frontiers in Plant Science

ORIGINAL RESEARCH
published: 17 June 2021
doi: 10.3389/fpls.2021.673505

Wheat Spike Blast Image Classification Using Deep Convolutional Neural Networks

*Mariela Fernández-Campos[1], Yu-Ting Huang[2], Mohammad R. Jahanshahi[2,3], Tao Wang[4], Jian Jin[4], Darcy E. P. Telenko[1], Carlos Góngora-Canul[1,5] and C. D. Cruz[1]**

[1] Department of Botany and Plant Pathology, Purdue University, West Lafayette, IN, United States, [2] Lyles School of Civil Engineering, Purdue University, West Lafayette, IN, United States, [3] School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, United States, [4] Department of Agricultural and Biological Engineering, Purdue University, West Lafayette, IN, United States, [5] Tecnológico Nacional de México /IT Conkal, Conkal, Yucatán, Mexico

Wheat blast is a threat to global wheat production, and limited blast-resistant cultivars are available. The current estimations of wheat spike blast severity rely on human assessments, but this technique could have limitations. Reliable visual disease estimations paired with Red Green Blue (RGB) images of wheat spike blast can be used to train deep convolutional neural networks (CNN) for disease severity (DS) classification. Inter-rater agreement analysis was used to measure the reliability of who collected and classified data obtained under controlled conditions. We then trained CNN models to classify wheat spike blast severity. Inter-rater agreement analysis showed high accuracy and low bias before model training. Results showed that the CNN models trained provide a promising approach to classify images in the three wheat blast severity categories. However, the models trained on non-matured and matured spikes images showing the highest precision, recall, and F1 score when classifying the images. The high classification accuracy could serve as a basis to facilitate wheat spike blast phenotyping in the future.

Keywords: wheat blast, convolutional neural networks, inter-rater agreement, severity classification, plant disease phenotyping, breeding, deep learning, controlled conditions

OPEN ACCESS

Edited by:
Michele Pisante, University of Teramo, Italy

Reviewed by:
Alexandr Muterko, Russian Academy of Sciences, Russia
Pouria Sadeghi-Tehran, Rothamsted Research, United Kingdom

***Correspondence:**
C. D. Cruz
cruz113@purdue.edu

Specialty section:
This article was submitted to Technical Advances in Plant Science, a section of the journal Frontiers in Plant Science

Received: *27 February 2021*
Accepted: *10 May 2021*
Published: *17 June 2021*

Citation:
Fernández-Campos M, Huang Y-T, Jahanshahi MR, Wang T, Jin J, Telenko DEP, Góngora-Canul C and Cruz CD (2021) Wheat Spike Blast Image Classification Using Deep Convolutional Neural Networks. Front. Plant Sci. 12:673505. doi: 10.3389/fpls.2021.673505

INTRODUCTION

Wheat blast is an emergent disease caused by the Ascomycetous fungus *Magnaporthe oryzae Triticum* (MoT). MoT was first detected in Brazil in 1985, with successive spread to Bolivia, Paraguay, and Argentina (Igarashi et al., 1986; Barea and Toledo, 1996; Viedma, 2005; Cabrera and Gutiérrez, 2007; Perello et al., 2015). In 2016, a wheat blast outbreak was first reported in Bangladesh, apparently due to the unintended importation of MoT-infected South American grain (Aman, 2016; Malaker et al., 2016). Many countries in South Asia are actively monitoring wheat fields for the presence of MoT (Bhattacharya and Pal, 2017; Mottaleb et al., 2018). In 2020, MoT presence was reported in Zambia, Africa, which summates another continent to the list (Tembo et al., 2020). Cruz et al. (2016b) predicted areas at risk in the United States (southern and pacific northwest states) for MoT establishment and the threat of this pathogen to soft- and hard-red winter wheat production.

MoT can infect leaves, stems, and seeds, although the most remarkable and studied symptoms are associated with the spike (Igarashi et al., 1986; Cruz et al., 2015; Cruz and Valent, 2017; Ceresini et al., 2019). The infection by MoT of the spike, spikelets, or rachis causes the wheat spike blast, inducing partial or complete bleaching of the spikes (Igarashi, 1986). Infection can cause shriveled grain reducing the grain quality and yield. A wide range of disease intensities can occur depending on the susceptibility of cultivars planted and the prevalent weather conditions (Goulart and Paiva, 1992).

Warm temperatures, excessive rain, long and frequent spike wetness, and limited fungicide efficacy exacerbate the intensity of wheat blast epidemics, especially in susceptible cultivars (Goulart et al., 2007). The optimum conditions for wheat blast development include a temperature range between 25 and 30°C and spike surface wetness between 25 and 40 h (based on controlled conditions) (Cardoso et al., 2008). Under conducive field conditions, the fungus can kill up to 100% of susceptible wheat spikes in a period of 2.5–3 weeks (Gongora-Canul et al., 2020).

Since 1985, when wheat spike blast was first detected, intense efforts have been undertaken to identify resistance (Igarashi et al., 1986; Urashima et al., 2004; Prestes et al., 2007; Cruz et al., 2016b; Ceresini et al., 2019; Cruppe et al., 2020). Recently, two new genes, Rmg8 and RmgGR119, were found to generate resistance to wheat blast (Wang et al., 2018). However, the only currently effective resistance provided by the $2N^VS$ translocation from *Aegilops ventricosa* (Tausch) confers useful yet partial and environment and/or genetic background-dependent resistance to wheat blast (Cruz et al., 2016a; Valent, 2016; Cruppe et al., 2019, 2020). Obtaining tissue samples from phenotyped wheat entries and testing for the presence or absence of the $2N^VS$ segment is relatively easy and routine (Cruz et al., 2016b; Yasuhara-Bell et al., 2018; Cruppe et al., 2019). Although there is evidence that $2N^VS$-based resistance may be overcome, additional sources of wheat spike blast resistance should be identified (Cruz et al., 2016b; Cruppe et al., 2019, 2020; Juliana et al., 2020). Thus, there is a continued need to find new sources of resistance to wheat blast.

Plant disease estimations, or phytopathometry, refer to the measurement and quantification of plant disease severity (DS)or incidence that is essential when studying and analyzing diseases at organ, plant, or population levels (Large, 1966; Bock et al., 2010). Plant disease estimations by human raters are the standard method used for plant disease phenotyping. Humans are trained to perform visual disease evaluations of incidence and severity, and their reliability can be improved with experience. These estimations are helpful, but they are subjective evaluations that can introduce variability and can be time-consuming and labor-intensive (Nutter et al., 1993; Madden et al., 2007; Bock et al., 2010, 2020). Due to issues associated with an agreement in data acquisition, inter-rater agreement among other statistical tests can be used to compare the consensus or agreement between estimations of raters of DS (Nutter et al., 1993; Madden et al., 2007; Bock et al., 2010, 2020). These agreement analyses are relevant in plant pathology and plant breeding since inaccurate disease estimations can cause imprecision and unreliability leading to incorrect conclusions (Chiang et al., 2016; Singh et al., 2021).

A bottleneck in the identification of novel sources of resistance is measuring disease intensity (i.e., plant disease phenotyping), which is considered a limiting factor in the assessment of genotype performance in plant breeding programs (Mahlein, 2015; Shakoor et al., 2017). Therefore, innovative and transformative solutions for the quantification of plant disease symptoms at the individual and host population levels are needed (Camargo and Smith, 2009; Kumar et al., 2020). Implementation of advanced computer vision and machine learning techniques could reduce the phenotyping bottleneck during breeding and enhance the understanding of genotype–phenotype relationships (Fiorani and Schurr, 2013; Kruse et al., 2014; Shakoor et al., 2017; Yang et al., 2020; Singh et al., 2021).

Computer vision, machine learning, and deep learning methods have recently been adapted to agriculture due to increased knowledge of algorithms and model capabilities that can learn and make predictions from images Red Green Blue (RGB), multispectral, or hyperspectral (Barbedo, 2016; Kersting et al., 2016; Mahlein et al., 2018). There are two ways in which these models are trained, one is supervised learning, which depends on an annotated dataset, and another is unsupervised learning, which does not rely on annotations (Mahlein et al., 2018). The most frequently used deep learning methods are the Convolutional Neural Networks (CNN). The CNN is characterized by high-accuracy metrics for image recognition and image segmentation. Recent studies have further enhanced the scope of a deep-learning-based approach for classifying, identifying, and quantifying plant diseases (Mahlein et al., 2018; Singh et al., 2018; Barbedo, 2019).

A variety of CNN classification models are available for plant diseases. These include models for bacterial pustule (*Xanthomonas axonopodis pv. glycines*), sudden death syndrome (SDS, *Fusarium virguliforme*), Septoria brown spot (*Septoria glycines*), bacterial blight (*Pseudomonas savastanoi pv. glycinea*), and several abiotic stresses in soybean (Ghosal et al., 2018). In tomato (*Solanum lycopersicum*), deep-learning models were developed with and without pre-training models with images from nine leaf tomato diseases from the website www.PlantVillage.org, obtaining better performance using pre-training models (Brahimi et al., 2018). A total of 54,306 leaf images from several crops with 26 diseases were obtained from PlantVillage.org and trained using AlexNet and GoogleLeNet pre-trained models with a leaf-segmented dataset, obtaining an accuracy of 99.35% (Mohanty et al., 2016). On wheat, an in-field automatic diagnosis system for powdery mildew (*Blumeria graminis f. sp. tritici*), smut (*Urocystis agropyri*), leaf blotch (*Septoria tritici*), black chaff (*Xanthomonas campestris pv. undulosa),* stripe rust (*Puccinia striiformis f. sp. tritici*), and leaf rust (*Puccinia recondita f. sp. tritici*) were developed using deep-learning, and multiple instances–learning techniques from the Wheat Disease Database 2017 (Lu et al., 2017). Although this database is a significant contribution to wheat disease identification based on images, aspects regarding the reliability of the labeler may be compromised (Lobet, 2017). It is appropriate that detection and quantification studies of

plant disease provide evidence of ("true") estimation agreement analysis before using the labeled images as a dataset for training deep-learning models. Currently, phenotyping of wheat spike blast DS relies on a visual estimation made by humans (Cruz et al., 2016a). We hypothesized that deep CNN models can be trained for wheat spike blast severity image classification under a controlled environment. To test this hypothesis, we focused on the following objectives:

i) Evaluate the agreement in data acquisition of the human rater who collected and classified datasets.
ii) Develop an accurate deep CNN model to detect and classify wheat spike blast symptoms in three severity categories.

