Detection of Fusarium Damaged Kernels in Wheat Using Deep Semi-Supervised Learning on a Novel WheatSeedBelt Dataset

Keyhan Najafian[1] Lingling Jin[1] H. Randy Kutcher[2] Mackenzie Hladun[2]
Samuel Horovatin[1] Maria Alejandra Oviedo-Ludena[2] Sheila Maria Pereira de Andrade[2]
Lipu Wang[*,2] Ian Stavness[*,1]

{keyhan.najafian, lingling.jin, randy.kutcher, mah486, s.horovatin, alejandra.oviedo, sheila.andrade, lipu.wang, ian.stavness}@usask.ca

[1]Department of Computer Science, University of Saskatchewan, Saskatoon, SK, Canada
[2]Department of Plant Sciences, University of Saskatchewan, Saskatoon, SK, Canada

Abstract

Fusarium head blight, caused by Fusarium spp., is a destructive disease of wheat worldwide. Fusarium damaged kernels (FDKs) significantly reduce grain yield and quality. Thus, FDK detection is a priority for wheat breeders seeking to develop high-grain quality and FDK-resistant wheat cultivars. However, traditional FDK measurement methods are time-consuming, labor-intensive, and of variable accuracy. Image-based phenotyping methods have the potential to efficiently detect FDK, but are challenging to develop due to the lack of large-scale damage-annotated wheat kernel datasets. Addressing this issue, we introduced WheatSeedBelt, a high-resolution large-scale dataset including $40,420$ *close-up top- and side-view single-kernel images of* 268 *wheat varieties with kernel damage annotations. Utilizing this dataset, we developed an image-processing pipeline to efficiently process images and extract the representative features for machine and deep-learning purposes. We also conducted three experiments on the dataset using pretraining and semi-supervised fine-tuning phases to classify wheat kernels into healthy, unhealthy but non-FDK, and FDK affected. Our best models achieved an F1-score of* 84.29% *for the Healthy-Unhealthy (including FDKs) task,* 56.35% *for the binary FDK-nonFDK, and* 68.30% *for the 3-class task (Healthy, Unhealthy, and FDK). We also conducted an inter-rater reliability study, which indicated that human experts do not outperform our model in FDK prediction, providing evidence that visual classification of FDK from RGB images is a challenging task.*

1. Introduction

Wheat is a staple food crop that provides a significant portion of the world's caloric intake, especially in developing countries. Wheat kernels are nutritious and contain essential vitamins, minerals, and dietary fiber. The ease of cultivation and relatively high yield make it an efficient and cost-effective crop for farmers across the globe [19, 15, 39]. Wheat kernels can be damaged by both abiotic stress, such as excess heat [43], and biotic stress, such as fungal pathogens [4] and insects [16]. For example, exposure to high temperatures results in germ- or heat-damaged kernels with distorted colors, fungi can cause damage such as discoloration and light, and insects usually damage the wheat kernels by chewing them.

Fusarium head blight (FHB) is a globally prevalent and destructive wheat disease caused by *Fusarium* spp. It adversely affects the development of wheat kernels, leading to the formation of lightweight, chalky white, and shrunken kernels. These affected kernels are commonly referred to as Fusarium damaged kernels (FDKs) [25, 37]. Infected kernels are usually contaminated with Fusarium-produced mycotoxins, especially deoxynivalenol (DON). FDK and DON contamination reduce grain quality, which limits marketing opportunities, adds to cleaning costs, and results in discounted prices. Further, FDK and DON can cause poor-quality food products, immunological and teratogenic problems in humans, and reduced livestock productivity due to toxicity or feed refusal. Achieving low FDK and DON in cereal germplasm is thus crucial to cereal breeders, cereal growers, and food producers.

The current tools to score or monitor FDK and DON are highly inefficient. For instance, FDK assessment is usually carried out by visual inspection, which requires extensive human labor, and the assessment accuracy can be dramatically reduced by fatigue and external distractions. Recently, machine learning and deep learning algorithms have been applied to inform precision agriculture [9, 17]. The effectiveness of these methods depends on the availability of large-scale annotated crop, plant, and seed datasets that are precisely organized [20, 45].

