https://doi.org/10.2352/EI.2022.34.6.IRIACV-264

DEEP LEARNING BASED WHEAT EARS COUNT IN ROBOT IMAGES FOR WHEAT PHENOTYPING

Ehsan Ullah, Mohib Ullah, Muhammad Sajjad, Faouzi Alaya Cheikh

Norwegian University of Science and Technology, Norway

ABSTRACT

The number of spikes, spikelets per spike, number of spikes per square meter are essential metrics for plant breeders and researchers in predicting wheat crop yield. Evaluating the crop yield based on wheat ears counting is still done manually, which is a labor-intensive, tedious and costly task. Thus, there is a significant need to develop a real-time wheat spikes/ears counting system for plant breeders for effective and efficient crop yield predictions. This paper proposed two deep learning-based methods based on EfficientDet and Faster-RCNN to detect and count the spikes. The images are taken using high-throughput phenotyping techniques under natural field conditions, and the algorithms localize and automatically count wheat spikes/ears. Faster R-CNN with Resnet50 as backbone architecture produced an overall accuracy of 88.7% on the test images. We also used recent state-of-the-art models EfficientDet-D5 and EfficientDet-D7, having backbone architectures EfficientNet-B5 and EfficientNet-B7, respectively. A comprehensive quantitative analysis is performed on the standard performance metrics. In the analysis, the EfficientDet-D5 model produces an accuracy of 92.7% on the test images, and EfficientDet-D7 produces an accuracy of 93.6%.

***Index Terms*—** Wheat Spikes, Deep Learning, Faster R-CNN, EfficientDet.

1. INTRODUCTION

Wheat is one of the most important and widely utilized crop species consumed daily by the public. 762.7 million tons of annual wheat production was recorded by [1] in 2020. Wheat is cultivated every year in around 215 million hectares, and the global trade of wheat is estimated at nearly 50 billion US dollars every year [2]. It is estimated that nearly 750.1 million tons of wheat is consumed every year globally [1]. Every coming year, the demand for grain is increasing. At the same time, extreme weather situations and variations in climate changes increase the risk of an uncertain supply of grains. Complex, multivariate, and unpredictable agricultural environments need to be better studied to solve these challenges by monitoring, measuring/analyzing, and constantly evaluating different physical aspects and phenomena. This helps researchers and plant breeders to know and recognize better-yielding and more stress-tolerant plant species. With the availability of large scale dataset, deep learning has revolutionized many fields including but limited to tracking [3–7], virtual reality [8, 9], cybersecurity [10, 11], crowd analysis [12–14], animal farming [15, 16], segmentation [17–19], classification [20, 21], facial emotion recognition [22, 23]. However, in agriculture sector, it is not exploited yet. In several research problems related to plant phenotyping, conventional Machine Learning (ML) methods have been used widely. ML models, including SVM, decision trees, Bayesian, and instance base model, have been used in crop yield prediction, disease detection, weed detection, plant species detection, and crop quality analysis [24]. Some ML-based techniques exist to automatically detect heading and flowering in wheat [25] to distinguish growth stages in field-grown wheat; a bag-of-visual-words method is used. Low-level characteristics are collected using the SIFT algorithm. Finally, to classify the growth levels in plants, the classification of support vector machines is used. Hyperspectral imaging systems with a five waveband of 20nm are also explored to examine symptoms of yellow rust disease, and nitrogen stress using hyperspectral features [26]. Crop growth characteristics are measured based on online multilayer soil data of satellite imagery, an unsupervised learning algorithm was used, and field variations in wheat yield were predicted [27]. However, Smart agriculture and plant phenotyping have progressed into the big data era. Massive data is sourced from open field trials, indoor plant phenotyping using advanced platforms such as UAV, satellite imagery, grounded robot vehicles, gantries, etc. With the availability of a large amount of data and recent high-end computing power of hardware [28]. Deep learning models are preferred as their performance increases with the increase in the data we provide to the model. This is one of the main reasons why deep learning approaches took over the traditional machine learning approaches. Secondly, Deep learning surpasses the need of manually selecting and defining handcrafted features [28]. Instead, deep learning approaches perform optimization in a complete end-to-end way by mapping input data samples to outputs targets. The detection of wheat heads from images in itself is a challenging task. It involves several factors like the observational conditions, genotypic differences, and development stages of the plant. Wheat