MATERIALS AND METHODS

Ethics

A written informed consent was obtained from the individual for the publication of any potentially identifiable images or data included in this article.

Plant Cultivation and Genetic Materials

Two experiments were conducted under controlled conditions in a growth room at the Asociación de Productores de Oleaginosas y Trigo (ANAPO) research facility in Santa Cruz de la Sierra, Bolivia. Wheat cultivars were planted in pots of 15 cm diameter, filled with vermicast:silt (3:1 [v/v]), and grown at 18−25°C, 14 h light/10 h dark photoperiod, and 50–60% relative humidity. Plants were fertilized, and insecticides were sprayed when needed. Plants were arranged in a randomized complete block design with wheat cultivars having various levels of resistance to MoT, two inoculation levels (inoculated and non-inoculated), and four replicates. Wheat cultivars with a range of sensitivity to the wheat blast were used for the experiments. Experiment one included Bobwhite and South American spring cultivars Atlax, BR-18, Motacú, Urubó, AN-120, Sossego, and San Pablo and for experiment two the cultivars included BR-18, San Pablo, Bobwhite, and Atlax (Baldelomar et al., 2015; Fernández-Campos et al., 2020).

Inoculation

Plants were inoculated at the growth-stage Feekes 10.5, when the spike had completely emerged, with MoT isolate 008-C (**Figure 1A**), according to a modified inoculation protocol previously published (Cruz et al., 2016a). A conidial suspension was adjusted to 20,000 spores/ml, and each spike received 1 ml of the spore suspension. Immediately after the spikes were sprayed with the MoT inoculum, plants were moved to a dew chamber (**Figure 1B**) to induce MoT infection (i.e., 24–26°C, 95–98% RH, and 14 h light photoperiod). Forty-eight hours after inoculation, plants were removed from the dew chamber and left under controlled environment room conditions [(24–26°C and relative humidity of 50–60%), until day 19 after inoculation; **Figure 1B**].

Data Collection, DS, and Disease Measurements

Following phytopathometry terminology, we used the term "estimate" for visual disease estimations made by humans and the term "measurement" for estimations made by image analysis (Bock et al., 2010; Gongora-Canul et al., 2020). Visual estimate of DS was obtained by observing the disease area covered in the spike and assigned a corresponding severity value from 0 to 100%. In this study, image analysis disease measurements were achieved by manually measuring spike disease area (pixels) using RGB color threshold segmentation with the image analysis software Fiji ImageJ v.1.52a (Schindelin et al., 2012; Sibiya and Sumbwanyambe, 2019). First, the measurement of the total spike area was obtained, then the diseased area was measured. Finally, the percentage of diseased severity (DS) of the individual spike was calculated (Equation 1), where $A_{Diseased}$ is the proportion of the area of spike that is diseased divided by the total area of the spike A_{Total} (See **Video 1** in **Supplementary Material**).

$$DS = \frac{A_{Diseased}}{A_{Total}} \times 100 \quad (1)$$

Visual estimations of wheat spike blast symptoms were taken seven times after inoculation in each experiment. In experiment one, visual estimations and images were collected 4, 6, 9, 12, 14, 16, and 19 days after inoculation (DAI) and in experiment two, 0, 5, 7, 10, 12, 14, and 19 DAI. Each spike side (four sides total) was visually estimated for DS by Rater 1 (a plant pathologist with experience on wheat blast, rice blast, and other diseases). Simultaneously, an image from each spike side was captured perpendicular to the spike with a distance of 50 cm approximately with a DSLR EOS 6D Canon camera (Canon Inc., Tokyo, Japan) (**Figure 1D**) using a photography studio set up

FIGURE 1 | Wheat blast image collection flow process: **(A)** *Magnaphorthe oryzae* pathotype *Triticum* inoculation, **(B)** After inoculation, plants were moved to the yellow dew chamber that provided optimal conditions for fungal infection for 48 h, later transfer to the black trays, **(C)** wheat spike imaging set up, and **(D)** an image was captured perpendicular to the spike.

with umbrellas, lights, and screens (Neewer 2.6 m × 3 m/8.5 ft × 10 ft Background Support System and 800 W 5,500 K Umbrellas Softbox Continuous Lighting Kit for Photo Studio Product) that helped create a uniform light and smooth environment (**Figure 1C**).

DS Categories

The total spike disease estimations of Rater 1 paired with the corresponding image were converted to a three-category scale according to the amount of severity that served to fed training and testing dataset of CNN model. The category selection was based on wheat blast results from published work conducted over the last decade (Baldelomar et al., 2015; Cruz et al., 2016b; Vales et al., 2018; Cruppe et al., 2020; Fernández-Campos et al., 2020). Category 1 (healthy spikes) was used as a baseline (i.e., negative control or fully immune). Category 2 showed 0.1–20% severity (resistant and moderately resistant/low levels of symptoms) corresponding to the selected putative population for successive trials under variable conditions (controlled environment or field). Category 3 showed 20.1–100% severity (moderately susceptible and susceptible/intermediate and high levels of symptoms) corresponding to the plant population that will not be selected for successive trials because of the high potential to be or to become susceptible to the disease studied (**Figure 2**).

Testing Reliability of Rater 1: Inter-rater Analysis of Wheat Spike Blast Severity Estimations

Rater 1 played a critical role in estimating DS and classifying into categories of all the images belonging to Dataset 1 and Dataset 2 (Datasets are described in the section, generation of data sets according to wheat spike physiological changes). Therefore, an inter-rater analysis was needed to determine the reliability of visual estimations of Rater 1. Inter-rater agreement assesses the degree of agreement between two and more raters who obtain independent ratings about the characteristics of a set of subjects. Subjects of interest include people, things, or events that are rated (Madden et al., 2007; Everitt and Anders, 2010; Bock et al., 2020).

To determine the agreement of disease estimations of Rater 1, we performed an inter-rater analysis including a second-rater, and ImageJ was used as an image analysis software baseline. Rater 2, is a plant pathologist and expert in the wheat blast. ImageJ is an image analysis software used to measure plant diseases from images.

We used the power analysis Wilcoxon signed-rank test to determine the sample size for the inter-rater agreement studies of the two training datasets. The test consisted of DS estimations or measurements of 31 and 29 images from the CNN training Dataset 1 and training Dataset 2, respectively. From now on, the 31 images selected from Dataset 1 will be called sample Dataset 1 and the 29 images from Dataset 2 will be referred to as sample Dataset 2. Rater 2, who is an experienced researcher with more than 4 years of working with the wheat blast disease, visually estimated DS from the sample Dataset 1 and Dataset 2. Additionally, disease measurements were obtained from the sample Dataset 1 and the sample Dataset 2 using ImageJ software as indicated above. Ultimately, the DS results of visual disease estimations of human raters and ImageJ measurements were compared. The estimated and measured DS values from both samples were analyzed for inter-rater agreement in two scenarios, one with a scale of 0–100% DS (continuous data), and the other with the images divided into three categories of DS (ordinal data). We, therefore, computed Lin's Concordance Coefficient, Fleiss kappa, and weighted kappa statistics.

The Lin's concordance coefficient (ρc or CCC) is used to estimate the accuracy$_1$ between two raters using continuous data. From the analysis, we obtained the estimation of accuracy$_1$, precision$_1$, and bias of the disease estimations and disease measurements between the two raters (Lin, 1989; Madden et al.,

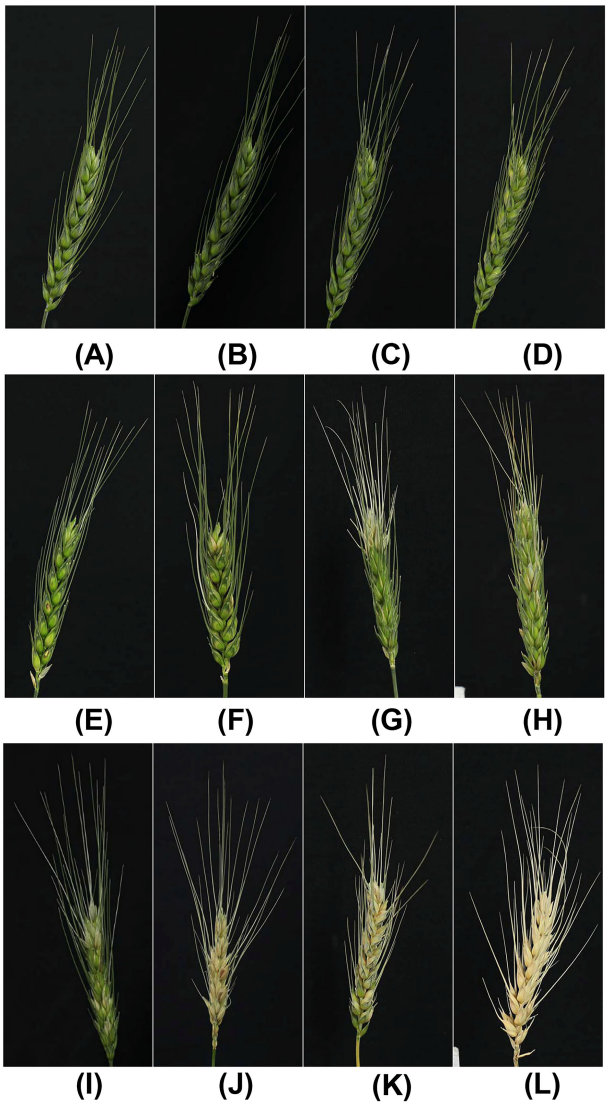

FIGURE 2 | Examples of images per category: **(A–D)** healthy wheat spikes no disease (0% severity, Category 1); **(E–H)** spikes with moderate severity (0.1–20%, Category 2); and **(I–L)** spikes with high severity (20.1–100%, Category 3).