In this study, we collected images of wheat kernels after harvest in a controlled environment utilizing a conveyor imaging system called BELT developed by [11]. We organized these images into the WheatSeedBelt dataset, consisting of $40,420$ high-resolution images depicting wheat kernels from a diverse selection of wheat varieties from two locations. Kernel images were manually labeled by expert wheat pathologists into three classes: Healthy (kernels that appear healthy), FDK (kernels that appear unhealthy due to FHB infection), and Unhealthy (kernels that appear unhealthy due to reasons other than FHB, e.g. damage from other biotic and abiotic stress). This dataset can assist agronomists in several wheat kernel-related analyses, such as kernel size/shape and kernel damage estimation.

There are two key contributions to our analyses. First, we developed a pipeline to automatically process the raw image data and extract important features from the WheatSeedBelt, preparing it for the development of machine and deep learning models. Second, we conducted semi-supervised [48, 27] deep-learning analyses with the goal of separating FDKs from healthy kernels and from unhealthy non-FHB kernels. In order to achieve this goal, we evaluated two binary tasks, Healthy versus Unhealthy (including FDK and non-FDK) and FDK versus non-FDK, as well as a 3-class recognition task for classifying Healthy, Unhealthy (unhealthy but non-FDK), and FDK. From 268 wheat varieties/packets in the WheatSeedDataset, we selected 36 packets for model development and evaluation. We divided these packets into training, validation, and test sets, including 11, 5, and 20 packets, respectively. We chose a packet-wise split strategy to avoid any information leaks among the split sets. Our optimal deep models were designed based on the EfficientNet [41] architecture and achieved $F1$ scores of 84.29, 53.62, and 68.30 for the three tasks, respectively.

We also conducted an inter-rater reliability study to evaluate the difficulty of assessing damage in wheat kernel images. A panel of three raters (LW, MO, and SP) independently assessed the wheat kernels in a subset of the WheatSeedBelt dataset. We then assessed the level of consensus among multiple raters in assigning labels to the data samples. This study shows that distinguishing FDK, healthy, and unhealthy kernels from each other is a difficult task for human experts, leading to a large proportion of disagreements in labels. This suggests the presence of labeling bias, which could impact model performance. Additionally, this enabled us to identify biases or variations in ratings, thereby augmenting the reliability and validity of our findings.

In summary, the main contributions of this work include (1) generating a large-scale and diverse dataset that captures $40{,}420$ RGB images of wheat kernels from 268 wheat varieties; (2) developing and evaluating an image processing pipeline that automatically extracts image features and Regions of Interest (ROIs) from conveyor-belt images of wheat kernels; (3) developing and evaluating deep learning models to automatically detect wheat kernels' health conditions, including healthy, unhealthy, and FDK; (4) conducting an inter-rater reliability study to judge the effectiveness and consistency of expert annotations of wheat kernel damage from wheat kernel images.

1.1. Related Work

Spectral imaging has been used in various studies [32, 47, 38] to capture wheat grain. Zhou et al. [47] conducted a comprehensive investigation by assembling a large-scale dataset that included $147,096$ low-resolution images, each with a channel size of 200, in 30 varieties. They utilized a Near-infrared (NIR) hyperspectral imaging system to capture patches of wheat kernels on a plate with low reflection and dark background, which facilitated the kernel segmentation process. Additionally, they developed a convolutional neural network (CNN)-based feature selector equipped with an attention mechanism to extract informative spectral channels for a fully supervised classification task of classifying images into 30 categories, and achieved an accuracy of 93% on the prediction set, which is considered an internal evaluation of the model. Polder et al. [32] studied the identification of Fusarium in individual wheat kernels by analyzing a dataset of 96 spectral images taken from a range of kernels that varied from heavily damaged to healthy. In this work, hyperspectral imaging was employed to collect spectral data from both healthy and Fusarium-inoculated wheat kernels. The images were analyzed using fuzzy c-means clustering and supervised partial least squares regression to describe the image information with quantitative information of Fusarium DNA concentrations.

