explant type, culture medium component and explant age etc, make the study extremely complex. It is obvious that linear programmes have limitations to interpret from big data combinations. Many recent studies showed effortless integration of these data with ANN models and other neural networksfor better predictions. ML algorithms can even be extended with accuracy for the procedures like artificial polyploidy induction, plant gene transformation techniques, gene editing methods, double haploid products etc.

3.7. Machine learning in plant phenomics and genomics

The genotype-to-phenotype gap is another important problem exist in modern plant breeding.[23]While genomics research has equipped with advanced tools to yield information about the genetic make-up of various plant species, but the analysis and processing of the big data generatedis always been a case of concern .

Advanced sequencing technologies allow longer reads of sequences in comparison to the typical short reads. However, a larger amount sequencing errors (5–15%) is existential in most of the times. To overcome this constraint, machine learning tools like artificial neural networks can prominently be used (Figure 4).Convolutional neural network (CNN) model to predict indel variants or SNP,zygosity, and indel length from aligned long reads. Supervised ML has been applied to study the rates of recombination in a target genomeand accurate prediction of the presence of a candidate variant.

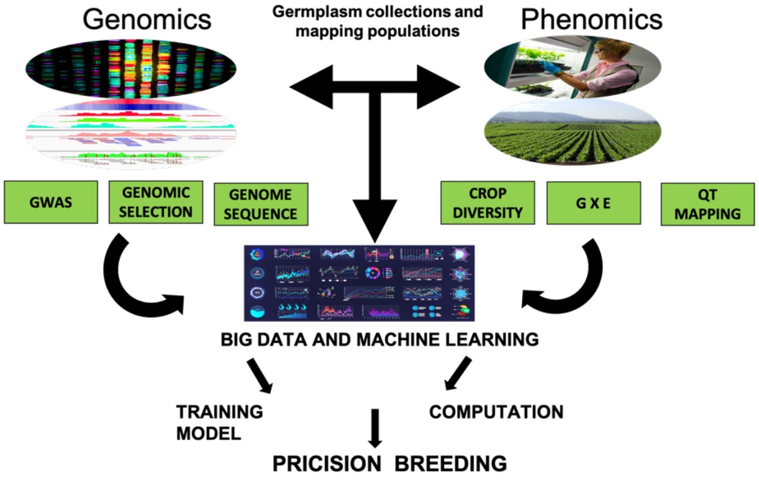

Figure 4. Integration of Genomic and Phenomics Data with Machine Learning

Classic phenotyping tools, prominently relay onmanual measurement of selected traitsfrom a small pool of plants, have very limited output and therefore impairinclusive analysis of attributes within a single plant and across cultivates. This so-called *phenotyping tailback* limits our ability to comprehend how expressed phenotypes linked with underlying genetic factors and environmental conditions.This obviously slowed down the success of breeding programmes. Different ML approaches, supervised and unsupervised, can be modified and applied for plant phenotyping.

3.8. Machine learning in image processing

The phenotype of plants is due to complicated interactions between the genotype and environment, analysis of which requires accurate determination of factors such as soil composition, weather conditions and availability of water. So it is important to establish a better correlation between environmental response, plant performance, and gene function. Image-based procedures, have a great potential to drastically increase the scale and output of plant phenotyping activities. Revolutionary applications of sensing devices in agriculture provide a better platform to act as a exemplary tool to correlate and predict the indoor and outdoor cultivation conditions to the major physiological changes in plants because of external strains.[9, 20-23]

Simple image processing methods tend to fail when complex non-linear, non-geometricphenotyping tasks are to be handled. Tasks such as vigour ratings, ,disease detection,pod/fruit/leaf counting, injury ratings,ageestimation,and mutant classification offer higher level of obstruction which requires more complicated image processing techniques.[9,24]Machine learning methods such as deep convolutional neural networks, integrate image feature extraction with regression.Connected components analysis,Machine learning techniques, and deep learning have potential to improve the reliability of image-based phenotyping. It allows processing of huge amount of data from sensors and phenotyping platforms, increasing the output and accuracy in analysis.