2007; Bock et al., 2010). For accuracy$_1$ (ρc) and precision$_1$ (r), values range from 0 to 1; values close to 1 indicate high accuracy$_1$ and precision$_1$. Bias (Cb) ranges from 0 to 1, and values close to 1 indicate less bias (Nita et al., 2003). Lin's concordance analysis was performed by using PROG REG ALL procedure on SAS v.9.4 (Cary, NC), based on the macro developed by Lawrence Lin and verified by Min Yang (Lin et al., 2002).

To determine the degree of association between the estimation of categorical information provided by the two raters (inter-rater agreement), the weighted kappa statistics were computed (Chmura, 1992; Graham and Jackson, 1993; Nelson and Edwards, 2015). The Fleiss kappa coefficient was used to compare the agreement of categorical information among all raters, (i.e., Rater 1, Rater 2, and ImageJ) (Fleiss et al., 2003). The values of both the weighted kappa and Fleiss kappa coefficients range from 0 to 1. Values from 0.5 to 1 indicate that the agreement is better than what is expected by chance (Nelson and Edwards, 2015; Tang et al., 2015; Mitani et al., 2017; Gamer et al., 2019). The Fleiss kappa statistics and weighted kappa were computed with the *irr* package of the R software (Team, 2017).

Generation of Datasets According to Wheat Spike Physiological Changes

Wheat was inoculated at the growth-stage Feekes 10.5 (spike completely emerged) of the host plant. Approximately every 2 days after the inoculation, the spike images were collected to capture the changes developed. Indirectly, progressive physiological changes in spikes were recorded, as maturing begins at wheat growth-stage Feekes10.5.4 (kernels watery ripe) and continues through the growth-stage Feekes 11.4 (mature kernels) (Large, 1954; Wise et al., 2011). During this period, the kernel hardened, and the green spike lose its color (maturing), which mimic the typically bleached spikes caused by wheat spike blast symptoms.

Two datasets were generated considering the (color) physiological changes that can lead to confusion when training the CNN model. Dataset 1, included maturing and non-matured wheat spikes; and Dataset 2 included only non-matured spikes (data available at: https://purr.purdue.edu/publications/3772/1). The proposed CNN model was trained using the two datasets. Each dataset was randomly separated into the training and testing datasets. The CNN model automatically extracted the features of each image in the training dataset to learn a good classifier, whereas the testing dataset was used to evaluate the performance of the trained CNN model. In general, an unseen dataset was applied to evaluate the CNN model to ensure that the model was not under-fitting or over-fitting. In this research, 80% of the images were categorized and used as the training set, and the remaining 20% as the testing set. **Table 1** lists the original distribution of the number of images in Dataset 1 and Dataset 2. Although Category 3 covers a large variability, it does not mean the number of the data in Category 3 is larger than the other two categories. The number of images in each category was extremely imbalanced and using them indiscriminately could have resulted in a biased model. Fortunately, there are several viable methods to cope with the disproportionate training data in each category.

TABLE 1 | Training and testing data distribution and the number of images used in Dataset 1 and Dataset 2.

Sets	Category 1	Category 2	Category 3
Dataset 1 (Maturing and non-matured spikes)			
Training	1,595	640	402
Augmented training	1,595	1,920	1,608
Testing	381	178	110
Dataset 2 (Non-matured spikes only)			
Training	1,430	386	307
Augmented training	1,430	1,544	1,535
Testing	327	120	90

Category 1: 0% severity, Category 2: 0.1–20% severity, Category 3: 20.1–100% severity.

Data augmentation is a common technique providing a viable solution to data shortage issues by adding copies of original images with modification or noise (Boulent et al., 2019). Data augmentation was used in this study to balance the number of images in each category. In this study, images were randomly flipped horizontally and vertically in order to increase the number of images in Categories 2 and 3. Thus, for Dataset 1, training data were triplicated in Category 2 and quadrupled in Category 3 (**Table 1**). For Dataset 2, training data were quadrupled in Category 2 and quintupled in Category 3 (**Table 1**).

Deep CNN Model

In recent years, the feasibility of using artificial intelligence, in particular deep learning, has been expanded into a variety of applications (Atha and Jahanshahi, 2018; Chen and Jahanshahi, 2018; Kumar et al., 2018; Wu and Jahanshahi, 2019). Deep learning is a subset of machine learning that enables computers to automatically extract features from a huge amount of data and learn to classify data.

In this study, wheat spike blast symptoms were automatically detected and classified into three severity categories using a pre-trained CNN model. This model may be more efficient than classifying images visually. To obtain a general and reliable CNN model, the network needed to be trained using a large labeled training dataset. The performance of the CNN model is highly dependent on the number and quality of the training data. However, it was hard to collect a wheat blast dataset having a million images in a short time. The performance CNN model can easily lead to under- or over-fitting due to the lack of a large dataset for training. To address this issue, transfer learning was used as a practical solution where a network was trained using a typically different larger dataset such as ImageNet. A major advantage of using transfer learning is that it can adapt the parameters trained from an abundant number of images. Transfer learning starts with a pre-trained model, e.g., VGG16 model, and replaces the fully-connected (FC) layers of the model with new FC layers. A network trained on the ImageNet dataset was used to initialize the network parameters, and the whole

network was fine-tuned since the nature of our dataset was very different from the ImageNet dataset. In this study, an FC layer that consisted of three nodes, representing three categories, were appended to the end of the network. A residual neural network architecture (ResNet101), a CNN model with 101 layers with recurrent connection trained on ImageNet data (He et al., 2015, 2016), was selected as the pre-trained model. Furthermore, as shown in **Table 1**, it was extremely difficult to obtain a large number of images in each category. An unbalanced dataset can result in a biased CNN model. To address this issue in the dataset, the loss function, which was used to optimize the parameter in a neural network, was transformed into a weighted loss function (Equation 2) by assigning individual weights to each category. Equation (2) defines the cross-entropy loss function in the CNN model, where ω_{category} is the assigned weight to each of the categories, the first term in Equation (2) is a negative log-likelihood loss, and the second term in Equation (2) is log-softmax. Four cases of study were tested with an individual weight set to the loss functions assigned to different categories. In the experiments, "cases" refer to specific combinations of weight loss functions for each of the three DS categories (**Table 2**). Case 1 was the non-weight set [1, 1, 1], with all categories sharing the same class weight. Case 2 used [1, 10, 1] class weights in the loss function, meaning that the highest weight was for Category 2, which includes plants at early disease stages and low levels of disease symptoms. Case 3 used [2, 5, 1] class weights in the loss function, meaning that the higher weight was assigned to Categories 1 (no symptoms) and 2 (early stages and low levels of disease symptoms). Case 4 had class weights [2, 1, 1] in the loss function, assigning a higher weight to category 1 (no symptoms) (**Table 2**).

$$loss\left(x,\ category\right) = -\omega_{\text{category}} * \log \frac{e^{x_{\text{category}}}}{\sum_{j=1}^{N} e^{x_j}}$$

$$= -\omega_{\text{category}}\ \left(-x_{\text{category}} + \log\left(\sum_{j}\exp(x_j)\right)\right) \quad (2)$$

The network was trained for 15 epochs using a stochastic gradient descent optimizer (Bottou, 2010), a learning rate of 0.0001 was used, and the batch size was 16. Additionally, 5-folds cross-validation was applied to the training process. The training took place on a Linux server with Ubuntu 14.04. The server included two Intel Xeon E5-2620 v4 CPUs, 256-GB DDR4 memories, and four NVIDIA Titan X Pascal GPUs. Pytorch (Paszke et al., 2017) was used to implement the CNN.

Model Performance Evaluation

The performance of the CNN model was evaluated *via* the classified results of the testing dataset. A 3 × 3 confusion matrix was used to describe the prediction result of the model. Each row of the confusion matrix represented the ground truth of the data, and each matrix column corresponded to a predicted category by the CNN model. Thus, the diagonal elements of the matrix, called true positive (TP), were the number of wheat images correctly classified into the ground truth. The false positive (FP) for each Category was the sum of all errors in that column. For example, the FP of Category 1 was the number of Category 2 and Category 3 severities that were incorrectly classified as Category 1. Based on the confusion matrix, additional evaluation metrics were calculated.

TABLE 2 | Two datasets trained the CNN model with four cases of the study through different weights in loss functions for each category.

Model	Values of weighted loss function per category [1, 2, 3]	
	Dataset 1 (Maturing and non-matured spikes)	Dataset 2 (Non-matured spikes)
Case 1	[1, 1, 1]	[1, 1, 1]
Case 2	[1, 10, 1]	[1, 10, 1]
Case 3	[2, 5, 1]	[2, 5, 1]
Case 4	[2, 1, 1]	[2, 1, 1]

[Category 1: 0% severity, Category 2: 0.1–20% severity, Category 3: 20.1–100% severity].

Accuracy$_2$ was defined as the total number of TP among three categories divided by the total number of the predictions. Precision$_2$ was defined as the total number of the TP instances divided by the total number of predicted positive examples, which was the summation of TP and FP instances in the binary classification task (Equation 3). Similarly, the precision$_2$ of the multi-classes task illustrates the number of instances that were correctly predicted given all the predicted labels for a given category. Recall was defined as the TP instance divided by all the positive samples (TP and FN) (Equation 4). F1 score is a single metric that encompasses both precision$_2$ and recall (Equation 5). Accuracy$_2$, precision$_2$, recall, and F1 score metrics ranged from 0 to 1, where higher values indicate the high predictive ability of the model.

$$Precision_2 = \frac{TP}{TP + FP} \quad (3)$$

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

$$F_1 score = 2 \times \frac{precision \times recall}{precision + recall} \quad (5)$$

RESULTS

Cultivar Response to Wheat Spike Blast Under Controlled Conditions

The final wheat spike blast severity was at day 19 after inoculation when cultivar Atlax reached 100% average DS, followed by Bobwhite (99.7%), San Pablo (32.9%), BR-18 (8.7%), Motacú (3.7%), AN-120 (3.31%), Urubó (1.9%), and Sossego (0.83%). Wheat spike blast symptoms developed on all tested cultivars, with reactions to MoT infection consistent with previous reports, except for cultivar San Pablo that showed moderate susceptibility (Baldelomar et al., 2015; Cruz et al., 2016b; Cruppe et al., 2020; Fernández-Campos et al., 2020; Gongora-Canul et al., 2020). Cultivar Atlax exhibited the highest DS of all the cultivars and had a high level of susceptibility to wheat spike blast.