Aside from the high cost of image collection and analysis with spectral imaging systems, the advances in deep learning models and the availability of cameras has increased interest in gathering massive RGB image datasets [36, 29, 46, 6, 28, 23] to support scholars and breeders to advance their theories and research in various fields of study including precision agriculture. Zheng et al. [46] introduced the CropDeep dataset, which consisted of $31,147$ greenhouse images, cropped to a maximum size of 1000×1000, of vegetables, fruits, and people in greenhouses, to be used in classification and detection tasks. The dataset contains over $49,000$ object-level annotated specimens belonging to 31 special classes. Rauf et al. [33] constructed the Citrus dataset, which consists of citrus fruits, leaves, and stems obtained at the point of fruit ripening and disease peak. The dataset encloses healthy and diseased citrus plant images affected by pathologies that include black spot, canker, scab, greening, and melanose.

Leaves and flowers are the plant parts most prone to disease damage. The 102 Flower Category Dataset, as described by Nilsback et al. [29], is comprised of $8,189$ images representing 102 flower varieties each including 40 to 258 samples chosen from flowers with various colors, shapes, and textures, and used for classification purposes. The PlantVillage dataset [14] also possesses $54,303$ healthy and unhealthy leaf images divided into 38 categories by species and diseases. The images depict 14 different crop species, such as apple and blueberry. Plant pathologists determined the stages of different diseases observed from the leaf images. The soybean dataset [24] holds $6,410$ 500×500 JPEG images of healthy and insect-damaged soybean leaves, captured with an Unmanned aerial vehicle (UAV) and phones. It includes three categories of data: healthy plants, caterpillar-damaged plants, and *Diabrotica Speciosa*-damaged plants.

Large-scale RGB datasets for wheat kernels are scarce. Halcro et al. [11] developed an imaging system that utilizes statistical image analysis software, designed specifically for small-seed optical analysis of a dataset of only 40 low-resolution RGB images captured on a black background. Ropelewska et al. [35] presented a study on single-seed image dataset to recognize healthy and FDKs of wheat. The author utilized machine learning models on approximately 200 selected textural features per color channel of 5 kg of kernels of two wheat varieties, a total of 240 hyperspectral images, and $4,490$ colored images of size 500 by 700 pixels captured by a flatbed scanner to train several models. The Global Wheat Head Detection (GWHD) dataset [6] was developed to provide RGB images of wheat spikes captured under field conditions for object detection purposes. The dataset contains $275,187$ wheat spikes in $6,422$ images developed by 16 institutions distributed across 12 countries. The dataset was evaluated under different applications such as wheat spike detection [27, 10], wheat spike segmentation [34, 26], and wheat spike counting [22, 18].

The datasets mentioned above are suitable for training deep models and performing extensive evaluations. Nevertheless, the WheatSeedBelt dataset we developed is unique. It includes high-resolution images depicting single kernels of wheat as shown in Figure 1. This dataset has the potential to provide researchers with a rich source of genetically diverse wheat kernels with varying phenotypic characteristics to develop their ideas using novel machine and deep learning approaches.

2. Materials and Methods

In this section, we provide a description of the data acquisition process and highlight the distinctive characteristics of the curated dataset. We deliver a detailed report on the pre-processing methodology developed for data preprocessing, cleaning, splitting, and enlargement, which were necessary to meet our model development requirements. Additionally, we discuss the development of deep learning models to detect the health status of wheat kernels.