head density (the number of wheat heads per unit ground area) is a significant yield component. However, because the process of evaluation of this parameter is still manual and labor-intensive, measurement errors of around 10% can be observed. [29] [30] Thus, developing automated image-based methods that can bring this error down is essential so that breeders can manipulate the balance between yield parameters in their breeding selections. In this paper, we use the Global Wheat Detection Dataset (GWHD) [31] which contains images taken at 90 degrees from above of a wheat field with the wheat head annotated using bounding boxes. These images contain occlusions, overlapped wheat ears, blurred background, etc., which makes it a perfect dataset for training any deep learning model. We used two different deep learning models, Faster-RCNN [32] and EfficientDet [33] for the detection of wheat ears and trained them with the Global wheat head dataset. The main objective of this study was to build a data-driven, efficient system that will detect the wheat ears with good performance and accuracy. The rest of the paper is organized as follows. In section 2, the detection model adopted in the study are briefly explained. In section 3 the data pre-processing, model architecture, training, and evaluation methods are elaborated. The quantitative results are listed in section 4 and section 5 concludes the paper with the final remarks.

2. DETECTION MODELS

In this paper, along with Faster R-CNN for detecting wheat ears, we used the recently published state-of-the-art deep learning model purposed by google brain researchers called EfficientDet which has a robust backbone architecture called EfficientNet. [34].

2.1. Faster R-CNN

Faster R-CNN, developed by Ren et al [32] is an object detection network composed of a feature extraction network which is typically a pre-trained CNN. It consists of two networks: a regional proposal network(RPN) for generating region proposals and a convolutional network which takes the proposed regions to detect objects almost in real-time. Thus, in addition to convolutional neural network, Faster R-CNN has a RPN which is inserted after the last convolutional layer making it different from its predecessors. RPN efficiently predicts region proposals with a wide range of scales and aspect ratios.

2.2. EfficientDet

EffificientDet consists of three parts as shown in Figure 1. The first part is the pre-trained EfficientNet as the backbone architecture of the model. The second part is BiFPN, which do the top-down and bottom-up feature fusion multiple times for the output characteristic of Level 3-7 in EfficientNet. The third part is the classification and detection box prediction network, to regress and classify the wheat ear frame.

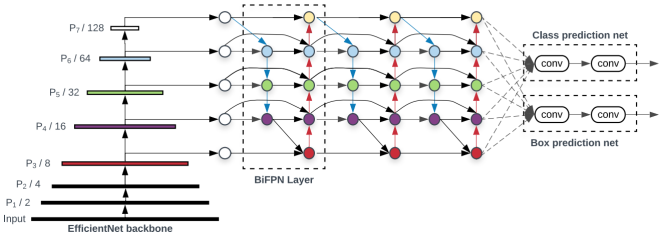

Fig. 1: Architecture of EfficientDet [34]

3. DATA PREPARATION & TRAINING

To perform the wheat head detection and counting, we followed three steps: starting with the exploratory data analysis and preprocessing, followed by training the deeplearning models, and finally using several evaluation metrics to evaluate the results.

3.1. Data Analysis and Preprocessing

In detecting objects of interest, such as wheat spikes, ambient noise poses significant challenges for computer vision-based techniques. Some challenges include the following:

- The movements of plants and/or the stability of handheld cameras are likely to cause blurred images.
- Due to natural conditions and light variations in the field, dark shadows or sharp brightness can appear in images.
- Overlaps between the ears due to the floppy attitude of the ears can also give rise to additional difficulties especially with the presence of awns in certain cultivars.
- Over development phases, spikes in various varieties change dramatically, as spikes display no correlation between the early and later growth phases.

