Received: 24 April 2023 | Accepted: 28 September 2023

DOI: 10.1002/ppj2.20087

The Plant Phenome Journal OPEN ACCESS

ORIGINAL ARTICLE

Phenotyping and predicting wheat spike characteristics using image analysis and machine learning

Mik Hammers[1] | **Zachary J. Winn**[1] | **Asa Ben-Hur**[2] | **Dylan Larkin**[3] | **Jamison Murry**[4] | **Richard Esten Mason**[1]

[1]Department of Soil and Crop Sciences, Colorado State University, Fort Collins, Colorado, USA

[2]Department of Computer Science, Colorado State University, Fort Collins, Colorado, USA

[3]Limagrain Cereal Seeds, Walla Walla, Washington, USA

[4]United States Department of Agriculture Natural Resource Conservation Service, Lonoke, Arkansas, USA

Correspondence
Mik Hammers and Richard Esten Mason, Department of Soil and Crop Sciences, Colorado State University, Fort Collins, CO, USA.
Email: hammers@colostate.edu and esten.mason@colostate.edu

Assigned to Associate Editor Michael Gore.

Funding information
Agriculture and Food Research Initiative of the National Institute of Food and Agriculture (WheatCAP), Grant/Award Numbers: 2022-68013-36439, 2017-67007-25939

Abstract

Improvements in trait phenotyping are needed to increase the quantity and quality of data available for genetic improvement of crops. In this study, we used moderate throughput image analysis and machine learning as a pipeline for phenotyping a key wheat spike characteristic: spikelet number per spike. A population of 594 soft red winter wheat inbred lines was evaluated in the field for 2 years and images of wheat spikes were taken and used to train deep-learning algorithms to predict spikelet number. A total of 12,717 images were used to train, test, and validate a basic regression convolutional neural network (CNN), a visual geometry group application regression model, VGG16, the ResNet152V2 model, and the EfficientNetV2L model. The EfficientNetV2L model was the most accurate, having the lowest mean absolute error, second lowest root mean square error, and highest coefficient of determination (mean absolute error [MAE] = 0.60, root mean square error [RMSE] = 0.79, and R^2 = 0.90). The ResNet152V2 model was slightly less accurate with a slightly better fit (MAE = 0.61,m RMSE = 0.78, and R^2 = 0.87), followed by the basic CNN (MAE = 0.75, RMSE = 1.00, and R^2 = 0.74) and finally by the VGG16 (MAE = 1.51, RMSE = 1.29, and R^2 = 0.076). With an average error of just above one half of a spikelet, utilizing image analysis and machine learning counting methods could be used for multiple breeding applications, including direct selection of spikelet number, to provide data to identify quantitative trait loci, or for training whole genome selection models.

1 | INTRODUCTION

Common wheat (*Triticum aestivum* L.) is one of the most widely used and versatile crops, with annual production exceeding 700 million tons globally (FAO, 2022). Wheat has a wide cultivation range (Shewry, 2009) due to its capacity for adaptation (Blake et al., 2009), particularly within the flowering pathway. The two largest components of the flowering pathway are the photoperiod response (*Ppd*), which contributes to 20%–25% of the variation in heading time, and the vernalization response (*Vrn*), which contributes to 70%–75% of the variation in heading time (Stelmakh, 1998). The

Abbreviations: CNN, convolutional neural network; HGAWN, Historical Gulf Atlantic Wheat Nursery; HTP, high-throughput phenotyping; MAE, mean absolute error; ML, machine learning; RMSE, root mean square error; SPS, spikelets per spike; SRWW, soft red winter wheat; UAV, unoccupied ariel vehicle; VGG, visual geometry group.

remaining 5% is determined by the earliness per se (*Eps*) gene system, which is not yet entirely understood.

There is evidence supporting the influence of heading time on spike architecture traits such as spikelets per spike (SPS) and spikelet differentiation (Chen et al., 2020; Miura & Worland, 1994). Spikelets are groups of three florets that move in an alternating pattern on rachis nodes along the length of the spike (Koppolu & Schnurbusch, 2019). Spikelets are the reproductive part of the spike, and SPS is positively correlated with grain yield, making it an important trait for increasing total grain yield (Chen et al., 2020). Understanding the genetic components underlying morphological traits, such as SPS, is difficult due to the tedious nature of phenotypic data collection. A pipeline to accurately and efficiently determine SPS, as well as other spike characteristics, would improve our understanding of these traits and aid in their genetic improvement. Phenotypes, such as SPS, are the resulting characteristics of a genotype and genotype by environment interaction. Accurate phenotyping is vital to genetic improvement through breeding and for discovery of marker-trait associations, genome-wide association studies, marker-assisted selection, and genomic selection (Furbank & Tester, 2011; Minervini et al., 2015). While advancements in technology for generating high-density genotypic data have increased the efficiency of genomic analysis (Koboldt et al., 2013), phenotyping methods are still lagging (Houle et al., 2010; Jin et al., 2021; Song et al., 2021). Recently, there has been increased emphasis on the development of high-throughput phenotyping (HTP) pipelines to increase the rate at which we can measure traits such as grain yield. Multiple nondestructive HTP approaches are now being evaluated and deployed in plant science research, including sensors (Milella et al., 2019), occupied or unoccupied ariel vehicles (UAVs) (Krause et al., 2020; Rutkoski et al., 2016), and imaging (Fitzgibbon et al., 2013; Yang et al., 2014).

Imaging in a controlled setting (as opposed to in field) allows for flexibility in terms of equipment cost, timing, and resolution. In general, these images are highly reproducible and can be utilized across a variety of applications ranging from microscopic to macroscopic subjects (Furbank & Tester, 2011; Minervini et al., 2015), capturing a single maturity point in time, or creating time-lapse data throughout a life cycle (Fahlgren et al., 2015; Gehan & Kellogg, 2017). While collection of the image data is important, an efficient analysis pipeline is also critical to maximize output and not just input of data.

Machine learning (ML) is a computerized model that can learn patterns from data and make decisions (Singh et al., 2016). While multiple types of models exist for ML, the main idea behind computer vision is the detection of similarities and dissimilarities between the provided images. ML has been used in conjunction with HTP in plant breeding programs due to its ability to automate the analysis process and increase efficiency (Gehan & Kellogg, 2017; Tsaftaris et al., 2016).