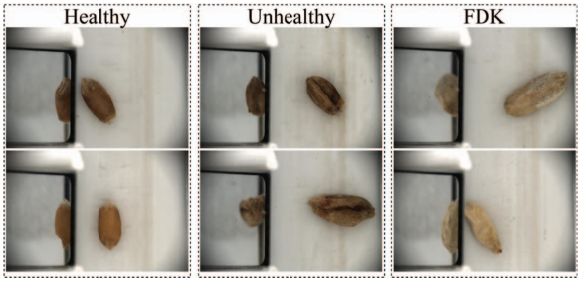


Figure 1: A few examples of the WheatSeedBelt images in different health conditions. Healthy, Unhealthy but non-FDK, and FDKs of wheat are shown in columns left to right.

2.1. Data Acquisition

A population of 300 spring wheat (*T. aestivum*) accessions were obtained from the Plant Gene Resources of Canada (PGRC) and inoculated in mist-irrigated FHB nurseries at the University of Saskatchewan (Saskatoon, SK), University of Manitoba (Carman, MB), and University of Guelph (Elora, ON) in 2021. Nurseries were artificially inoculated with a liquid inoculum of *F. graminearium* macrospores prepared at each location. FHB incidence, severity, plant height, and heading date were recorded. Accessions were harvested at all locations and were threshed with fans set low to maintain all damaged kernels.

Wheat kernels from the same variety were organized into the same packets. To capture kernel images, we employed an automatic and portable conveyor imaging system, equipped with *Chameleon3 CM3-U3-50S5C-C5 USB* cameras, devised in [11]. Wheat kernels from each packet were loaded into an imaging chamber and each kernel was individually imaged to acquire its top and side view (via an oblique mirror). We refer to the original, unprocessed images as the WheatSeedBelt dataset, which is comprised of $44,710$ images organized into 293 genetically distinct breeding packets, each packet contains wheat kernels ranging in number from 2 to 294 with an average of 150.8. Additionally, the image intensities across all packets demonstrated an average range of $314,53.89$ to $366,94.49$, with an overall mean of $346,59.62$. Figure 1 displays examples of original images capturing wheat kernels in the three conditions: healthy, unhealthy (non-FDK), and FDK.

2.2. Data Preprocessing

The original high-resolution images in the WheatSeedBelt dataset pose challenges such as higher computational costs, more complexity, and longer processing time. Therefore, optimizing the resolution of images for deep learning algorithms is important. To accomplish this goal, we

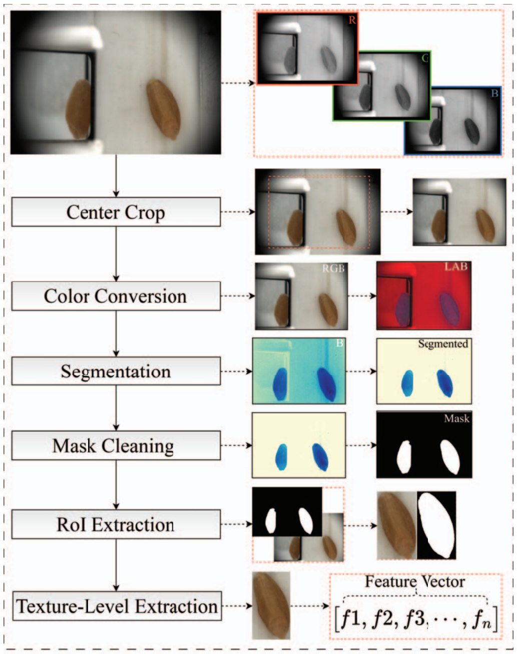


Figure 2: An overview of the pre-processing pipeline developed for wheat kernel images. First, we used a center crop to remove boundary noise or artifacts that may adversely affect segmentation accuracy. This was followed by color conversion, segmentation, mask cleaning, then the extraction of regions of interest and relevant texture-level features recorded as feature vectors.

developed a strategic image processing pipeline specifically to optimize the high-resolution images of the WheatSeedBelt dataset and extract the most informative texture-level features and regions of interest (ROI) and texture-level features that are efficient in developing machine or deep learning models. Figure 2 showcases the pipeline we developed for processing the WheatSeedBelt images. The source code for the data processing pipeline is publicly accessible at `https://github.com/USask-BINFO/WheatSeedBelt`.