Pre-processing is a preliminary phase in the analysis of images, which helps to arrange data properties in order to enable subsequent steps and also to achieve fair final results. At first, the GWHD dataset was analysed. The dataset is gather from several parts of the world, with a total of 4698 squared patches extracted from the 2219 original high-resolution RGB images. It contains 188,495 labelled heads with an average of 20 to 60 heads per image. There are also around 100 images that don't contain any heads to represent actual capturing conditions and make the task more difficult. We tried several data augmentation techniques to improve the performance of our models. In addition to the usual data augmentation methods employed in normal computer vision tasks along with other

transform methods, the ones used in our approach were horizontal/vertical flips, cropping and resizing, change to gray, cutout [35], cutmix [36], hue/saturation value changes, and brightness/contrast changes. The data augmentation and image pre-processing will help in producing more samples and variations and help in training the models to decrease overfitting and increase the generalization of our models.

3.2. Training

3.2.1. Faster R-CNN

For training, we used a normal simple random sampling from the dataset we obtained from the above step. We used 80%-20% splitting for training set and validation data. Initially, we started from the pre-trained model on the pedestrian images and did some fine-tuning to adapt to our use case.

We studied the results of the model using a Resnet50 backbone, learning rate of 0.005 and CosineAnnealing scheduler [37] with Stochastic Gradient Descent (SGD) as the optimizer. The main reason behind using a cosine function for the learning function is the idea that for each batch of the SGD, the network should get very close to the global minimum value for the loss, means we don't want the algorithm to overshoot and the learning rate should get smaller helping the loss value settle to some point. Cosine annealing decreases the learning rate following the cosine function and helps in making this global minimum stable.

We also tried the Adam optimizer and tried to see how it performs in comparison to SGD. Using these parameters, we trained the model for 40 epochs with the batch size of 8.

The results are represented in the plots below.

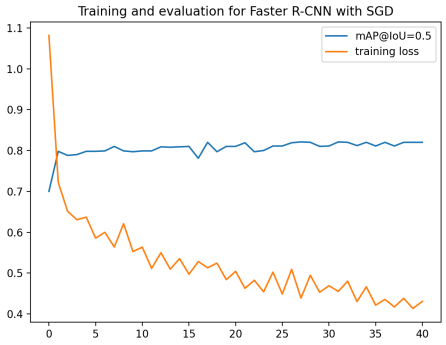

Fig. 2: Faster R-CNN training and evaluation with SGD

3.2.2. EfficientDet

EfficientDet- D5 and EfficientDet-D7 were used as our detection models for detecting wheat ears effectively. We used wheat ears GWHD dataset with images of 15 different wheat varieties captured under different environment conditions.

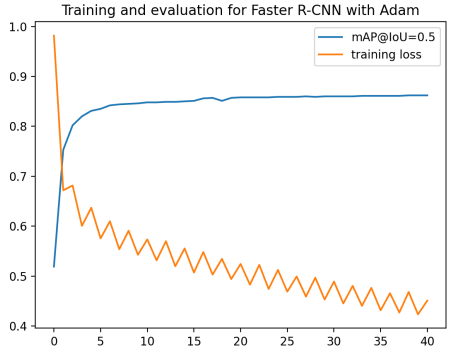

Fig. 3: Faster R-CNN training and evaluation with Adam

We utilized all those images for training and validation of the model. In this study, EffificientDet-D5 and EffificientDet-D7 were trained respectively. We used Pytorch framework version 1.6.0 and Python 3.7. we use the CUDA/10.0.130 version for graphics cards. we trained our model using Idun high performance computing cluster at NTNU Trondheim using only one GPU which was NIVIDIA V100 Tensor Core. The images with input size of 512x512 was introduced to the model and the model is trained for 40 Epochs. The Average loss error on both of the model is saved and its shown in the below figures 9, 10.