Core Ideas

- High-throughput phenotyping methods are needed to make larger quantities of phenotypic data accessible to researchers and plant breeders.
- Imaging is valuable for high-throughput phenotyping and can be used for collecting quantitative plant traits.
- Machine learning algorithms can be used as an efficient analysis pipeline to create high-throughput phenotyping methods for collecting trait data from wheat spikes.

ML models can be trained using both supervised methods, with labels for images during the time of training, or unsupervised methods, which allows the model to decide the most meaningful features of an image (Singh et al., 2016). For training, the input dataset is partitioned into three parts, including (1) a training set where the model will learn how to identify relevant aspects based on its task, (2) a validation set for testing the accuracy of the model, and (3) a test set of data that the model has not seen before. The validation set allows for fine-tuning of the model hyperparameters to increase accuracy to the desired level and then used to analyze the test set, which the model did not utilize during training or validation (Singh et al., 2016). For supervised learning, training sets labeled with the associated information are an input to the model so that it can learn the associations between the provided image and label.

Deep learning is a type of ML that does not rely on "feature engineering" (LeCun et al., 2015). Instead, deep learning models have layers of nonlinearity that transform data into a representation that makes it easy to classify. This structure allows deep learning models to excel with larger datasets and more complex data analytics and is becoming more common in plant phenotyping and image analyses (LeCun et al., 2015; Ubbens & Stavness, 2017). Convolutional neural networks (CNNs) are a class of deep learning methods whose primary components are convolutional and pooling layers (LeCun et al., 2015). Convolutional layers scan an image with a collection of small matrices called "kernels" or "filters" to find locations in the image that match the patterns learned by each filter. Following a convolutional layer, we typically employ a pooling layer that computes summary statistics of the output of its preceding convolutional layer. This serves to downsample the data and build locational invariance into the learned model (Montesinos López et al., 2022). These processing

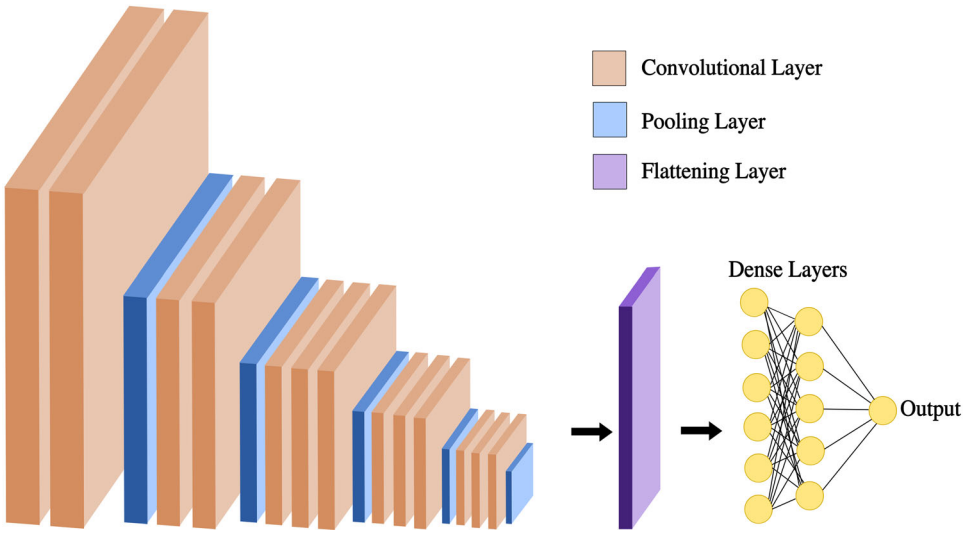

FIGURE 1 Example structure of convolutional, pooling, flattening, and dense layers in a convolutional neural network (CNN). The above example is the architecture of a VGG16 model. In CNNs, convolutional layers and pooling layers, which are shown in orange and blue, respectively, are alternated to extract features from an input image. The outputs from the final layer of feature extracted are then passed on to the flattening layer, shown in purple. Finally, the outputs from the flattening layer are passed on to the dense layers, which will compute the predicted spikelet count.

layers are able to extract various features from an input image, assign importance to them, and differentiate them. The final layer of a CNN is the fully connected layer, which takes the output from the processing layers and reshapes them back into a single column for processing based on the type of model being used, such as classification or regression (Ubbens & Stavness, 2017). An example of a CNN architecture with each of these different layers is depicted in Figure 1.

Convolutional neural networks have the ability to find edges and motifs in images that improve their ability for image analysis for certain plant phenotyping tasks (Tsaftaris et al., 2016) such as counting (Khaki et al., 2020; Ubbens & Stavness, 2017). Deep learning can be used for several image-based tasks including segmentation, regression, classification, and detection (LeCun et al., 2015). Regression can be used for counting tasks over other methods, such as classification, to have a better understanding of how close the predicted values are to the true values of an image, providing an error rate that can help assess model accuracy. Some CNN models are known for their use in vision problems and are readily available. Some well-known architectures include VGG16, ResNet models, and EfficientNet models. Many studies utilize these complex models in lieu of generating a new model from scratch (Kanchanadevi & Sandhia, 2023; Khaki et al., 2020; Nigam et al., 2023; Rao et al., 2022).

The objectives of this study were to; (1) evaluate soft red winter wheat (SRWW) genotypes using imaging to determine SPS and (2) develop deep learning models for a high-throughput analysis pipeline of wheat spike images. These results will aid in further genetic analysis and improvement of wheat spike characteristics and a pipeline of imaging and image analysis that could be adapted to other crop species.

2 | MATERIALS AND METHODS

2.1 | Plant materials

The genetic material used in this study was the Historical Gulf Atlantic Wheat Nursery (HGAWN), a population consisting of 594 SRWW lines from public breeding institutions located in the southeastern United States. The HGAWN includes varieties from the University of Arkansas ($n = 103$), Louisiana State University ($n = 109$), University of Georgia ($n = 105$), North Carolina State University ($n = 104$), Texas A&M University ($n = 60$), Virginia Polytechnic Institute ($n = 44$), Clemson University ($n = 19$), and the United States Department of Agriculture Agricultural Research Service (USDA-ARS, $n = 9$).