During the initial stage of data processing, we aimed to convert the image intensities into a conventional range to avoid potential information loss in the rest of the preprocessing steps. As a result of this transformation, the dataset was reduced to $40,420$ high-resolution color images of size 1500×2200 (height and width, respectively), which were categorized into 268 packets. It is worth noting that the intensity of the images was converted into a conventional range of 0 to 255 that is well supported by Python packages, such as Sikit-image [42], Albumentations [2], and PyTorch [31], whereas prior to conversion, the intensity range was approximately $3,583$ to $65,535$.

In the next steps, we used traditional image processing techniques such as color space conversion, thresholding, and morphological operations, to segment and accurately extract the wheat kernels. We first cropped the central region of the images to a fixed size of 1200×1800 to remove the dark boundaries, which add extra noise to the final segmentation. We only extracted the top-view kernels that best fit our goals. Note that breeders and scientists who are interested in evaluating the visual statistical information of genetically diverse wheat species could benefit from the information in the side view of the kernels and therefore it is included in the WheatSeedBelt dataset.

The segmentation phase includes a few steps. We converted the *RGB* images into *CIELAB* color space and used the *B* channel, which more precisely differentiated the wheat kernels from the background. After that, we used Otsu thresholding [30] to segment the image, then improved the segmentation by removing clutter or undesirable elements operating morphological Opening [8, 40], eliminating small objects, and removing small holes in order. Having the accurate mask for the image, we extracted the top-view wheat kernel object by overlaying the segmentation mask on top of the image. The newly processed dataset is referred to as SS (Figure 3) in which each image contains a single kernel. The SS dataset was used for model development, which only includes highly compressed and concentrated samples. It reduces the demand for more VRAM, and the processing costs of high-resolution images and accelerates the learning procedure of deep models.

Finally, we extracted texture-level features from the top-view instance. By doing so, we minimized the amount of noise introduced into our extracted texture-level features by obtaining only the regions of interest (rather than the whole image). Particularly, we obtained histogram [3], Local Binary Patterns [1], and Gray-Level Co-occurrence features [12, 7] of the top-view instances within the SS. Using texture-level features allows more computationally efficient, robust to noise, and interpretive descriptions of the images for model training. We refer to the texture-level features data set as FS, which is of the same size as SS.

2.3. Dataset Split

Annotating a large-scale dataset, such as the SS dataset with more than forty thousand images, can be impractical and costly. In order to develop deep models with reasonable annotation efforts, we only annotated a small proportion of the samples (approximately 10%) that were selected from a limited number of packets (36 packets). The annotated dataset contains $3,923$ images, representing a useful subset of the SS dataset for deep model development. In order to maximize the annotation accuracy, two raters labeled each image as Healthy, Unhealthy, and FDK with mutual agreement. This ensured that the labeling process, which has a significant impact on the entire model development

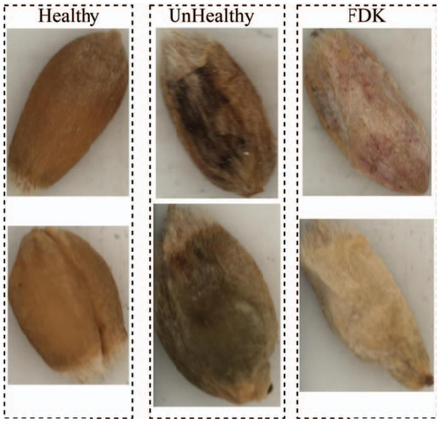


Figure 3: Examples of single kernel images for each kernel damage category that were extracted by our pre-processing pipeline.

Table 1: The number of images in the training (tr), validation (va), and testing (te) splits of our dataset for each category. The dataset split for the SS and FS series is identical, where SS includes images and the FS includes texture-level feature vectors extracted from SS images.