3.9. Machine learning assisted breeding and natural recourses

Every single plant demands light, water, soil, land, nutrients and growth medium. Data science processing tools like machine learning reduce the number of plants needed to generate a trait, saving not just money and time but energy and vital natural resources. More data is collected precisely on every product, allows to makes better predictions, that are customized to their unique surroundings.

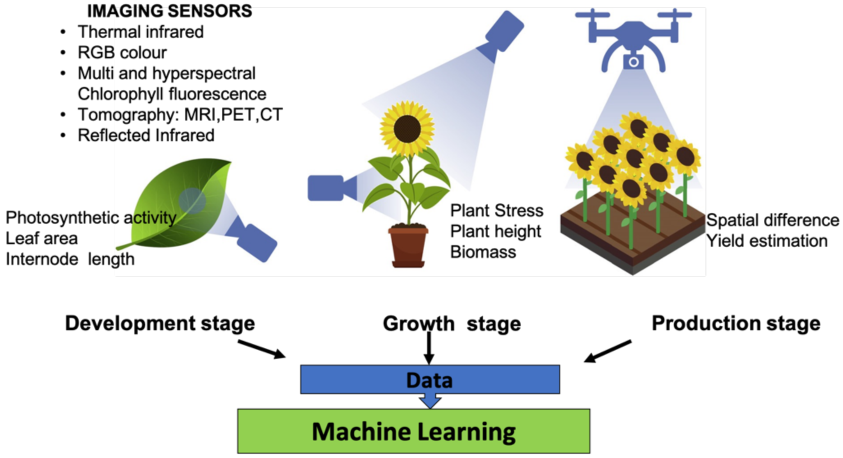

Figure 5. Different imaging tools and phenotype characteristics analysed for plant breeding.

IV. CONCLUSION

With the integration of advanced technologies of computation, especially machine learning (ML) and artificial intelligence (AI), precise breeding with 100% accuracy is likely to be achieved at a faster pace. Eventhough traditional statistical methods predominates the biological research and plant breeding, the new intense and robust machine learning tools hold great promises to impart insights from big and heterogenous research data. Although, the development of a reliable and suitable individual ML algorithm for each complex biological question is difficult to pursue with, the researchers are not leaving any stonesunturned to make it possible because of the rather important problem of feeding the world, to be addressed responsibly.

REFERENCES

[1] S. Ceccarell,"Plant breeding and climate changes", *Journal of Agricultural Science,* vol.148, pp. 627–637,2010.

[2] D.William and C.L. LaxmipathiGowda, "Declining Agricultural Productivity and Global Food Security", *Journal of Crop Improvement,* vol.27, no. 2, pp. 242-254,2013.

[3] S. Esposito, D. Carputo, T.Cardi and P. Tripodi, "Applications and Trends of Machine Learning in Genomics and Phenomics for Next-Generation Breeding", *Plants,* vol.9, no.34, pp.1-18, 2020.

[4] G.R Chegini, J.Khazaei, B.Ghobadian, A.M.Goudarzi, "Prediction of Process and Product Parameters in an Orange Juice Spray Dryer using Artificial Neural Networks". *Journal of Food Engineering,*vol.84, pp. 534–543,2008.

[5] M.Niazian, and G. Niedbała, "Machine Learning for Plant Breeding and Biotechnology",*Agriculture,* vol.10, no.436, pp.1-24,2020.

[6] H. Zheng, W. J. LiJiang,Y. Li, T. Cheng, Y. Tian and Y. Zhu, "Comparative Assessment of Different Modeling Algorithms for Estimating Leaf Nitrogen Content in Winter Wheat Using Multispectral Images from an Unmanned Aerial Vehicle",*Remote Sensing*,vol 10,pp. 20- 26, 2018.

[7] M.Hesami, R. Naderi, M. Y. Najafabadi and M. Rahmati, "Data-Driven Modelling in Plant Tissue Culture", *Journal of Applied Environmental and Biological Sciences,* vol.7,pp. 37–44, 2017.

[8] M.Salehi, S. Farhadi, A. Moieni, N. Safaie, andH. Ahmadi, "Mathematical Modeling of Growth and Paclitaxel Biosynthesis in Corylus avellana Cell Culture Responding to Fungal Elicitors Using Multilayer Perceptron-Genetic Algorithm", *Frontiers in Plant Science*, vol.11,pp. 11-48, 2018.