2.2 | Experimental design

The HGAWN was evaluated during the 2019 and 2020 growing seasons at the Milo J. Shult Agricultural Research & Extension Center in Fayetteville, AR. Plots were drill seeded at a rate of 250 seed m^{-2} in a randomized complete block design with two replications. Each plot consisted of a single 1.20-m long row with 0.38 m horizontally between rows and 0.60 vertically between each range of plots. During both seasons, pre-plant recommendations were followed for

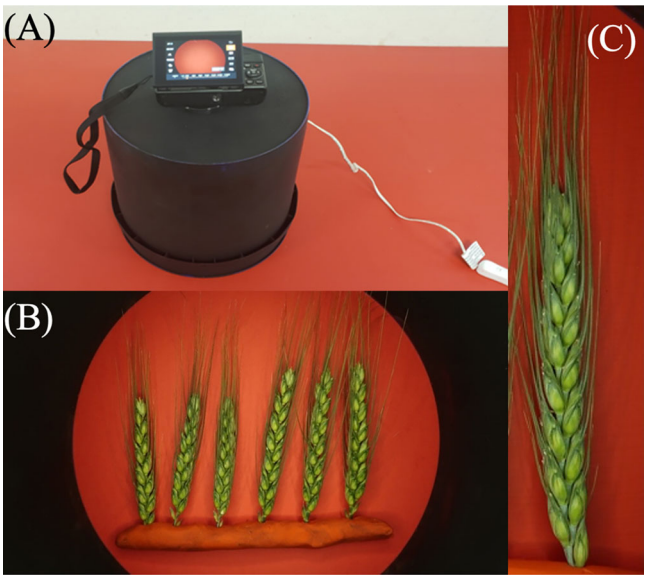

FIGURE 2 A Phenotyping device allowed for a controlled environment and consistency for photos of wheat spikes. (A) A camera rests on top of the device, which was used on the same red background for each photo. The black interior of the cover and attached light strip create uniform lighting and minimal reflection from camera flash while taking pictures. (B) Controlled environment images of six wheat spikes from each plot. Each spike was inserted into clay by the peduncle, rachis side up. Spikelets were fully visible for each spike to allow for phenotyping. (C) Example input image for specified deep learning algorithms. Each image of six spikes was cropped down to images of individual spikes before having their size reduced and padding the images to ensure identical sizing.

phosphorus and potassium and 100 kg ha^{-1} of nitrogen in the form of urea was applied in a split application in the spring (February and March). Harmony Extra (0.28 kg ha^{-1}) and Axial (0.6 kg ha^{-1}) were applied for control of annual ryegrass (*Lolium multiflorum* L.) and other weed species.

2.3 | Imaging

An imaging device was used to achieve consistent lighting and camera height across wheat spike image capture as described by Winn et al. (2021) (Figure 2A). The imaging device was designed for a Canon Powershot G1 X Mark II camera. A 4.5 cm circular hole was drilled into the bottom of a 0.46-L bucket using a 4.5-cm drill bit on an electric drill. The hole was drilled into the center of the bottom of the bucket with the mouth of the bucket facing downward. After drilling, the edges were filed down to protect the camera lens from scratches. A 2-cm hole was then drilled into the side of the bucket approximately 10 cm above the base of the mouth while the mouth of the bucket faced downward.

The inside of the bucket was painted by spraying with RUST-OLEUM Camouflage Ultra Flat Black so that the interior was fully coated. The paint was cured for 5 min and was coated until exterior light was no longer visible through the walls of the bucket. After the final coat of paint had cured, tape lights from a 1.8-m GoodEarth Self-Adhesive Tape Lighting Kit were cut to fit the circumference of the bucket. The protective paper strip was removed from the back of the lights, allowing it to be directly applied to the inside of the bucket. The lights were adhered to the interior of the bucket starting at the 2-cm hole, making sure they remained parallel to the mouth of the bucket. Any necessary paint touch-ups were performed using an artistic brush.

For an imaging surface, a 1.29-m by 0.81-m marker board with a smooth particleboard back was used. The board was painted red using Classic Red Valspar Ultra Interior Flat Paint until the color was opaque and texture was no longer apparent. The red background was chosen for visibility of the heads during imaging. Each image was taken on the marker board surface using the imaging cap and the Canon Powershot G1 X Mark II camera. Each of the images were 20.1 megapixels and had a focal length of 9 mm, F-stop of f/4.0, and a shutter speed of 1/60 s at an ISO of 200 with a white balance of 3000 K. This allowed for clarity in the images and definition of the spike for analysis.

2.4 | Image analysis: Image J

For determination of SPS, six heads were collected from each plot at 30–33 days after the plot was fully headed but before the onset of senescence. All images were taken the same day as sampling. The peduncle of all six heads was inserted into a piece of red Van Aken Plastalina Modeling Clay, ensuring visibility of the heads in the image. Each head was placed in the clay in a uniform manner so that the rachis faced upward toward the lens to ensure each individual spikelet could be observed (Figure 2B).

SPS was manually measured using ImageJ version 1.52o (Caroline et al., 2012). Each image of six wheat heads was repeatedly cropped to create six separate photos of individual spikes. The cropped photos were padded to the same size of 140 by 600 pixels (Figure 2C) using the Python package Pillow version 8.4.0 (Clark, 2015). The spikelet number in each image was counted using the multitool on ImageJ, recorded in a comma-separated values (CSV) file, and used as the associated label for the image. After preparation, 12,717 total images were used for model development.

2.5 | Deep learning

Images were divided randomly into three different groups, with 70% used as a training set, 20% used for the test set, and the remaining 10% as a validation set. The training, validation,

and test sets remained the same for all models, and images of heads originating from the same plot were grouped so that they did not appear in more than one of the three datasets. Four different regression CNNs were trained, with the first being trained from scratch and having five alternating sets of convolutional and max-pooling layers. The second model was based on a pre-trained visual geometry group (VGG)16 network (Simonyan & Zisserman, 2014). The VGG16 model was selected since it has been utilized in several image-counting studies (Khaki et al., 2020; Ubbens & Stavness, 2017). Next was the pre-trained ResNet152V2 model (He et al., 2016), which is the most complex and recent model architectures available of their respective model series. The ResNet152V2 model was selected due to its prevalence in image classification literature in a variety of different fields from disease ratings in agriculture (Kanchanadevi & Sandhia, 2023; Nigam et al., 2023) to medical research (Sulaiman et al., 2023). The final model was the pre-trained EfficientNetV2L model (Tan & Le, 2021), which was selected due to its recent use in plant disease detection (Shovon et al., 2023; Ulutaş & Aslantaş, 2023). Mean squared error (MSE) was used as the loss function for each model, which evaluates differences between values predicted by the model and their true value. All of the models were implemented using the TensorFlow Keras package (Abadi et al., 2016) and trained on a local computer server utilizing no graphic processing units and 500 gigabytes of RAM.