Dataset	Health Condition	Number of Images
SS_{tr} (FS_{tr})	Healthy	302
	Unhealthy	637
	FDK	392
SS_{va} (FS_{va})	Healthy	164
	Unhealthy	194
	FDK	127
SS_{te} (FS_{te})	Healthy	976
	Unhealthy	883
	FDK	248

process, is accurate and reliable.

We then categorized the images in each annotated packet into three categories: Healthy, Unhealthy, and FDK. To avoid information leaks, we devised a packet-wise division strategy to split the labeled packets into training, validation, and test sets. We selected images from 11 packets to serve as our training set denoted as SS_{tr}, picked 5 packets for validation called SS_{va}, and the remaining 20 packets for testing purposes named SS_{te}. We also reserved the unlabeled samples from the SS dataset as SS_{un} for further pseudo-label refinement modeling. Table 1 provides more information on the number of samples annotated and split for model development. SS_{tr}, SS_{va}, and SS_{te} can be accessed at `https://binfo.usask.ca/Projects/WheatSeedBelt/WHEATSEEDBELT.zip`.

We designed a rater reliability study to evaluate the consistency of classifying the wheat seeds into Healthy, Unhealthy, and FDK categories by three raters. In the inter-rater reliability study, we randomly chose 10% of the SS_{te} to be labeled by three experts, independently. This dataset was named SS_{ra}.

2.4. Data Augmentation

We utilized a relatively limited number of examples for model development (Table 1). Given the large-scale data, which included a broad range of genetically distinct variations, it was critical to develop a model that was generalizable across domains and resistant to overfitting on small sets of packets. To achieve this, we employed augmentation techniques to enhance the diversity of the training data.

To mitigate the effect of data imbalance, we expanded the SS_{tr} set using the rotation operation to expose the model with more computationally generated samples for each class. We employed a rotation augmentation technique to rotate each image in the SS_{tr} set at specific intervals in degrees between 0 and 360, as outlined in [26]. The degrees were obtained from $\left\{t \times n \mid t \in \{0, 1, \cdots, \lfloor \frac{360}{n} \rfloor\}\right\}$, where $n = 15$ for Healthy and FDK categories, and $n = 30$ for the Unhealthy class.

Despite utilizing a camera with established settings and within a controlled environment, the SS dataset still contains wheat kernels of varying sizes and shapes, both horizontally and vertically oriented. This was caused by the differences in kernel size and position in the conveyor imaging system and their distance from the camera. However, deep learning models require input images of consistent size. Considering the potentially large difference between the height and width of wheat kernel images, we used padding instead of resizing to ensure that all images were of the same size. This was because resizing can distort the aspect ratio of rectangular kernel images when converted to a square shape, while padding simply adds extra pixels to the image boundaries, thus enlarging the image without altering the aspect ratio. Additionally, padding preserves the spatial information that can be lost through the resize process, which sacrifices the informative fine-grained features.

In addition to padding, we applied normalization in model development phases, including training, validation, and testing. To further increase the variability in the training set, we employed spatial transformations, such as Random Rotation, Flip, a list of Random Crops with sizes between 256 to 512, and pixel-level augmentations such as Color Jitter, Channel Shuffle, Random Noise, and Blurring. We leveraged the Albumentations library [2] to implement the above-mentioned augmentation transformations.

3. Model Development and Training

In this study, we conducted three semi-supervised experiments using EfficientNet-B0 [41] and ResNet18 [13], two of the best-performing architectures pretrained on ImageNet [36]. The first experiment involved training the models on a binary classification task of Healthy vs. Unhealthy using the split datasets SS_{tr} and SS_{va}. The second experiment involved training the models for a binary task to detect FDKs on wheat, again using the SS_{tr} and SS_{va} datasets. Finally, we conducted a 3-class classification experiment involving all Healthy, Unhealthy, and FDK categories.