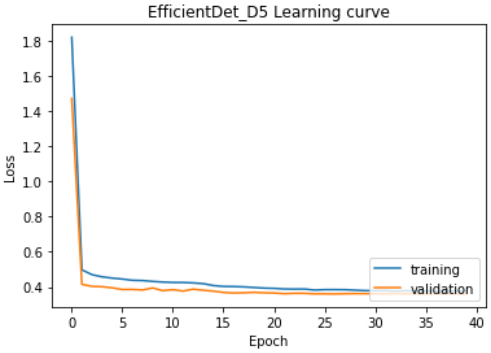

Fig. 4: EfficientDet-D5 training and loss error

4. EVALUATION AND RESULTS

For the evaluation of the results, we used the training error along with the mean average precision (mAP) from the standard MS COCO metrics [38] for the validation set. The mAP values relies on the Intersection over Union(IoU) values. The IoU value is the area of intersection between the actual bounding box divided by their union's area. A True Positive prediction is the one with IoU $>$ threshold, whereas False Positive

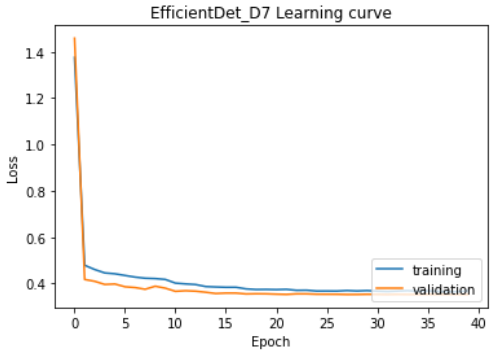

Fig. 5: Efficient-D7 training and loss error

refers to one with IoU $<$ threshold.

$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}} \tag{1}$$

Similarly, we used the regular precision, recall, and accuracy for the test set.

4.0.1. Precision and Recall

Precision is the ratio between true positives and all positives, whereas recall is the measure of a model identifying true positives.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{\text{TP}}{\text{\# ground truths}} \tag{2}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} = \frac{\text{TP}}{\text{\# predictions}} \tag{3}$$

4.0.2. Accuracy

Accuracy, the simplest of the metrics, is the ratio of total number of correct predictions to the number of predictions.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FP} + \text{FN}} \tag{4}$$

We used 10 test images that were not used during the training or evaluation phase and calculated the above metrics.We achieved overall 87.4% accuracy using Faster-RCNN with SGD optimizer and 88.7% accuracy using Adam optimizer. For EfficeintDet Models, The EfficientDet-D5 achieved overall accuracy of 92.7% . EfficientDet-D7 produce better results than Faster-RCNN with Resnet as its backbone architecture and EfficientDet-D5.The EfficeintDet-D7 model achieved 93.6% accuracy on the test images. The results are represented in tables below.

According to our findings, trained models struggle to detect wheat ears that are extensively overlapped and occluded, and the shape structure of the wheat ears is completely morphed amongst each other. Also the trained models completely neglect the clipped wheat ears in test images.

Table 1: Faster R-CNN with SGD Optimizer on Test Data

ImageID	GT	Detected	Precision	Recall	Accuracy
2fd875eaa	27	24	1.0	0.89	88.9%
51b3e36ab	27	29	0.86	0.93	80.6%
51f1be19e	18	18	1.0	1.0	100.0%
53f253011	31	29	1.0	0.94	93.5%
348a992bb	37	36	0.97	0.95	92.1%
796707dd7	31	23	1.0	0.74	74.2%
aac893a91	24	21	0.95	0.83	80.0%
cb8d261a3	24	21	1.0	0.88	87.5%
cc3532ff6	26	29	0.9	1.0	89.7%
f5a1f0358	28	31	0.9	1.0	90.3%
Total	273	261	0.95	0.91	87.4%