The VGG16, ResNet152v2, and EfficientNetV2L models were made available through Keras (Chollet, 2015). Image arrays for the first model were normalized by dividing them by 255, the number of color values possible for red, green, and blue in an image, to put vector values between 0 and 1, while images for the VGG16, ResNet152V2, and EfficientNetV2L models were preprocessed using their respective TensorFlow Keras preprocessing functions.

The MSE was used as the loss function for all models. The MSE was calculated using the following equation:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2,$$

where N is the number of samples, y_i is the true value of the label, and $\hat{y}_i$ is the predicted label. Mean absolute error (MAE) gives the average absolute error between the predicted value and true value for spikelet number in an image and was the second metric used to evaluate model performance and accuracy.

The MAE was calculated using the following equation:

$$\text{MAE} = \frac{\sum_{i=1}^{n}\left|\hat{y}_i - y_i\right|}{N}$$

MAE and root mean square error (RMSE), which is the square root of MSE, were used to evaluate the average deviation of the estimated values for spikelet numbers from the true value across images. Each model was set to run for a maximum of 100 epochs but utilized early stopping-to-end training before 100 epochs if the validation loss had not improved for five consecutive epochs.

TABLE 1 The model mean absolute error (MAE) and root mean square error (RMSE) for each regression model.

Model	MAE	RMSE
Basic model	0.75	1.00
VGG16	1.51	1.29
ResNet152V2	0.61	0.78
EfficientNetV2L	0.60	0.79

3 | RESULTS

The mean spikelets per image was 17.5 with a range of 11–32 spikelets for all images. The basic CNN model had an MAE of 0.75 and an RMSE of 1.00 (Table 1). The VGG16 model had an MAE of 1.51 and an RMSE of 1.29. The ResNet152v2 model had an MAE of 0.61 and an RMSE of 0.78. The EfficientNetV2L model had an MAE of 0.60 and an RMSE of 0.79. The MAE values indicated that the EfficientNetV2L had the lowest average error between the true spikelet number in the image and the spikelet number predicted by the model, deviating 0.60 spikelets from the true value on average. The VGG16 model had an MAE over twice that of both the EfficientNetV2L and ResNet152V2 models. The VGG16 also had a much larger RMSE than each of the other models, indicating it is not as well fit.

Estimated spikelet values and actual spikelet values for each image were correlated for each model to compare the predicted spikelet number to the true value for each image in the test set. The regressions for the three best models, the basic CNN, ResNet152V2, and EfficientNetV2L, are depicted in Figure 3. The coefficient of determination (R^2) was 0.74 for the basic CNN (Figure 3A), 0.0076 for the VGG16, 0.87 for the ResNet152V2 (Figure 3B), and 0.90 for the EfficientNetV2L model (Figure 3C). To evaluate the error of each model, the error for each prediction was calculated by subtracting the true values from the predicted value for each test set image. The error distributions of the three best models are depicted in Figure 4. The mean and standard deviation of error for the basic CNN were −0.19 and 1.75, respectively. The range in errors for the model was −7.80 to 5.30 spikelets (Figure 4A). The average and standard deviation of errors for the VGG16 model were −0.17 and 1.89, with a range of −6.80 to 5.00 spikelets. The average and standard deviation of errors for the ResNet152V2 model were −0.35 and 0.69, respectively, with a range of −5.11 to 3.65 spikelets (Figure 4B). Finally, the average and standard deviation of errors for the