For all deep learning experiments, we fine-tuned the models' backbones and trained newly replaced classifiers from scratch. We used the same classifier consisting of a batch normalization, a dropout, and a linear layer for both models in all experiments. We trained the models for 10 epochs by deploying the Adam optimizer [44] with a coupled learning rate and weight decay of ($1e-3$, $1e-3$) and ($1e-4$, $1e-5$), the batch size of 32, and loss functions of BinaryCrossEntropy for the binary tasks and CrossEntropy for the multi-class one. In each epoch, we assessed model performance on the validation set using the $F1$ metric and saved the best-performing model with the highest score. In section 4, we evaluated the chosen best-performing models using the Accuracy, Precision, Recall, and $F1$ metrics, calculated on the SS_{te} set.

To enhance our models' generalizability, we adopted a semi-supervised learning approach, where we utilized the pre-trained models to predict pseudo-labels for the unlabeled portion of the dataset, SS_{un}, comprised of wheat kernels with a broader spectrum of genotypic and phenotypic characteristics. We then fine-tuned the models on the pseudo-labeled datasets. Similar to the pre-training stage, we preserved the models with the highest $F1$ score on the validation set, SS_{va} as the best-performing model.

The trained models are denoted by a shorthand notation, which consists of the model name, initials of the task name, and the training phase. For instance, *EffB0-HU-A* represents the EfficientNet-B0 model trained for the binary task of Healthy vs Unhealthy, where letters A and B correspond to the pre-training and semi-supervised fine-tuning stages, respectively. The letter S is used to refer to models that have been trained from scratch; the keyword None indicates that no pretraining was performed.

For computation, we used a system with an NVIDIA Tesla V100S PCIe GPU, featuring 32 GB of RAM, and an Intel(R) Xeon(R) Gold 5220 CPU @ 2.20GHz. We implemented both image-processing and model development pipelines in Python "3.8" along with the Scikit-image version "0.18.3" and PyTorch "1.10.1+CU102" as the main packages.

4. Results

The results obtained from all the deep-learning experiments in this research are presented in Table 2, which shows the performance of each model in terms of accuracy, precision, recall, and $F1$ score.

We conducted three deep-learning experiments, namely Healthy vs Unhealthy (including FDKs), FDK vs non-FDK, and a 3-class classification task including Healthy, FDK, and Unhealthy (neither healthy nor FDK) categories. For these tasks, we employed two state-of-the-art deep-learning models, EfficientNet-B0 [41] and ResNet18 [13].

Figure 4 illustrates the confusion matrices of all three tasks for the fine-tuned models. These matrices provide a detailed representation of the model's performance for each class in the defined task. The confusion matrices allow us to evaluate the model's classification performance by showing true positive, true negative, false positive, and false negative percentages for each class.

Table 2 presents a comprehensive comparison between the developed models. Particularly, the fine-tuned EfficientNet model *EffB0-HU-B* achieved a high level of performance on the Healthy vs Unhealthy classification task, attaining an $F1$ score of 84.29, as evidenced by the other three evaluation metrics. Notably, this model demonstrated an outstanding ability to accurately differentiate healthy kernels from unhealthy and FDKs.