[9] A.D.J.Dijk, G. Kootstra, W.Kruijer, and D.Ridder, "Machine Learning in Plant Science and Plant Breeding", *iScience*, vol. 24, no.1, pp. 1-24, 2021.

[10] C. Xuand Scott and A. Jackson, "Machine Learning and Complex Biological Data, *Genome Biology*, vol. 20, no. 76,2020.

[11] R. Fischer and R. Maurer, "Drought Resistance in Spring Wheat Cultivars. I. Grain Yield Responses",*Australian Journal of Agricultural Research,* vol. 29, no.5,pp. 897 – 912, 1978.

[12] R.Fischer and J. Wood, "Drought Resistance in Spring Wheat Cultivars. III. Yield Associations with Morpho-Physiological Traits",*AustralianJournal of Agricultural Research,*vol.30, no. 6,pp.1001-1020, 1979.

[13] A. Rosielle and J. Hamblin, "TheoreticalAspectsofSelectionforYieldinStressandNon-StressEnvironment", *Crop Science*, vol. 21,pp. 943–946,1981.

[14] M. Bouslama and W.T. Schapaugh, "StressToleranceinSoybeans.I.EvaluationofThreeScreeningTechniques for Heat and Drought Tolerance", *Crop Science,* vol.24, pp. 933–937,1984.

[15] G.Fernandez, "EffectiveSelectionCriteriaforAssessingStressTolerance", InProc.Int Sym. Adapt. Veg. Food Crop. Tem. Water Stress, Taiwan, 1992.

[16] P. Gavuzzi, F. Rizza, M. Palumbo, R. G. Campanile, G. L. Ricciardi and B. Borghi, "Evaluation of field and laboratory predictors of drought and heat tolerance in winter cereals", *Canadian journal of plant science*, vol.77,pp. 523–531,1997.

[17] K.A. Schneider, R. Rosales-Serna, F. Ibarra-Perez, B. Cazares-Enriquez and J.A. Acosta-Gallegos,"Improving Common Bean Performance under Drought Stress",*Crop Science,*Vol 37,pp.43–50, 1997.

[18] E. Farshadfar and J. Sutka, "Screening Drought Tolerance Criteria in Maize", *Acta Agronomica Hungarica*, vol. 50,pp. 411–416, 2002.

[19] C. Pandolfi, S.Mugnai, E. Azzarello, S. Bergamasco, E. Masi andS. Mancuso, "Artificial Neural Networks as a Tool for Plant Identification: A Case Study on Vietnamese Tea Accessions",*Euphytica,*vol. 166, pp. 411–421, 2009.

[20] S. Lee, Y.Jeong, S. Son and B. Lee, "A Self-Predictable Crop Yield Platform (SCYP) Based on Crop Diseases Using Deep Learning", *Sustainability,* Vol.11, 3637- 3648, 2019.

[21] G. Jedbala, D. Kurasiak-Popowska, K. Stuper-Szablewska and J. Nawracała, "Application of Artificial Neural Networks to Analyze the Concentration of Ferulic Acid, Deoxynivalenol, and Nivalenol in Winter Wheat Grain", *Agriculture,* vol. 10, no.127, pp. 88- 101,2020.

[22] A.L. Harfouche, D.A. Jacobson, D. Kainer, J.C. Romero, A.H. Harfouche, G. S. Mugnozza andM. Moshelion, "Accelerating Climate Resilient Plant Breeding by Applying Next-Generation Artificial Intelligence". Trends in Biotechnology., vol. 37, pp. 1217–1235, 2019.

[23] R. Jordan and I. Stavness, "Deep Plant Phenomics: A Deep Learning Platform for Complex Plant Phenotyping Tasks", *Frontiers in Plant Science,*vol. 8, no 1190pp. 1-25, 2017.

[24] J. M. Papeand C. Klukas, "3-D histogram-based segmentation and leaf detection for rosette plants". *Lecture in Computer* Science.,vol. 89, no.28, pp.61–74, 2010.