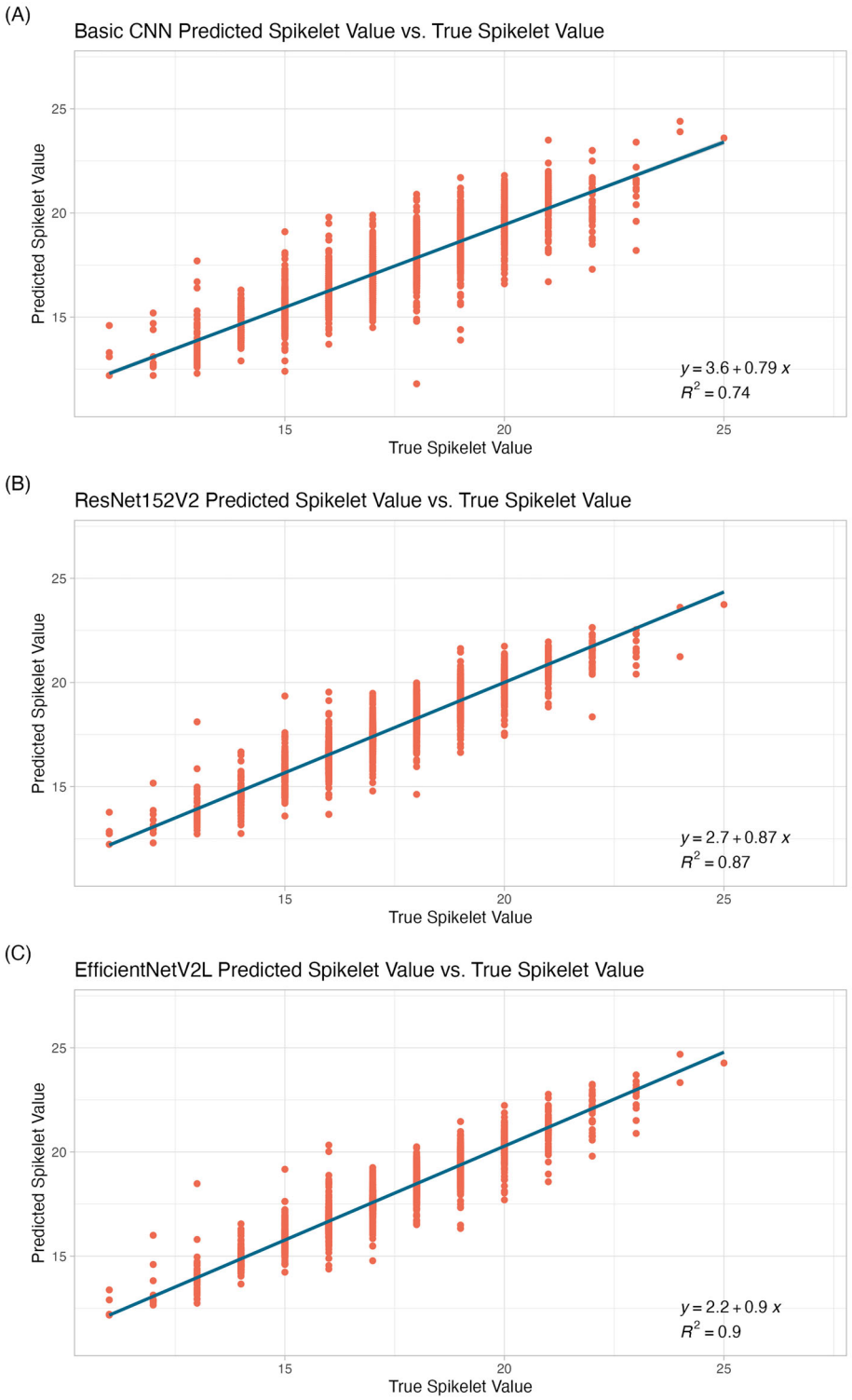

FIGURE 3 Regression of predicted spikelet number versus true spikelet value for each image for the basic (A) convolutional neural network (CNN), (B) ResNet152V2, and (C) EfficientNetV2L regression models. The *x* axis represents the true values of spikelet number present in each image and the *y* axis is the spikelet number value predicted by each respective model and image. Displayed within each graph is the equation of the linear regression and the coefficient of determination (R^2).

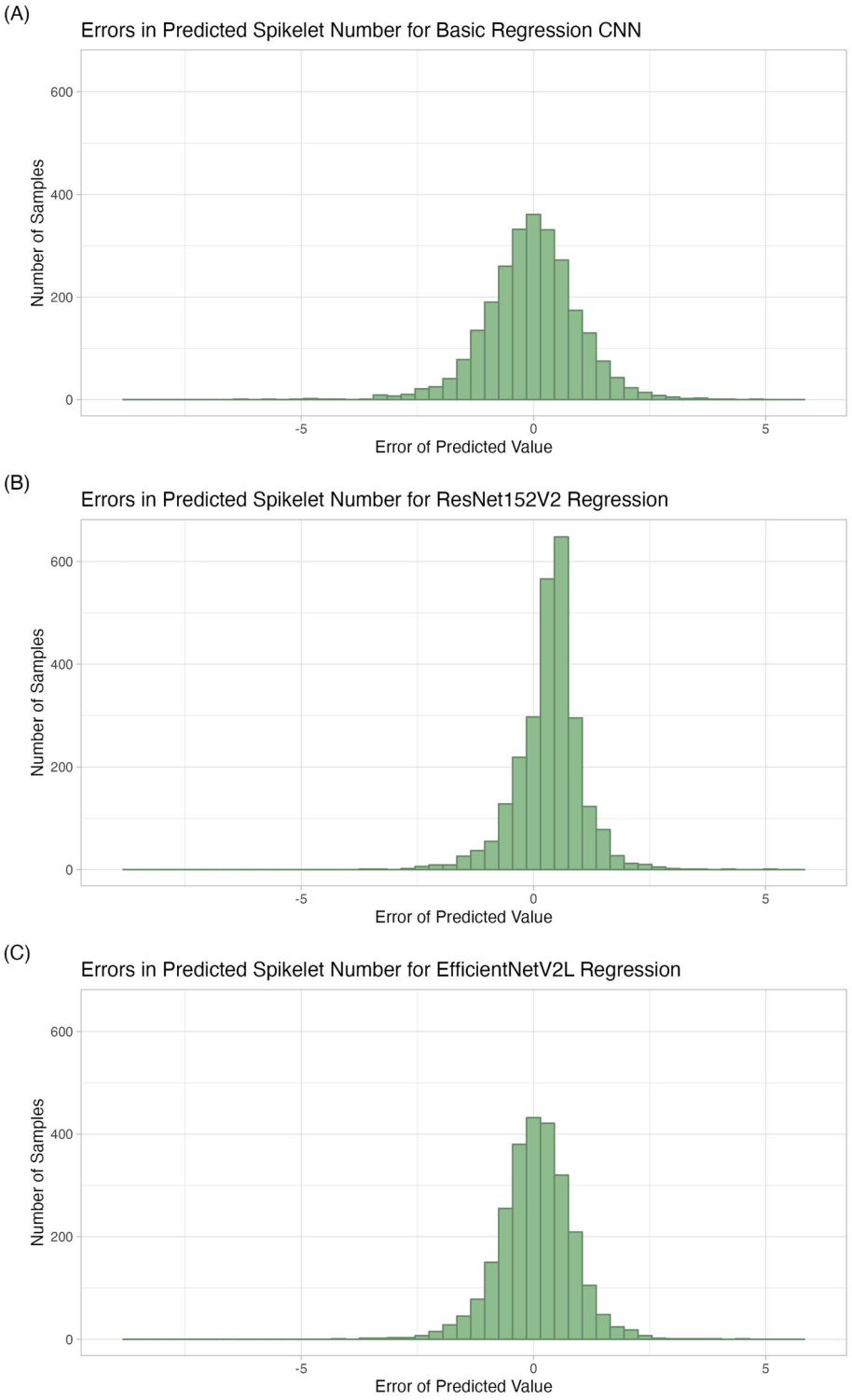

FIGURE 4 Histograms of error for each predicted spikelet number for the basic convolutional neural network (CNN), ResNet152V2, and EfficientNetV2L models. The *x* axis shows the difference between the predicted value for each image in the test set and the true spikelet number value in the image. The *y* axis shows the number of images represented for a given error range.

EfficientNetV2L were −0.06 and 0.79, respectively, with a range of −4.36 to 4.05 spikelets (Figure 4C).

4 | DISCUSSION

Developing phenotyping pipelines that allow for greater data collection is crucial to improving the efficiency of plant breeding programs. Imaging could serve as an affordable, accurate, and easily reproducible method for high-to-medium throughput phenotyping. The use of deep learning for image analysis is also promising, but further development and fine-tuning of hyperparameters is needed to generate optimal models. While it has been shown that the amount of data produced is more important than the accuracy of the data, acceptable thresholds for analyses in plant breeding programs must be determined prior to implementation (Lane & Murray, 2021).

The objective of this study was to demonstrate the capability of imaging for high-throughput counting of a critical wheat spike characteristic, SPS using deep learning models. Counting has a multitude of potential uses in agriculture and is a common method in ML models designed in other studies such as counting leaves (Miao et al., 2021; Ubbens & Stavness, 2017) and corn stalks in fields (Khaki et al., 2020). Generating a model for SPS not only provides a resource for phenotyping wheat spike characteristics but can also contribute to the development of models designed for similar tasks in agriculture. Both regression (Miao et al., 2021; Ubbens & Stavness, 2017) and detection-based methods (Khaki et al., 2020) have been implemented using sets of annotated images.