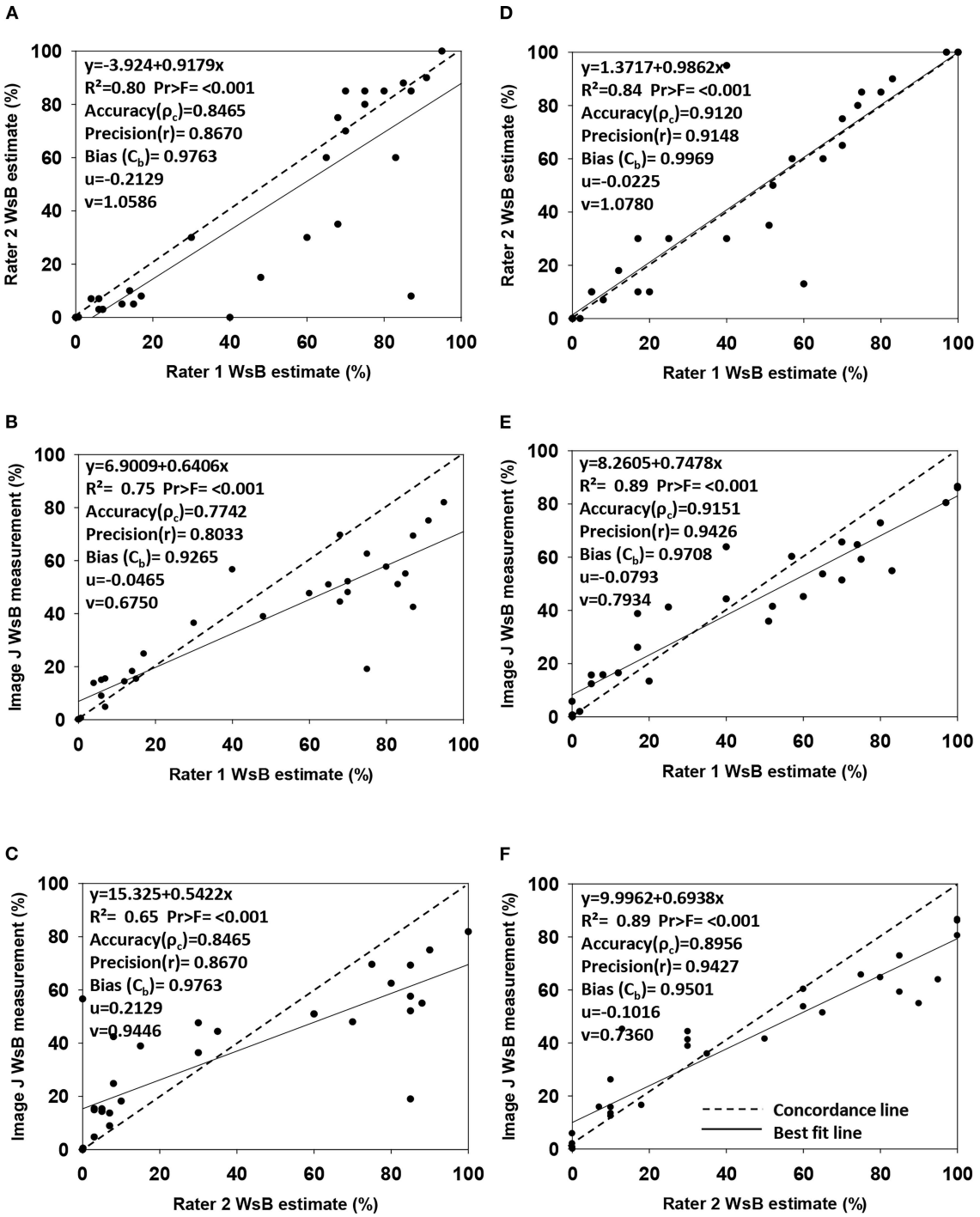

FIGURE 3 | Regression analysis of wheat spike blast DS estimations made by Rater 1 (responsible to estimate the severity of total image dataset) *vs.* Rater 2 (expert in wheat blast) and ImageJ DS measurements (image analysis software). Graphs show accuracy (ρ_c), precision(r), bias (Cb), scale shift (ν), and location shift (u) for wheat spike blast continuous Dataset 1 **(A–C)** ($n = 31$ images) and Dataset 2 **(D–F)** ($n = 29$ images). **(A)** Disease estimation comparison from images Dataset 1 between Rater 1 and Rater 2. **(B)** Disease estimation and disease measurement comparison from images Dataset 1 between Rater 1 and ImageJ. **(C)** Disease estimation and disease measurement comparison from images Dataset 1 between Rater 2 and ImageJ. **(D)** Disease estimation comparison from images Dataset 2 between Rater 1 and Rater 2. **(E)** Disease estimation and disease measurement comparison from images Dataset 2 between Rater 1 and ImageJ. **(F)** Disease estimation and disease measurement comparison from images Dataset 2 between Rater 2 and ImageJ.

TABLE 3 | Values of weighted Kappa (κ) analysis for inter-rater agreement between raters and ImageJ in Dataset 1 (maturing and non-matured spikes) and Dataset 2 (non-matured spikes) of wheat spike blast under controlled environment.

	Dataset 1		Dataset 2	
Categories	**κ**	**z**	**κ**	**z**
Rater 1[x] *vs.* ImageJ[z]	0.882**	4.93	0.822**	4.45
Rater 2[y] *vs.* ImageJ	0.727**	4.13	0.776**	4.32
Rater 1 *vs.* Rater 2	0.747**	4.32	0.849**	4.65

**$p < 0.01$.
[x]*Rater 1: Responsible to estimate the severity of the total image dataset.*
[y]*Rater 2: Expert in the wheat blast.*
[z]*ImageJ: Image analysis software.*

Inter-rater Agreement Analysis

The Lin's concordance correlation analysis showed a high accuracy$_1$ (ρc = 0.89–0.91), high precision$_1$ (r = 0.91–0.94), and less bias (Cb = 0.95–0.99) in the sample Dataset 2 than in the sample Dataset 1 (ρc = 0.77–0.85, precision$_1$ r = 0.80–0.87, and bias Cb = 0.93–0.98) (**Figure 3**). In the sample Dataset 1, the highest accuracy$_1$ was between Rater 1 and Rater 2 (ρc = 0.85) and between Rater 1 and ImageJ (ρc = 0.85). In the sample Dataset 2, the highest accuracy$_1$ value was between Rater 1 and ImageJ (ρc = 0.92), followed by between Rater 1 and Rater 2 (ρc = 0.91). In both sample datasets, strong accuracy$_1$, high precision$_1$, and low bias involved Rater 1, providing evidence that ratings of disease based on continuous data were done correctly for further classification of the images into categories for model training.

The weighted kappa statistics (κ), used to quantify inter-rater agreement, were higher in the sample Dataset 1 than in the sample Dataset 2, with κ = 0.72–0.88 ($p < 0.01$) and κ = 0.78–0.85 ($p < 0.01$), respectively (**Table 3**). In the sample Dataset 1, the highest agreement occurred between Rater 1 and ImageJ (κ = 0.88), and in the sample Dataset 2, the highest agreement was between Rater 1 and Rater 2 (κ = 0.85). In both sample datasets, the substantial agreement involved the ground truth (Rater 1), providing evidence that ratings were done correctly for further classification of the images into categories for model training.

The Fleiss kappa coefficient (Fκ), which compared the association of ordinal categorical information of two or more raters, showed an Fκ = 0.771 ($n = 31$, z = 9.26, $p < 0.001$) for the sample Dataset 1 and 0.697 ($n = 29$, z = 8.1, $p < 0.001$) for the sample Dataset 2, indicating substantial agreement among the human raters and ImageJ in both datasets. However, the sample Dataset 1 possessed a higher Fleiss kappa coefficient index than the sample Dataset 2, both presented substantial agreement between the rates and ImageJ. Yet, the evidence supported the fact that the three raters correctly estimated the amount of the disease from the same image.

Deep CNNs Model Performance

To train the proposed CNN model, two different datasets were used. As mentioned above in the section *Generation of Datasets According to Wheat Spike Physiological Changes*, testing reliability of Rater 1, Dataset 1 included matured and non-matured wheat spikes and Dataset 2 included only non-matured spikes (**Table 1**). Four cases applied different weight set of loss functions in both Datasets (**Table 2**, **Supplementary Figures 1, 2**). The performance of the CNN model was evaluated *via* the classified result of the testing data.

The testing accuracy$_2$ of the model trained with Dataset 1 was 90.1% in Case 1, 90.4% in Case 2, 90.0% in Case 3, and 87.7% in Case 4. The testing accuracy$_2$ of Dataset 2 was 98.4% in Case 1, 93.9% in Case 2, 95.0% in Case 3, and 94.2% in Case 4. Dataset 2 presented higher accuracy$_2$ values compared to Dataset 1, suggesting that the model was accurate. However, it was not sufficient to claim that the model was reliable based on accuracy$_2$ alone since the dataset in this study was unbalanced. In addition to accuracy$_2$, other metrics can help evaluate the performance of the CNN model, such as precision$_2$, recall, and F_1 score.