Table 2: Faster R-CNN with Adam optimizer on Test Data

ImageID	GT	Detected	Precision	Recall	Accuracy
2fd875eaa	27	24	1.0	0.89	88.9%
51b3e36ab	27	29	0.9	0.96	86.7%
51f1be19e	18	18	1.0	1.0	100.0%
53f253011	31	29	1.0	0.94	93.5%
348a992bb	37	36	0.97	0.95	92.1%
796707dd7	31	25	1.0	0.81	80.6%
aac893a91	24	21	0.95	0.83	80.0%
cb8d261a3	24	21	1.0	0.88	87.5%
cc3532ff6	26	29	0.9	1.0	89.7%
f5a1f0358	28	31	0.9	1.0	90.3%
Total	273	263	0.96	0.92	88.7%

Table 3: EfficientDet-D5 results on Test Data

Precision, Recall and Accuracy of the EfficientDet-D5 Model					
ImageID	GT	Detected	Precision	Recall	Accuracy
2fd875eaa	27	24	0.88	0.88	88%
53f253011	31	30	0.96	0.96	96%
51b3e36ab	27	25	0.92	0.92	92%
51f1be19e	18	18	1.0	1.0	100%
348a992bb	37	38	0.97	1.0	97%
796707dd7	31	26	0.83	0.83	83%
aac893a91	24	19	0.79	0.79	79%
cb8d261a3	24	24	1.0	1.0	100%
cc3532ff6	26	25	0.92	0.96	92%
f5a1f0358	28	28	1.0	1.0	100%
Total	273	257	92.7%	93.4%	92.7%

Table 4: EfficientDet-D7 results on Test Data

Precision, Recall and Accuracy of the EfficientDet-D7 Model					
ImageID	GT	Detected	Precision	Recall	Accuracy
2fd875eaa	27	24	0.88	0.88	88%
53f253011	31	30	0.96	0.96	96%
51b3e36ab	27	25	0.92	0.92	92%
51f1be19e	18	18	1.0	1.0	100%
348a992bb	37	35	0.94	0.94	94%
796707dd7	31	26	0.83	0.83	83%
aac893a91	24	21	0.87	0.87	87%
cb8d261a3	24	24	1.0	1.0	100%
cc3532ff6	26	25	0.96	0.96	96%
f5a1f0358	28	28	1.0	1.0	100%
Total	273	256	93.6%	93.6%	93.6%

5. CONCLUSIONS

Agriculture plays a critical role in the global economy, and pressure on the agricultural system will increase with the continuing expansion of the human population. Digital Agriculture or precision farming has arisen as new scientific fields that use intense data approaches to drive agricultural productivity while minimizing its environmental impact. The data generated in modern agricultural operations is provided by various sensors that enable researchers to understand the morphological properties of the crops better, leading to more accurate and faster crop yield predictions. In this study, we use a data-driven deep learning approach to accurately identify and count wheat ears/spikes in digital images taken in an open field environment. We used two variants of Faster-RCNN, EfficientDet-D5, and EfficientDet-D7, to detect the target ears/spikes in the wheat crop images. We achieved an accuracy of 88.7% using Faster-RCNN, 92.7% accuracy on EfficientDet-D5, and 93.6% accuracy on efficientDet-D7, respectively. The proposed model performance can be enhanced by introducing more data during the training phase with varying illuminations and environmental conditions (occlusions, overlapping, blur) to learn rich feature representations of wheat ears. Introducing an attention module to backbone architecture can be another way of enhancing the performance of these models. Also, increasing the contrast between the wheat canopy and wheat's ears will boost the accuracy of the already trained models.

6. REFERENCES

[1] "FAO world food situation," https://www.fao.org/worldfoodsituation/csdb/en/, Accessed: 05-11-2020.

[2] "CGIAR wheat in the world," https://wheat.org/wheat-in-the-world/, Accessed: 05-11-2020.

[3] Mohib Ullah, Ahmed Kedir Mohammed, Faouzi Alaya Cheikh, and Zhaohui Wang, "A hierarchical feature model for multi-target tracking," in *2017 IEEE international conference on image processing (ICIP)*. IEEE, 2017, pp. 2612–2616.