Miao et al. (2021) had an agreement rate, or proportion of exact predictions, ranging from 0.33 to 0.45 and MSEs ranging from 0.92 to 1.72 for leaf counting. Ubbens and Stavness (2017) used regression to count Arabidopsis and tobacco leaves and had MAEs of 0.41 and 0.61. These results are comparable to the accuracy of the basic CNN, ResNet152V2, and EfficientNetV2L models used for counting spikelets in this study which had MAEs of 0.75, 0.61, and 0.60, respectively.

One of the pre-trained models utilized in this study, VGG16, performed much more poorly than each of the other models. This may indicate that some models with more complex architecture, such as VGG16, may not be best suited for a task such as counting spikelets, or that the model was not optimally utilized. Counter to this, one of the most complex pre-trained models, EfficientNetV2L had the lowest MAE and best fit of all of the models, closely followed by the pre-trained ResNet152V2 and the basic CNN model. Even though the ResNet152V2 model had a slightly better fit than the EfficientNetV2L model, the EfficientNetV2L model still had the highest R^2 and MAE of all four models. This could be due to the slight overfitting of the ResNet152V2 model, or due to the fact that the EfficientNetV2L had a smaller range of

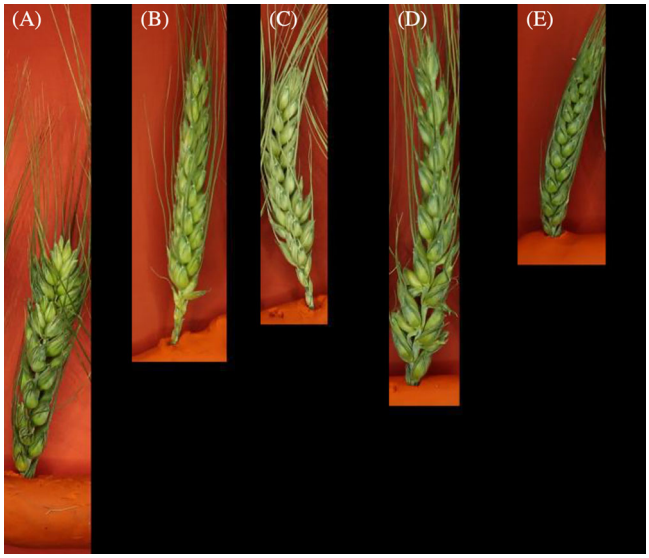

FIGURE 5 Examples of images in the test set that resulted in underestimation or overestimation of spikelet number. Common observations of images that produced larger spikelet estimation errors include images where the rachis was not fully visible due to turned spikelets (A) or the spike being slightly turned at the insertion point of the peduncle (B), crowding of a spike with awns of other neighboring spikes (C), supernumerary spikelets (D), and small images that contained large amounts of padding (E).

errors. The difference in fit between these two models is negligible since the RMSE values are almost equal (0.78 and 0.79). Since the basic CNN architecture is so simple, it also had the fastest runtime and required less computing power. The increased efficiency of the basic CNN model due to decreased runtime and memory requirements makes the basic CNN the more "user-friendly" option.

For all four models, the largest cause for overestimating spikelets appeared to be spikes that were slightly turned in the image so that the entire side view of the rachis was not perfectly aligned and visible (Figure 5A,B). For all three models, the largest errors for underestimating the number of spikelets were images with awns from neighboring spikes still present in the image (Figure 5C) or spikes containing supernumerary spikelets (Figure 5D). For the VGG16 model, spikes that were slightly turned in the image resulted in underestimations. Occasionally, larger error margins were identified for images that contained larger amounts of padding and smaller spikes (Figure 5E) compared to the other two models. The ResNet152V2 model did not show any patterns for large margins of error associated with abnormal spikes in images, and the EfficientNetV2L had the most trouble with excessive awns and slightly turned spikes.

While the errors found in this study were comparable to results from other studies, there is still room to improve accuracy of the models. Manual inspection found that images with large errors were consistently observed in patterns that may

have reduced the accuracy of the model. While some images appeared to be outliers, others that produced high error were very similar to other images in the set. For many images with high error, no issues could be identified with the image or spike present in the image.

For this method of imaging and phenotyping to be optimally used, the images would need to be gathered in a more efficient way, such as in the field. This would impose other challenges or potential sources of error. Taking images in the field would not allow for the same consistent lighting across spikes and images that is possible in a controlled environment. Time of day, the angle of lighting, and cloud cover could all vary either day to day, or multiple times within one data collection session. There would also be many spikes close together in the field, with each being turned in varying directions or occluding others. Analysis pipelines used for HTP of images taken in the field would need to be able to handle variability between images with regard to image quality and spike positioning to accurately phenotype spikelets. This would require much larger training sets in which these sources of variability can be observed.

There are limited datasets of annotated images for specific ML tasks in agriculture, making the use of pre-trained models or transfer learning more accessible for counting algorithms (Wang et al., 2017). Use of these methods could help further improve the accuracy of the models, allowing for the efficient and reliable use of ML models for HTP, both in and out of the field.

5 | CONCLUSIONS

Many methods are implemented by plant breeders for phenotyping crop traits, with little currently existing for wheat spike characteristics. This study used imaging techniques to analyze wheat spike characteristics and developed deep learning models for high-throughput analysis of wheat spike images. Images and techniques used in this study could be utilized for training models to analyze phenotypic information collected from rovers in fields or from UAVs, further increasing the efficiency of the phenotyping pipeline for spike morphology traits. Improving the efficiency of analyzing these traits will improve breeders' ability to discover and understand genetic components behind wheat spike architecture and their relationship to yield component traits. This will allow for the development of more environmentally efficient and higher yielding cultivars to meet future demands.

AUTHOR CONTRIBUTIONS

Mik Hammers: Data curation; formal analysis; investigation; methodology; project administration; software; visualization; writing—original draft; writing—review and editing. **Zachary Winn:** Conceptualization; data curation; methodology; project administration; writing—review and editing. **Asa Ben-Hur:** Methodology; supervision; validation; writing—review and editing. **Dylan Larkin:** Data curation; writing—review and editing. **Jamison Murry:** Data curation; writing—review and editing. **Richard Esten Mason:** Conceptualization; funding acquisition; project administration; resources; supervision; writing—review and editing.