Precision$_2$ indicates the ability to correctly classify an instance in all predicted positive instances. The focus was on the performance of the CNN model in Category 2 as this was the category that breeders and pathologists will concentrate on for breeding purposes. Dataset 1 Case 2 showed the lowest precision$_2$ (75.4%) among all cases values (**Table 4**). Moreover, the confusion matrix of Dataset 1 Case 2 showed that the model misclassified 38 images of Category 1 (no symptoms) as Category 2 (early disease stages and low levels of disease symptoms), which was the highest number of wrongly classified images among all the cases (**Figure 4B**). This suggested that the class weight of Category 2 might be too high since its misclassified images that belonged to other categories as Category 2. Hence, the class weight combination was modified by lowering the weight in Category 2 and increasing the weight in Category 1 as to not overemphasize the impact from Category 2. Precision$_2$ of Category 2 significantly increased from 75.4% in Case 2 to 84.1% in Case 3, and to 85.0% in Case 4 (**Table 4**). In Case 2, precision$_2$ of Category 2 significantly increased from 75.4% in Dataset 1 to 90.2% in Dataset 2 (**Table 4**). Precision$_2$ of Category 2 significantly increased from 90.2% in Case 2 to 92.7% in Case 3 and from 90.2% in Case 2 to 94.1% in Case 4 (**Table 4**).

The recall metric for evaluating the CNN model that indicates the ability to correctly recognize a category was also used. In datasets 1 and 2, the recall of Category 2 was the lowest, illustrating the challenge of the model to classify images of Category 2 (early disease stages and low levels of disease symptoms) (**Table 4**). The highest recall of Dataset 1 Category 2 was 86.0% in Case 2, and the lowest was 74.2% in Case 1 (**Table 4**). This was expected given that Case 2 had a higher weight in the loss function of Category 2 compared to Case 1 (non-weighted loss function). In Case 2, Dataset 1, the recall values were similar among the three categories (**Table 4**). In Dataset 2 Category 2, the lowest recall was 75.0% in Case 1, and the highest recall was 84.2% in cases 2 and 3 (**Table 4**). The model in these two cases had the highest weight in loss function of Category 2 (early disease stages and low levels of disease symptoms).

F_1 score is a common indicator of the overall performance of the CNN model. In datasets 1 and 2, the F_1 score of Category 2 was the lowest, reaffirming the difficulty of classifying images of

TABLE 4 | Classification performance of the CNN model when classifying the testing set of Dataset 1 (maturing and non-matured spikes) and Dataset 2 (non-matured spikes) in the cases of the study presented different weights in the loss function [weight in Category 1, weight in Category 2, weight in Category 3].

		Dataset 1			Dataset 2		
Model	**Performance Index**	**Category 1**	**Category 2**	**Category 3**	**Category 1**	**Category 2**	**Category 3**
Case 1[(A)]	Precision	0.891	0.852	0.955	0.923	0.918	0.967
	Recall	0.945	0.742	0.955	0.985	0.750	0.967
	F-1 score	0.917	0.793	0.955	0.953	0.826	0.967
Case 2[(B)]	Precision	0.926	0.754	0.950	0.952	0.902	0.936
	Recall	0.890	0.860	0.864	0.963	0.842	0.978
	F-1 score	0.908	0.803	0.905	0.957	0.871	0.957
Case 3[(C)]	Precision	0.915	0.841	0.938	0.953	0.927	0.967
	Recall	0.929	0.803	0.955	0.985	0.842	0.967
	F-1 score	0.922	0.822	0.946	0.968	0.882	0.967
Case 4[(D)]	Precision	0.915	0.850	0.946	0.942	0.941	0.946
	Recall	0.937	0.798	0.964	0.991	0.792	0.967
	F-1 score	0.926	0.823	0.955	0.966	0.860	0.956

[(A)] Case 1 [1, 1, 1], [(B)] Case 2 [1, 10, 1], [(C)] Case 3 [2, 5, 1], and [(D)] Case 4 [2, 1, 1]. The performance measures per class considered were precision, recall, and F_1 score.

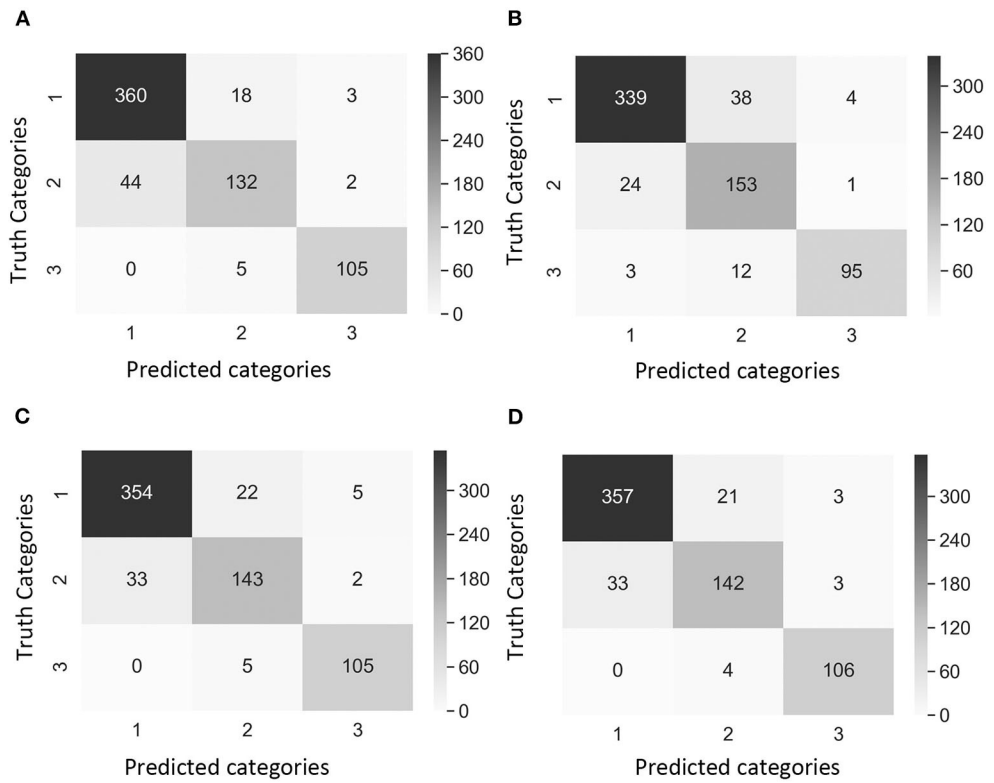

FIGURE 4 | Confusion matrix of the images of Dataset 1 (non-matured spikes only) showing "true" categories by Rater 1 (y-axis) and predicted categories by the CNN model (x-axis). Category 1: contained images with 0% severity, Category 2: 0.1–20% severity, Category 3: 20.1–100% severity. The cases of study presented different weights in the loss function [weight in Category 1, weight in Category 2, weight in Category 3]. **(A)** Case 1 [1, 1, 1], **(B)** Case 2 [1, 10, 1], **(C)** Case 3 [2, 5, 1], and **(D)** Case 4 [2, 1, 1]. Values and color intensity represent number of images.

Category 2 by the model (**Table 4**). The lowest F_1 score of Dataset 1 Category 2, was 79.3% in Case 1, while the highest was 82% in both Case 3 and Case 4 (**Table 4**). In Dataset 2 Category 2, the lowest F_1 score was 82.6% in Case 1, and the highest F_1 score was 88.2% in Case 3 followed by Case 2 with 87.1% (**Table 4**).

A comparison of outcomes revealed that Category 2 was the most difficult category to classify correctly (**Figure 4**). This difficulty was attributed to the disease symptoms being barely visible at the early stage of infection, and some wheat spikes in Category 1 were maturing, and their color was similar to

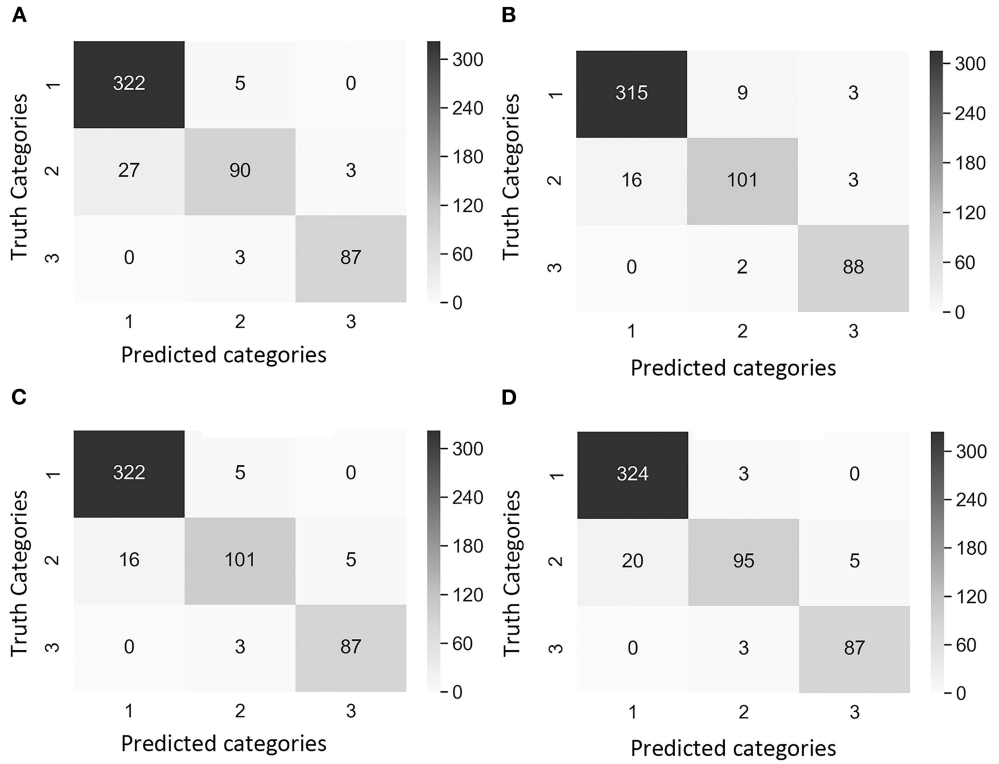

FIGURE 5 | Confusion matrix of the images of Dataset 2 (non-matured spikes only) showing "true" categories by Rater 1 (y-axis) and the predicted categories by the CNN model (x-axis). Category 1: contained images with 0% severity, Category 2: 0.1–20% severity, Category 3: 20.1–100% severity. The cases of study presented different weights in the loss function [weight in Category 1, weight in Category 2, and weight in Category 3]. **(A)** Case 1 [1, 1, 1], **(B)** Case 2 [1, 10, 1], **(C)** Case 3 [2, 5, 1], and **(D)** Case 4 [2, 1, 1]. Values and color intensity represent number of images.

that of MoT infected spikes. We observed that the highest number of images exactly classified as Category 2 was obtained with the Case 2 Dataset 1 (**Figure 4B**). These results suggested that Case 2 was the most appropriate to classify wheat spike blast images in Dataset 1 because it was capable of detecting the infection at an early stage. Even though Case 2 had a slightly lower precision, this is considered the usual trade-off between $precision_2$ and recall for disease classification purposes. The recall, $precision_2$, and F_1 score increased after the images of maturing spikes were omitted when training the model with Dataset 2 (**Figure 5**). The cases 2 and 3 of Dataset 2 presented the highest number of images exactly classified as Category 2 (**Figures 5B,C**). Cases 2 and 3 were the most appropriate to detect the wheat spike blast in Dataset 2 because the model was capable of detecting the infection in the early stages. Additionally, in all the cases, the model was more stable predicting Category 3, which is relevant because it covers DS from 20.1 to 100%, potentially aiding breeders and pathologists to discern higher levels of susceptibility among cultivars. Although the CNN model misclassified some images of Category 2, it still provided a promising approach to classify the severity of the disease. It demonstrated that the CNN model is potentially a good method for breeders and pathologists.