[4] Zolbayar Shagdar, Mohib Ullah, Habib Ullah, and Faouzi Alaya Cheikh, "Geometric deep learning for multi-object tracking: A brief review," in *2021 9th European Workshop on Visual Information Processing (EUVIP)*. IEEE, 2021, pp. 1–6.

[5] Mohib Ullah and Faouzi Alaya Cheikh, "Deep feature based end-to-end transportation network for multi-target tracking," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 3738–3742.

[6] Shoudong Han, Piao Huang, Hongwei Wang, En Yu, Donghaisheng Liu, and Xiaofeng Pan, "Mat: Motion-aware multi-object tracking," *Neurocomputing*, 2022.

[7] Mohib Ullah, Faouzi Alaya Cheikh, and Ali Shariq Imran, "Hog based real-time multi-target tracking in bayesian framework," in *2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 2016, pp. 416–422.

[8] Stellan Ohlsson, Thomas G Moher, and Andrew Johnson, "Deep learning in virtual reality: How to teach children that the earth is round," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, 2000, vol. 22.

[9] Habib Ullah, Sultan Daud Khan, Mohib Ullah, and Faouzi Alaya Cheikh, "Social modeling meets virtual reality: An immersive implication," in *International Conference on Pattern Recognition*. Springer, 2021, pp. 131–140.

[10] Theyazn HH Aldhyani and Hasan Alkahtani, "Attacks to automatous vehicles: a deep learning algorithm for cybersecurity," *Sensors*, vol. 22, no. 1, pp. 360, 2022.

[11] Jun Zhang, Lei Pan, Qing-Long Han, Chao Chen, Sheng Wen, and Yang Xiang, "Deep learning based attack detection for cyber-physical system cybersecurity: A survey," *IEEE/CAA Journal of Automatica Sinica*, vol. 9, no. 3, pp. 377–391, 2021.

[12] Habib Ullah, Ihtesham Ul Islam, Mohib Ullah, Muhammad Afaq, Sultan Daud Khan, and Javed Iqbal, "Multi-feature-based crowd video modeling for visual event detection," *Multimedia Systems*, vol. 27, no. 4, pp. 589–597, 2021.

[13] Qian Wang and Toby P Breckon, "Crowd counting via segmentation guided attention networks and curriculum loss," *IEEE Transactions on Intelligent Transportation Systems*, 2022.

[14] Mohib Ullah, Habib Ullah, Nicola Conci, and Francesco GB De Natale, "Crowd behavior identification," in *2016 IEEE international conference on image processing (ICIP)*. IEEE, 2016, pp. 1195–1199.

[15] Cheng Fang, Tiemin Zhang, Haikun Zheng, Junduan Huang, and Kaixuan Cuan, "Pose estimation and behavior classification of broiler chickens based on deep neural networks," *Computers and Electronics in Agriculture*, vol. 180, pp. 105863, 2021.

[16] Akif Quddus Khan, Salman Khan, Mohib Ullah, and Faouzi Alaya Cheikh, "A bottom-up approach for pig skeleton extraction using rgb data," in *International Conference on Image and Signal Processing*. Springer, 2020, pp. 54–61.

[17] Chungang Zhuang, Zhe Wang, Heng Zhao, and Han Ding, "Semantic part segmentation method based 3d object pose estimation with rgb-d images for bin-picking," *Robotics and Computer-Integrated Manufacturing*, vol. 68, pp. 102086, 2021.

[18] Habib Ullah, Mohib Ullah, and Muhammad Uzair, "A hybrid social influence model for pedestrian motion segmentation," *Neural Computing and Applications*, vol. 31, no. 11, pp. 7317–7333, 2019.

[19] Mohib Ullah, Ahmed Mohammed, and Faouzi Alaya Cheikh, "Pednet: A spatio-temporal deep convolutional neural network for pedestrian segmentation," *Journal of Imaging*, vol. 4, no. 9, pp. 107, 2018.