ACKNOWLEDGMENTS

This work was supported by the USDA-ARS, the Agriculture and Food Research Initiative (AFRI) of the USDA National Institute of Food and Agriculture (NIFA) Grants 2022-68013-36439 and 2017-67007-25939 (Wheat-CAP), and in collaboration with SunGrains.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

Code utilized for this study can be found in a public folder on the author's GitHub. Image data are available by request to the author due to the file sizes being too large for a GitHub repository. https://github.com/mikhammers/Predicting-SPS.

ORCID

Mik Hammers https://orcid.org/0000-0003-2661-8389
Zachary J. Winn https://orcid.org/0000-0003-1543-1527

REFERENCES

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., … Zheng, X. (2016). *Tensorflow: Large-scale machine learning on heterogeneous distributed systems*. arXiv. https://doi.org/10.48550/arXiv.1603.04467

Blake, N. K., Lanning, S. P., Martin, J. M., Doyle, M., Sherman, J. D., Naruoka, Y., & Talbert, L. E. (2009). Effect of variation for major growth habit genes on maturity and yield in five spring wheat populations. *Crop Science*, *49*(4), 1211–1220. https://doi.org/10.2135/cropsci2008.08.0505

Burra, L. R., Bonam, J., Tumuluru, P., & Rao, B. N. K.(2022). Fine-tuning for transfer learning of ResNet152 for disease identification in tomato leaves. In B. N. K. Rao, R. Balasubramanian, S.-J. Wang, & R. Nayak (Eds.), *Intelligent computing and applications* (Vol. 315, pp. 295–302).Springer. https://doi.org/10.1007/978-981-19-4162-7_28

Caroline, A. S., Wayne, S. R., & Kevin, W. E. (2012). NIH image to imageJ: 25 years of image analysis. *Nature Methods*, *9*(7), 671–675. https://doi.org/10.1038/nmeth.2089

Chen, Z., Cheng, X., Chai, L., Wang, Z., Du, D., Wang, Z., Bian, R., Zhao, A., Xin, M., Guo, W., Hu, Z., Peng, H., Yao, Y., Sun, Q., & Ni, Z. (2020). Pleiotropic QTL influencing spikelet number and heading date in common wheat (*Triticum aestivum* L.). *Theoretical and Applied Genetics*, *133*(6), 1825–1838. https://doi.org/10.1007/s00122-020-03556-6

Chollet, F. (2015). *Keras*. GitHub. https://github.com/keras-team/keras

Clark, A. (2015). *Pillow (PIL Fork) documentation*. Release10.1.0.dev0. https://buildmedia.readthedocs.org/media/pdf/pillow/latest/pillow.pdf

Fahlgren, N., Feldman, M., Gehan, M. A., Wilson, M. S., Shyu, C., Bryant, D. W., Hill, S. T., McEntee, C. J., Warnasooriya, S. N., Kumar, I., Ficor, T., Turnipseed, S., Gilbert, K. B., Brutnell, T. P., Carrington, J. C., Mockler, T. C., & Baxter, I. (2015). A versatile phenotyping system and analytics platform reveals diverse temporal responses to water availability in *Setaria*. *Molecular Plant*, *8*(10), 1520–1535. https://doi.org/10.1016/j.molp.2015.06.005

FAO. (2022). *FAO cereal supply and demand brief*. https://www.fao.org/worldfoodsituation/csdb/en/

Fitzgibbon, J., Beck, M., Zhou, J., Faulkner, C., Robatzek, S., & Oparka, K. (2013). A developmental framework for complex plasmodesmata formation revealed by large-scale imaging of the *Arabidopsis* leaf epidermis. *The Plant Cell*, *25*(1), 57–70. https://doi.org/10.1105/tpc.112.105890

Furbank, R. T., & Tester, M. (2011). Phenomics: Technologies to relieve the phenotyping bottleneck. *Trends in Plant Science*, *16*(12), 635–644. https://doi.org/10.1016/j.tplants.2011.09.005

Gehan, M. A., & Kellogg, E. A. (2017). High-throughput phenotyping. *American Journal of Botany*, *104*(4), 505–508. https://doi.org/10.3732/ajb.1700044

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Identity mappings in deep residual networks. arXiv. https://doi.org/10.48550/arXiv.1603.05027

Houle, D., Govindaraju, D. R., & Omholt, S. (2010). Phenomics: The next challenge. *Nature Reviews Genetics*, *11*(12), 855–866. https://doi.org/10.1038/nrg2897

Jin, X., Zarco-Tejada, P. J., Schmidhalter, U., Reynolds, M. P., Hawkesford, M. J., Varshney, R. K., Yang, T., Nie, C., Li, Z., Ming, B., Xiao, Y., Xie, Y., & Li, S. (2021). High-throughput estimation of crop traits: A review of ground and aerial phenotyping platforms. *IEEE Geoscience and Remote Sensing Magazine*, *9*(1), 200–231. https://doi.org/10.1109/MGRS.2020.2998816

Kanchanadevi, K., & Sandhia, G. K. (2023). A comparative survey of maize leaf diseases using pre-trained convolutional neural networks. In *Third international conference on advances in electrical, computing, communication and sustainable technologies (ICAECT)* (pp. 1–6). IEEE. https://doi.org/10.1109/TNNLS.2021.3057958

Khaki, S., Pham, H., Han, Y., Kent, W., & Wang, L. (2020). *High-throughput image-based plant stand count estimation using convolutional neural networks*. arXiv. https://arxiv.org/abs/2010.12552

Koboldt, D. C., Steinberg, K. M., Larson, D. E., Wilson, R. K., & Mardis, E. R. (2013). The next-generation sequencing revolution and its impact on genomics. *Cell*, *155*(1), 27–38. https://doi.org/10.1016/j.cell.2013.09.006

Koppolu, R., & Schnurbusch, T. (2019). Developmental pathways for shaping spike inflorescence architecture in barley and wheat. *Journal of Integrative Plant Biology*, *61*(3), 278–295. https://doi.org/10.1111/jipb.12771

Krause, M. R., Mondal, S., Crossa, J., Singh, R. P., Pinto, F., Haghighattalab, A., Shrestha, S., Rutkoski, J., Gore, M. A., Sorrells, M. E., & Poland, J. (2020). Aerial high-throughput phenotyping enables indirect selection for grain yield at the early generation, seed-limited stages in breeding programs. *Crop Science*, *60*(6), 3096–3114. https://doi.org/10.1002/csc2.20259