DISCUSSION

Wheat blast is spreading worldwide, the identification of durable and broad-spectrum resistance is urgently needed (Valent et al., 2021). There are a few known sources of effective resistance, and therefore it is crucial to identify more genetic resources. Plant disease phenotyping is a bottleneck in the identification of novel sources of resistance. We developed the first deep CNNs model for wheat spike blast phenotyping under controlled environment.

This study results demonstrated that the agreement between disease estimations and disease measurements was more significant than what could have been expected to occur by chance. Rater 1 (a pathologist with expertise in multiple diseases besides blast) consistently obtained the higher kappa coefficient (substantial agreement), higher accuracy, and lower bias in all the performed analyses than disease estimations of an expert (Rater 2) in the wheat blast and the disease measurements of ImageJ software. These results are relevant because Rater 1 estimated the DS and classified the entire image dataset into three categories. Therefore, the agreement analysis supports an accurate classification of the images before they were used to train and test the CNN model. The inter-rater agreement analysis also showed that accuracy, precision, and bias are highly dependent on the nature of the dataset. Dataset 1 included images

showing disease symptoms and natural plant physiological changes. However, although Dataset 2 was preferred due to higher concordance, results showed that DS assessments among raters were never perfect.

In the present study, the applicability of CNNs for wheat spike blast severity classification from spring wheat images was investigated. Currently, the CNN approach can classify three severity levels (0%, 0.1–20%, and 20.1–100% severity) and was trained using a reliable wheat spike blast dataset. The advantage of this three categories CNN model is that it detects the infected wheat spike and provides further information on the corresponding blast severity level. It is useful to have such a model to classify different infection levels and identify the resistant cultivars from the susceptible ones. Despite the wheat blast dataset comprising of imbalanced data that could have led to a biased CNN model, two techniques, including data augmentation and weighted loss function, were applied to the training process. The loss function is a function map of the difference between the ground truth and predicted output of the model. The importance of a category with a larger error can be enhanced by assigning a weighted variable in the loss function. The results indicate that the performance of the model has a significant improvement when the weighted loss function is applied. In particular, the model has gained the ability to detect Category 2 using a weighted loss function. These encouraging results demonstrate that the proposed CNN model can distinguish Category 1 and Category 2 even though there is a relatively little difference between both the categories. More significant, the CNN could classify the images of Category 3 with low error, which contained infected spikes with severities higher than 20%.

The results showed that the CNN models trained in both datasets (Fernandez-Campos et al., 2021) presented good performance classifying the wheat spike blast images in the corresponding severity categories. However, the models trained without images of wheat maturing spikes showed higher precision$_2$, recall, and F_1 score when classifying the images than the models trained with maturing and not matured wheat spikes. The performance of the model trained with maturing and non-matured spikes is a critical finding from a biological/physiological point of view. These symptoms on spikes are often reported when wheat has reached the medium milk-to-dough growth stage (Cruz et al., 2016b). The reason why the rating is often stopped at the milk-to-dough stage is that from that point forward, physiological maturity starts to kick in. Our findings will serve to provide future and explicit guidelines to potential users of the preferred model. Users will need to acknowledge the natural wheat maturity process (which alters spike color from green to yellow/white), which can confuse the CNN model. This statement applies when phenotyping for wheat blast or similar diseases with symptoms characterized by spike bleaching [e.g., *Fusarium graminearum* (Fusarium head blight)].

Different software based on image analysis are currently available to measure DS (Lamari, 2002; Vale et al., 2003). We used ImageJ, a free image-processing software, and manually thresholded images to measure wheat spike blast severity. de Melo et al. (2020) indicated the inevitable error when delineating the disease area with image analysis software (Bock et al., 2008). This is a challenge that future research needs to address when disease symptoms are not well-defined.

Researchers could benefit from the proposed approach promising for wheat spike blast severity measurements under controlled environmental conditions. Results are supported by a substantial agreement with "true" data obtained from Rater 1, compared against disease estimations of Rater 2, and disease measurements of ImageJ. In collaboration with data scientists, breeders could pre-select wheat cultivars under controlled environments by automatically analyzing and classifying images using the wheat spike blast CNN model preferably trained with Dataset 2. Next, the breeders can focus on the cultivars that fall into categories 1 and 2, which in general terms, are considered resistant or moderately resistant. This may reduce the high number of cultivars tested under field conditions, accelerating the cultivar screening process. A limitation of the study is that the CNN was trained to classify only images of wheat spike blast (spring wheat) under controlled conditions. Further research is required to improve the generalizability of the CNN model using a greater wheat spike blast dataset consisting of controlled and field images. In addition, the results in this study show an opportunity that could be applied similar to other pathogens.

The next step in this research is to validate the model with other images with a similar background and deploy it in a Web application. This future option might allow breeders and pathologists to submit their images and have the model classify them by categories automatically. As more images of various cultivars infected with different isolates can be added to the dataset, increasing symptom variability, a more refined and robust model can be developed. To our knowledge, this is the first study presenting a deep CNN model trained to detect and classify wheat spike blast symptoms. The model might help in the pre-screening of wheat cultivars against the blast fungus under controlled conditions in the future.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article and corresponding models are available at: https://purr.purdue.edu/projects/wheatspikeblastcnn/publications/3772.

AUTHOR CONTRIBUTIONS

MF-C, CDC, MJ, Y-TH, TW, and JJ contributed to the study's conception and design. MF-C and CDC conducted the experiment. MF-C collected data and wrote the first draft of the manuscript. MF-C and CG-C performed the statistical analysis. Y-TH and TW wrote the code for the model. Y-TH wrote sections of the manuscript. CDC, MJ, Y-TH, DT, and CG-C edited the manuscript. All authors approved the submitted version.

FUNDING

This work was supported by the USDA National Institute of Food and Agriculture Hatch Project accession number 1016253. We thank Purdue University start-up funds for supporting this work. This work was partially supported

by the Borlaug Fellowship Program (Grant nos. FX18BF-10777R014 and FX19BF-10777R001) from the USDA Foreign Agricultural Service.

ACKNOWLEDGMENTS

We thank the Asociación de Productores de Oleaginosas y Trigo (ANAPO) and the Centro de Investigación Agrícola Tropical (CIAT) for the support provided with the experiments conducted in Bolivia and M. G. Rivadeneira from CIAT for the help with inoculum preparation. We acknowledge the Iyer-Pascuzzi Lab and A. P. Cruz for their guidance in plant phenotyping and D. F. Baldelomar, L. Calderón, J. Cuellar, F. Cortéz, and D. Coimbra from ANAPO for their help with research activities. We also thank Dr. Barbara Valent (Kansas State University) and Gary Peterson for their contribution and commitment to the wheat blast work in South America. Borlaug Fellows Drs. Mr. Kabir and S. Das participated and were trained while conducting experiments associated with this work.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2021.673505/full#supplementary-material

REFERENCES

Aman, A. (2016). Wheat blast' threatens yield-farmers in 6 districts complain of infection. *Dailystar*. Available online at: https://www.thedailystar.net/backpage/wheat-blast-threatens-yield-784372 (accessed June 6, 2020).

Atha, D. J., and Jahanshahi, M. R. (2018). Evaluation of deep learning approaches based on convolutional neural networks for corrosion detection. *Struct. Health Monit.* 17, 1110–1128. doi: 10.1177/1475921717737051

Baldelomar, D., Cardenas, S., and Quispe, K. (2015). *Caracterización de genotipos de trigo (Triticum aestivum) con resistencia a piricularia durante el verano 2014/2015 en Quirusillas*, Vol. 1. Santa Cruz: Revista científica de investigación INFO-INIAF, 17–21.

Barbedo, J. G. A. (2016). A review on the main challenges in automatic plant disease identification based on visible range images. *Biosyst. Eng.* 144, 52–60. doi: 10.1016/j.biosystemseng.2016.01.017

Barbedo, J. G. A. (2019). Plant disease identification from individual lesions and spots using deep learning. *Biosyst. Eng.* 180, 96–107. doi: 10.1016/j.biosystemseng.2019.02.002

Barea, G., and Toledo, J. (1996). *Identificación y zonificación de pyricularia (Pyricularia oryzae) en el cultivo de trigo en el departamento de Santa Cruz*, Informe Técnico. Proyecto de Investigación Trigo, 76–86.

Bhattacharya, R., and Pal, S. (2017). Deadly wheat blast symptoms enters India through the Bangladesh border, Bengal govt burning crops on war footing. *Hindustan Times*. Kolkata. Available online at: http://www.hindustantimes.com/kolkata/deadly-wheat-blast-symptoms-enters-india-through-the-bangladesh-border-bengal-govt-burning-crops-on-war-footing/story-3zoWQ0H7sdMU4HxQyzWUsN.html (accessed August 2020).