[20] Tom Vermeire, Dieter Brughmans, Sofie Goethals, Raphael Mazzine Barbossa de Oliveira, and David Martens, "Explainable image classification with evidence counterfactual," *Pattern Analysis and Applications*, pp. 1–21, 2022.

[21] Yun Ding, Jinpeng Feng, Yanwen Chong, Shaoming Pan, and Xiaohui Sun, "Adaptive sampling toward a dynamic graph convolutional network for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, 2021.

[22] Carmen Bisogni and Chiara Pero, "Ifepe: On the impact of facial expression in head pose estimation," in *International Conference on Pattern Recognition*. Springer, 2021, pp. 486–500.

[23] Abdulrahman Alreshidi and Mohib Ullah, "Facial emotion recognition using hybrid features," in *Informatics*. Multidisciplinary Digital Publishing Institute, 2020, vol. 7, p. 6.

[24] Konstantinos G Liakos, Patrizia Busato, Dimitrios Moshou, Simon Pearson, and Dionysis Bochtis, "Machine learning in agriculture: A review," *Sensors*, vol. 18, no. 8, pp. 2674, 2018.

[25] Pouria Sadeghi-Tehran, Kasra Sabermanesh, Nicolas Virlet, and Malcolm J Hawkesford, "Automated method to determine two critical growth stages of wheat: heading and flowering," *Frontiers in Plant Science*, vol. 8, pp. 252, 2017.

[26] Xanthoula Eirini Pantazi, Dimitrios Moshou, Roberto Oberti, Jon West, Abdul Mounem Mouazen, and Dionysios Bochtis, "Detection of biotic and abiotic stresses in crops by using hierarchical self organizing classifiers," *Precision Agriculture*, vol. 18, no. 3, pp. 383–393, 2017.

[27] Xanthoula Eirini Pantazi, Dimitrios Moshou, Thomas Alexandridis, Rebecca L Whetton, and Abdul Mounem Mouazen, "Wheat yield prediction using machine learning and advanced sensing techniques," *Computers and Electronics in Agriculture*, vol. 121, pp. 57–65, 2016.

[28] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[29] Simon Madec, Xiuliang Jin, Hao Lu, Benoit De Solan, Shouyang Liu, Florent Duyme, Emmanuelle Heritier, and Frederic Baret, "Ear density estimation from high resolution rgb imagery using deep learning technique," *Agricultural and forest meteorology*, vol. 264, pp. 225–234, 2019.

[30] Md Mehedi Hasan, Joshua P Chopin, Hamid Laga, and Stanley J Miklavcic, "Detection and analysis of wheat spikes using convolutional neural networks," *Plant Methods*, vol. 14, no. 1, pp. 100, 2018.

[31] Etienne David, Simon Madec, Pouria Sadeghi-Tehran, Helge Aasen, Bangyou Zheng, Shouyang Liu, Norbert Kirchgessner, Goro Ishikawa, Koichi Nagasawa, Minhajul Arifin Badhon, et al., "Global wheat head detection (gwhd) dataset: a large and diverse dataset of high resolution rgb labelled images to develop and benchmark wheat head detection methods," *arXiv preprint arXiv:2005.02162*, 2020.

[32] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 6, pp. 1137–1149, 2016.

[33] Mingxing Tan, Ruoming Pang, and Quoc V Le, "Efficientdet: Scalable and efficient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10781–10790.

[34] Mingxing Tan and Quoc V Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," *arXiv preprint arXiv:1905.11946*, 2019.

[35] Terrance DeVries and Graham W. Taylor, "Improved regularization of convolutional neural networks with cutout," 2017.

[36] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," 2019.

[37] Ilya Loshchilov and Frank Hutter, "Sgdr: Stochastic gradient descent with warm restarts," 2017.

[38] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár, "Microsoft coco: Common objects in context," 2015.