Lane, H. M., & Murray, S. C. (2021). High throughput can produce better decisions than high accuracy when phenotyping plant populations. *Crop Science*, *61*(5), 3301–3313. https://doi.org/10.1002/csc2.20514

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Miao, C., Guo, A., Thompson, A. M., Yang, J., Ge, Y., & Schnable, J. C. (2021). Automation of leaf counting in maize and sorghum using deep learning. *Plant Phenome Journal*, *4*(1), 20022. https://doi.org/10.1002/ppj2.20022

Milella, A., Marani, R., Petitti, A., & Reina, G. (2019). In-field high throughput grapevine phenotyping with a consumer-grade depth camera. *Computers and Electronics in Agriculture*, *156*, 293–306. https://doi.org/10.1016/j.compag.2018.11.026

Minervini, M., Scharr, H., & Tsaftaris, S. A. (2015). Image analysis: The new bottleneck in plant phenotyping [Applications Corner]. *IEEE Signal Processing Magazine*, *32*(4), 126–131. https://doi.org/10.1109/msp.2015.2405111

Miura, H., & Worland, A. (1994). Genetic control of vernalization, day-length response, and earliness per se by homoeologous group-3 chromosomes in wheat. *Plant Breeding*, *113*(2), 160–169. https://doi.org/10.1111/j.1439-0523.1994.tb00718.x

Montesinos López, O. A., Montesinos López, A., & Crossa, J. (2022). Convolutional neural networks. In *Multivariate statistical machine learning methods for genomic prediction* (pp. 533–577). Springer International Publishing. https://doi.org/10.1007/978-3-030-89010-0_13

Nigam, S., Jain, R., Marwaha, S., Arora, A., Haque, M. A., Dheeraj, A., & Singh, V. K. (2023). Deep transfer learning model for disease identification in wheat crop. *Ecological Informatics*, *75*, 102068. https://doi.org/10.1016/j.ecoinf.2023.102068

Rutkoski, J., Poland, J., Mondal, S., Autrique, E., Pérez, L. G., Crossa, J., Reynolds, M., & Singh, R. (2016). Canopy temperature and vegetation indices from high-throughput phenotyping improve accuracy of pedigree and genomic selection for grain yield in wheat. *G3 Genes|Genomes|Genetics*, *6*(9), 2799–2808. https://doi.org/10.1534/g3.116.032888

Shewry, P. R. (2009). Wheat. *Journal of Experimental Botany*, *60*(6), 1537–1553. https://doi.org/10.1093/jxb/erp058

Shovon, M. S. H., Mozumder, S. J., Pal, O. K., Mridha, M. F., Asai, N., & Shin, J. (2023). PlantDet: A robust multi-model ensemble method based on deep learning for plant disease detection. *IEEE Access*, *11*, 34846–34859. https://doi.org/10.1109/ACCESS.2023.3264835

Simonyan, K., & Zisserman, A. (2014). *Very deep convolutional networks for large-scale image recognition*. arXiv. https://doi.org/10.48550/arXiv.1409.1556

Singh, A., Ganapathysubramanian, B., Singh, A. K., & Sarkar, S. (2016). Machine learning for high-throughput stress phenotyping in plants. *Trends in Plant Science*, *21*(2), 110–124. https://doi.org/10.1016/j.tplants.2015.10.015

Song, P., Wang, J., Guo, X., Yang, W., & Zhao, C. (2021). High-throughput phenotyping: Breaking through the bottleneck in future crop breeding. *The Crop Journal*, *9*(3), 633–645. https://doi.org/10.1016/j.cj.2021.03.015

Stelmakh, A. F. (1998). Genetic systems regulating flowering response in wheat. *Euphytica*, *100*(1), 359–369. https://doi.org/10.1023/A:1018374116006

Sulaiman, A., Kaur, S., Gupta, S., Alshahrani, H., Reshan, M. S. A., Alyami, S., & Shaikh, A. (2023). ResRandSVM: Hybrid approach for acute lymphocytic leukemia classification in blood

smear images. *Diagnostics*, *13*(12), 2121. https://doi.org/10.3390/diagnostics13122121

Tan, M., & Le, Q. V. (2021, April 1). *EfficientNetV2: smaller models and faster training*. Presented at the International Conference on Machine Learning. https://doi.org/10.48550/arXiv.2104.00298

Tsaftaris, S. A., Minervini, M., & Scharr, H. (2016). Machine learning for plant phenotyping needs image processing. *Trends in Plant Science*, *21*(12), 989–991. https://doi.org/10.1016/j.tplants.2016.10.002

Ubbens, J. R., & Stavness, I. (2017). Deep plant phenomics: A deep learning platform for complex plant phenotyping tasks. *Frontiers in Plant Science*, *8*, 1190–1190. https://doi.org/10.3389/fpls.2017.01190

Ulutaş, H., & Aslantaş, V. (2023). Design of efficient methods for the detection of tomato leaf disease utilizing proposed ensemble CNN model. *Electronics*, *12*(4), 827. https://doi.org/10.3390/electronics12040827

Wang, G., Sun, Y., & Wang, J. (2017). Automatic image-based plant disease severity estimation using deep learning. *Computational Intelligence and Neuroscience*, *2017*, Article 2917536. https://doi.org/10.1155/2017/2917536

Winn, Z. J., Larkin, D. L., Murry, J. T., Moon, D. E., & Mason, R. E. (2021). Phenotyping anther extrusion of wheat using image analysis. *Agronomy*, *11*(6), 1244. https://doi.org/10.3390/agronomy11061244

Yang, W., Guo, Z., Huang, C., Duan, L., Chen, G., Jiang, N., Fang, W., Feng, H., Xie, W., Lian, X., Wang, G., Luo, Q., Zhang, Q., Liu, Q., & Xiong, L. (2014). Combining high-throughput phenotyping and genome-wide association studies to reveal natural genetic variation in rice. *Nature Communications*, *5*(1), 5087–5087. https://doi.org/10.1038/ncomms6087

How to cite this article: Hammers, M., Winn, Z. J., Ben-Hur, A., Larkin, D., Murry, J., & Mason, R. E. (2023). Phenotyping and predicting wheat spike characteristics using image analysis and machine learning. *The Plant Phenome Journal*, *6,* e20087. https://doi.org/10.1002/ppj2.20087