Bock, C. H., Barbedo, J. G. A., Del Ponte, E. M., Bohnenkamp, D., and Mahlein, A. K. (2020). From visual estimates to fully automated sensor-based measurements of plant disease severity: status and challenges for improving accuracy. *Phytopathol. Res.* 2, 1–30. doi: 10.1186/s42483-020-00049-8

Bock, C. H., Parker, P. E., Cook, A. Z., and Gottwald, T. R. (2008). Visual rating and the use of image analysis for assessing different symptoms of citrus canker on grapefruit leaves. *Plant Dis.* 92, 530–541. doi: 10.1094/PDIS-92-4-0530

Bock, C. H., Poole, G. H., Parker, P. E., and Gottwald, T. R. (2010). Plant disease severity estimated visually, by digital photography and image analysis, and by hyperspectral imaging. *Crit. Rev. Plant Sci.* 29, 59–107. doi: 10.1080/07352681003617285

Bottou, L. (2010). "Large-scale machine learning with stochastic gradient descent," in *Proceedings of compstat,* eds Y. Lechevallier and G. Saporta (Heidelberg: Physica-Verlag HD), 177–186. doi: 10.1007/978-3-7908-2604-3_16

Boulent, J., Foucher, S., Théau, J., and St-Charles, P. L. (2019). Convolutional neural networks for the automatic identification of plant diseases. *Front. Plant Sci.* 10:941. doi: 10.3389/fpls.2019.00941

Brahimi, M., Arsenovic, M., Laraba, S., Sladojevic, S., Boukhalfa, K., and Moussaoui, A. (2018). "Deep learning for plant diseases: detection and saliency map visualization," in *Human Machine Learning*, eds J. Zhou, and F. Chen (Cham: Springer), 93–117. doi: 10.1007/978-3-319-90403-0_6

Cabrera, M. G., and Gutiérrez, S. (2007). "Primer registro de Pyricularia grisea en cultivos de trigo del NE de Argentina," in *Jornada de Actualización en Enfermedades de Trigo*, Vol. 60 (Buenos Aires: IFSC Press).

Camargo, A., and Smith, J. S. (2009). Image pattern classification for the identification of disease causing agents in plants. *Comput. Electron. Agr.* 66, 121–125. doi: 10.1016/j.compag.2009.01.003

Cardoso, C. A. A., Reis, E. M., and Moreira, E. N. (2008). Development of a warning system for wheat blast caused by *Pyricularia grisea*. *Summa Phytopathol.* 34, 216–221. doi: 10.1590/S0100-54052008000300002

Ceresini, P., Castroagudín, V. L., Rodrigues, A. F., Rios, J., Aucique-Pérez, C., Moreira, S. I., et al. (2019). Wheat blast: from its origins in South America to its emergence as a global threat. *Phytopathology* 104, 95–107. doi: 10.1111/mpp.12747

Chen, F., and Jahanshahi, M. R. (2018). Deep learning-based crack detection using convolutional neural network and naïve bayes data fusion. *IEEE Trans. Ind. Electron.* 65, 4392–4400. doi: 10.1109/TIE.2017.2764844

Chiang, K. S., Bock, C. H., El Jarroudi, M., Delfosse, P., Lee, I. H., Liu, H. I. (2016). Effects of rater bias and assessment method on disease severity estimation with regard to hypothesis testing. *Plant Pathol.* 65, 523–535. doi: 10.1111/ppa.12435

Chmura, H. (1992). Measurement of reliability for categorical data in medical research. *Stat. Methods Med. Res.* 1, 183–199. doi: 10.1177/096228029200100204

Cruppe, G., Cruz, C., Peterson, G., Pedley, K., Mohammad, A., Fritz, A., et al. (2019). Novel sources of wheat head blast resistance in modern breeding lines and wheat wild relatives. *Plant Dis.* 104, 35–43. doi: 10.1094/PDIS-05-19-0985-RE

Cruppe, G., Silva, P., da Silva, C. L., Peterson, G., Pedley, K. F., Cruz, C. D., et al. (2020). Genome wide association reveals limited benefits of pyramiding the 1B and 1D loci with the 2NvS translocation for wheat blast control. *Crop Sci.* 61, 1–15. doi: 10.1002/csc2.20397

Cruz, C. D., Bockus, W., Stack, J., Valent, B., Maciel, J. N., and Peterson, G. L. (2016a). A standardized inoculation protocol to test wheat cultivars for reaction to head blast caused by *Magnaporthe oryzae (Triticum* pathotype). *Plant Health Prog.* 17, 186–187. doi: 10.1094/PHP-BR-16-0041

Cruz, C. D., Kiyuna, J., Bockus, W., Todd, T. C., Stack, J. P., and Valent, B. (2015). *Magnaporthe oryzae* conidia on basal wheat leaves as a potential source of wheat blast inoculum. *Plant Pathol.* 64, 1491–1498. doi: 10.1111/ppa.12414

Cruz, C. D., Peterson, G. L., Bockus, W. W., Kankanala, P., Dubcovsky, J., Jordan, K. W., et al. (2016b). The $2N^{v}S$ translocation from *Aegilops ventricosa* confers resistance to the *triticum* pathotype of *Magnaporthe oryzae*. *Crop Sci.* 56, 990–1000. doi: 10.2135/cropsci2015.07.0410

Cruz, C. D., and Valent, B. (2017). Wheat blast disease: danger on the move. *Trop. Plant Pathol.* 42, 210–222. doi: 10.1007/s40858-017-0159-z

de Melo, V. P., da Silva Mendoca, A. C., de Souza, H. S., Gabriel, L. C., Bock, C. H., Eaton, M. J., et al. (2020). Reproducibility of the development and validation process of standard area diagram by two laboratories: an example using the Botrytis cinerea/Gerbera jamesonii pathosystem. *Plant Dis.* 104, 2440–2448. doi: 10.1094/PDIS-08-19-1708-RE

Everitt, B., and Anders, S. (2010). *The Cambridge Dictionary of Statistics*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511779633

Fernández-Campos, M., Góngora-Canul, C., Das, S., Kabir, M., Valent, B., and Cruz, C. (2020). Epidemiological criteria to support breeding tactics against the emerging, high-consequence wheat blast disease. *Plant Dis*. 104, 1–38. doi: 10.1094/PDIS-12-19-2672-RE

Fernandez-Campos, M. S., Huang, Y., Jahanshahi, M., Wang, T., Jin, J., Telenko, D., et al. (2021). Wheat spike blast image classification using deep convolutional neural networks. *Purdue University Research Repository*. West Lafayette doi: 10.4231/P0Y7-3428

Fiorani, F., and Schurr, U. (2013). Future scenarios for plant phenotyping. *Annu. Rev. Plant Biol*. 64, 267–291. doi: 10.1146/annurev-arplant-050312-120137

Fleiss, J. L., Levin, B., and Paik, M. C. (2003). *Statistical Methods for Rates and Proportions*. John Wiley and Sons, Inc.

Gamer, M., Lemon, J., Fellows, I., and Singh, P. (2019). *Package 'irr' Title Various Coefficients of Interrater Reliability and Agreement*. Avaialble online at: https://CRAN.R-project.org/package=irr

Ghosal, S., Blystone, D., Singh, A. K., Ganapathysubramanian, B., Singh, A., and Sarkar, S. (2018). An explainable deep machine vision framework for plant stress phenotyping. *Proc. Natl. Acad. Sci. U.S.A*. 115, 4613–4618. doi: 10.1073/pnas.1716999115

Gongora-Canul, C., Salgado, J., Singh, D., Cruz, A., Cotrozzi, L., Couture, J. J., et al. (2020). Temporal dynamics of wheat blast epidemics and disease measurements using multispectral imagery. *Phytopathology* 110, 393–405. doi: 10.1094/PHYTO-08-19-0297-R

Goulart, A., and Paiva, F. (1992). Incidência da brusone (Pyricularia oryzae) em diferentes cultivares de trigo (Triticum aestivum) em condições de campo. *Fitopatol bras*. 17, 321–325.

Goulart, A. C. P., Sousa, P. G., and Urashima, A.S. (2007). Damages in wheat caused by infection of Pyricularia grisea. *Summa Phytopath*. 33, 358–63. doi: 10.1590/S0100-54052007000400007

Graham, P., and Jackson, R. (1993). The analysis of ordinal agreement data: beyond weighted kappa. *J. Clin. Epidemiol*. 46, 1055–1062. doi: 10.1016/0895-4356(93)90173-X

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition. *arXiv* [Preprint] 495arXiv:1512.03385.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016* (IEEE), 770–778. doi: 10.1109/CVPR.2016.90

Igarashi, S. (1986). "Brusone no trigo- histórico e distribuição geográfica no Paraná," in *Reunião nacional de pesquisa de trigo* (Passo Fundo), 15.

Igarashi, S., Utiamada, C. M., Kazuma, I. L. C., H., A., and Lopes, R. S. (1986). Pyricularia em trigo. 1. Ocorrencia de *Pyricularia sp*. no estado do Parana. *Fitopatol. Bras*. 11, 351–352.

Juliana, P., He, X., Kabir, M. R., Roy, K. K., Anwar, M. B., Marza, F., et al. (2020). Genome-wide association mapping for wheat blast resistance in CIMMYT's international screening nurseries evaluated in Bolivia and Bangladesh. *Sci Rep*. 10:15972. doi: 10.1038/s41598-020-72735-8

Kersting, K., Bauckhage, C., Wahabzada, M., Mahlein, A. K., Steiner, U., Oerke, E. C., et al. (2016). "Feeding the world with big data: uncovering spectral characteristics and dynamics of stressed plants," in *Computational Sustainability*, Vol. 645 eds J. Lässig, K. Kersting, and K. Morik (Cham: Springer), 99–120. doi: 10.1007/978-3-319-31858-5_6

Kruse, O. M., O., Prats-Montalbán, J. M., Indahl, U. G., Kvaal, K., Ferrer, A., et al. (2014). Pixel classification methods for identifying and quantifying leaf surface injury from digital images. *Comput. Electron. Agr*. 108, 155–165. doi: 10.1016/j.compag.2014.07.010

Kumar, S., Kashyap, P. L., Mahapatra, S., Jasrotia, P., and Singh, G. P. (2020). New and emerging technologies for detecting *Magnaporthe oryzae* causing blast disease in crop plants. *Crop Protect*. 143:105473. doi: 10.1016/j.cropro.2020.105473

Kumar, S. S., Abraham, D. M., Jahanshahi, M. R., Iseley, T., and Starr, J. (2018). Automated defect classification in sewer closed circuit television inspections using deep convolutional neural networks. *Autom. Construc*. 91, 273–283. doi: 10.1016/j.autcon.2018.03.028

Lamari, L. (2002). *ASSESS: Image Analysis Software for Plant Disease Quantification*. St. Paul, MN: APS Press.

Large, E. C. (1954). Growth stages in cereals illustration of the Feeks scale. *Plant Pathol*. 3, 128–129. doi: 10.1111/j.1365-3059.1954.tb00716.x

Large, E. C. (1966). Measuring plant disease. *Ann. Rev. Phytopathol*. 4, 9–28. doi: 10.1146/annurev.py.04.090166.000301

Lin, L., Hedayat, A. S., Sinha, B., and Yang, M. (2002). Statistical methods in assessing agreement. *J. Am. Stat. Assoc*. 97, 257–270. doi: 10.1198/016214502753479392

Lin, L. I. K. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 45, 255–268. doi: 10.2307/2532051

Lobet, G. (2017). Image analysis in plant sciences: publish then perish. *Trends Plant Sci*. 22, 559–566. doi: 10.1016/j.tplants.2017.05.002

Lu, J., Hu, J., Zhao, G., Mei, F., and Zhang, C. (2017). An in-field automatic wheat disease diagnosis system. *Compt. Electron. Agr*. 142, 369–379. doi: 10.1016/j.compag.2017.09.012

Madden, L. V., Hughes, G., and van den Bosch, F. (2007). *The Study of Plant Disease Epidemics*. St. Paul, MN: APS Press.

Mahlein, A. K. (2015). plant disease detection by imaging sensors – parallels and specific demands for precision agriculture and plant phenotyping. *Plant Dis*. 100, 241–251. doi: 10.1094/PDIS-03-15-0340-FE

Mahlein, A. K., Kuska, M. T., Behmann, J., Polder, G., and Walter, A., (2018). Hyperspectral sensors and imaging technologies in phytopathology: state of the art. *Ann. Rev. Phytopathol*. 56, 535–558. doi: 10.1146/annurev-phyto-080417-050100

Malaker, P. K., Barma, N. C. D., Tiwari, T. P., Collis, W. J., Duveiller, E., Singh, P. K., et al. (2016). First report of wheat blast caused by Magnaporthe oryzae pathotype triticum in Bangladesh. *Plant Dis*. 100, 2330–2330. doi: 10.1094/PDIS-05-16-0666-PDN

Mitani, A. A., Freer, P. E., and Nelson, K. P. (2017). Summary measures of agreement and association between many raters' ordinal classifications. *Ann. Epidemiol*. 27, 677–685.e4. doi: 10.1016/j.annepidem.2017.09.001

Mohanty, S. P., Hughes, D. P., and Salathé, M. (2016). Using deep learning for image-based plant disease detection. *Front. Plant Sci*. 7, 14–19. doi: 10.3389/fpls.2016.01419

Mottaleb, K. A., Singh, P. K., Sonder, K., Kruseman, G., Tiwari, T. P., Barma, N. C. D., et al. (2018). Threat of wheat blast to South Asia's food security: an ex-ante analysis. *PLoS ONE*. 13:e0197555. doi: 10.1371/journal.pone.0197555

Nelson, K., and Edwards, D. (2015). Measures of agreement between many raters for ordinal classifications. *Stat. Med*. 34, 3116–3132. doi: 10.1002/sim.6546

Nita, M., Ellis, M. A., and Madden, L. V. (2003). Reliability and accuracy of visual estimation of Phomopsis leaf blight of strawberry. *Phytopathology* 93, 995–1005. doi: 10.1094/PHYTO.2003.93.8.995

Nutter, F. W., Gleason, M., Jenco, J., and Christians, N. (1993). Assessing the accuracy, intra-rater repeatability, and inter-rater reliability of disease assessment systems. *Phytopathology* 83, 806–812. doi: 10.1094/Phyto-83-806

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., et al. (2017). *Automatic Differentiation in PyTorch*. Nips autodiff workshop.

Perello, A., Martinez I., and Molina, M. (2015). First report of virulence and effects of Magnaporthe oryzae isolates causing wheat blast in Argentina. *Plant Dis*. 99, 1177–1178. doi: 10.1094/PDIS-11-14-1182-PDN

Prestes, A. M., Arendt, P. F., Fernandes, J. M. C., and Scheeren, P. L. (2007). Resistance to *Magnaporthe grisea* among Brazilian wheat genotypes. *Wheat Prod. Stress. Environ*. 16, 119–123. doi: 10.1007/1-4020-5497-1_16

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., et al. (2012). Fiji: an open-source platform for biological-image analysis. *Nat. Methods*. 9, 676–682. doi: 10.1038/nmeth.2019

Shakoor, N., Lee, S., and Mockler, T. C. (2017). High throughput phenotyping to accelerate crop breeding and monitoring of diseases in the field. *Curr. Opin. Plant Biol*. 38, 184–192. doi: 10.1016/j.pbi.2017.05.006

Sibiya, M., and Sumbwanyambe, M. (2019). An algorithm for severity estimation of plant leaf diseases by the use of colour threshold image segmentation and fuzzy logic inference: a proposed algorithm to update a "leaf doctor" application. *Agri. Eng*. 1, 205–219. doi: 10.3390/agriengineering1020015

Singh, A., Ganapathysubramanian, B., Sarkar, S., and Singh, A. (2018). Deep learning for plant stress phenotyping: trends and future perspectives. *Trends Plant Sci*. 23, 883–898. doi: 10.1016/j.tplants.2018.07.004

Singh, A., Jones, S., Ganapathysubramanian, S., Mueller, D., Sandhu, K., and Nagasubramanian, K. (2021). Challenges and opportunities in machine-augmented plant stress phenotyping. *Trends Plant Sc*. 26, 53–69. doi: 10.1016/j.tplants.2020.07.010

Tang, W., Hu, J., Zhang, H., Wu, P., and He, H. (2015). Kappa coefficient: a popular measure of rater agreement. *Shanghai Arch. Psychiatry*. 27, 62–67. doi: 10.11919/j.issn.1002-0829.215010

Team, R. C. (2017). *R: A Language and Environment for Statistical Computing (Computer Software Manual)*. Vienna.

Tembo, B., Mulenga, R. M., Sichilima, S., M'siska, K. K., Mwale, M., Chikoti, P. C., et al. (2020). Detection and characterization of fungus (*Magnaporthe oryzae* pathotype *Triticum*) causing wheat blast disease on rain-fed grown wheat (*Triticum aestivum L.*) in Zambia. *PLoS ONE* 15:e0238724. doi: 10.1371/journal.pone.0238724

Urashima, A., Lavorent, N. A., Goulart, A. C. P., and Mehta, Y. R. (2004). Resistance spectra of wheat cultivars and virulence diversity of *Magnaporthe grisea* isolates in brazil. *Fitopatol. Bras*. 29, 511–518. doi: 10.1590/S0100-41582004000500007

Vale, F. X. R., Fernandes-Filho, E. I., and Liberato, J. R. (2003). "QUANT: a software for plant disease severity assessment," in *Proceedings of the Eighth International Congress of Plant Pathology* (Christchurch), 105.

Valent, B. (2016). *Novel Strategies for Managing Blast Diseases on Rice and Wheat*. Progress report 01/01/15 to 12/31/15. Available online at: https://portal.nifa.usda.gov/web/crisprojectpages/0231543-novel-strategies-for-managing-blast-diseases-on-rice-and-wheat.html (accessed August 2020).

Valent, B., Cruppe, G., Stack, J. P., Cruz, C. D., Farman, M. L., Paul, P. A., et al. (2021). *Recovery Plan for Wheat Blast Caused by Magnaporthe Oryzae Pathotype Triticum*. Plant Health Prog.

Vales, M., Anzoátegui, T., Huallpa, B., and Cazon, M. I. (2018). Review on resistance to wheat blast disease (*Magnaporthe oryzae Triticum*) from the breeder point-of-view: use of the experience on resistance to rice blast disease. *Euphytica* 214, 1–19. doi: 10.1007/s10681-017-2087-x

Viedma, L. (2005). Wheat blast occurrence in Paraguay. *Phytopathology* 95:152.

Wang, S., Asuke, S., Vy, T.T.P., Inoue, Y. Chuma, I. Win, J., et al. (2018). A new resistance gene in combination with Rmg8 confers strong resistance against triticum isolates of Pyricularia oryzae in a common wheat landrace. *Phytopathology*. 108, 1299–1306. doi: 10.1094/PHYTO-12-17-0400-R

Wise, K., Shaner, G., and Mansfield, C. (2011). *Purdue Extension Managing Wheat by Growth Stage*. Retrieved from: www.the-education-store.com

Wu, R. T., and Jahanshahi, M. R. (2019). Deep convolutional neural network for structural dynamic response estimation and system identification. *J. Eng. Mech*. 145:04018125. doi: 10.1061/(ASCE)EM.1943-7889.0001556

Yang, W., Feng, H., Zhang, X., Zhang, J., Doonan, J. H., Batchelor, W. D., et al. (2020). Crop phenomics and high-throughput phenotyping: past decades, current challenges, and future perspectives. *Mol. Plant*. 13, 187–214. doi: 10.1016/j.molp.2020.01.008

Yasuhara-Bell, J., Pedley, K. F., Farman, M., Valent, B., and Stack, J. P. (2018). Specific detection of the wheat blast pathogen (*Magnaporthe oryzae Triticum*) by loop-mediated isothermal amplification. *Plant Dis*. 102, 2550–2559. doi: 10.1094/PDIS-03-18-0512-RE

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.