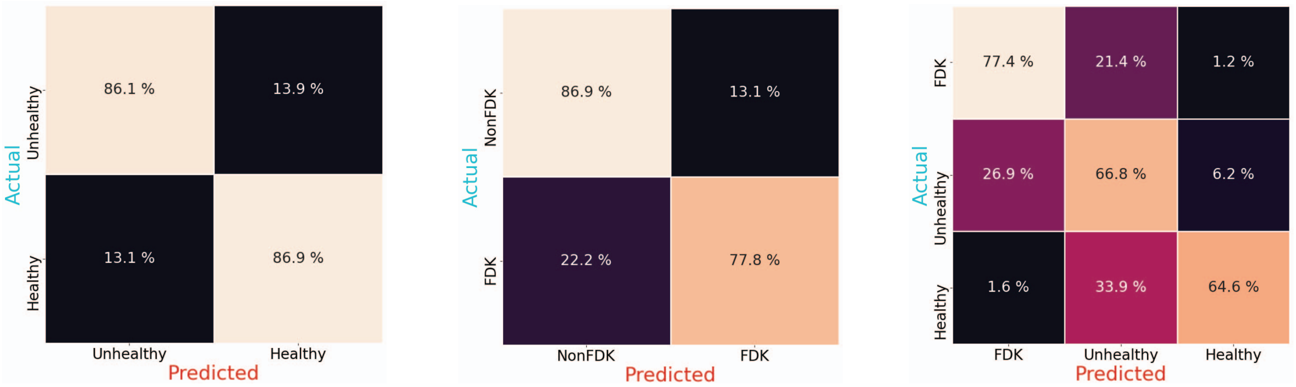
However, when comparing the performance of the model trained from scratch to that of the pretrained models, the latter exhibited a significant performance improvement. Similarly, the model *EffB0-HUF-A* achieved better performance on the Healthy-Unhealthy-FDK task. Notably, the precision and recall scores demonstrated comparable values. On the other hand, the model trained from scratch performed best compared to the other models on the FDK vs nonFDK classification task. This was mainly due to the high number of False Positives, leading to a low precision score, indicating the model predicts many of the nonFDK samples as FDK.

Compared to the 3-class classification task, binary tasks achieved higher accuracy in detecting positive and negative class samples, as shown in the confusion matrices (Figure 4). The confusion matrix (4a) for model *EffB0-HU-B* obtained the highest accuracy rates of 86.9% and 86.1% for the positive and negative classes, respectively, and a total $F1$ score of 84.29% (Table 2). The *EffB0-FN-S* model (4b) also exhibited high performance in detecting a high percentage of FDKs and non-FDKs. Note that the low $F1$ score for the *FN* models (Table 2), was due to the low number of FDK images in the SS_{te} set, which only numbered 248 images compared to the high number of non-FDKs. Moreover, the performance differences between pretrained and unpretrained EfficientNet models, *EffB0-FN-S* and *EffB0-FN-A*, are insignificant. However, the developed 3-class classification task models encountered challenges in distinguishing unhealthy kernels from healthy and FDKs while categorizing healthy and FDK samples with high accuracy (4c).

In addition to measuring model performance, we aimed to evaluate the inter-rater reliability of the SS_{ra} dataset, which was annotated by three raters (Table 3). We em-

Table 2: The performance measure of the developed models on the test set, SS_{te}, for all three experiments, Healthy vs Unhealthy, FDK vs non-FDK, and 3-class Healthy-Unhealthy-FDK. Performance measured by Accuracy (Acc), Precision (Pre), Recall (Rec), and $F1$ score ($F1$) weighted metrics for all experiments. Model abbreviations are as follows: using EfficientNet (*EffB0*), using ResNet18 (*Res18*), *HU* stands for Healthy-Unhealthy, *FN* stands for FDK-nonFDK, *HUF* stands for Healthy-Unhealthy-FDK, *A* stands for the models trained in the pre-training phase, *B* represents the fine-tuned models in the semi-supervised pseudo-label stage, and *S* highlights the models trained from scratch. The top-performing models and their corresponding metrics in each experiment are highlighted in bold.

Experiment	Model	Pretraining	Acc	Pre	Rec	F1
Healthy vs Unhealthy	*EffB0-HU-S*	None	58.71	51.35	27.97	36.22
	EffB0-HU-A	ImageNet	84.86	80.59	84.14	82.33
	EffB0-HU-B	***EffB0-HU-A***	**86.43**	**81.86**	**86.86**	**84.29**
	Res18-HU-A	ImageNet	83.20	86.79	70.67	77.90
	Res18-HU-B	*Res18-HU-A*	82.68	86.89	69.08	76.97
FDK vs nonFDK	***EffB0-FN-S***	**None**	**85.81**	**44.16**	**77.82**	**56.35**
	EffB0-FN-A	*ImageNet*	83.58	40.16	80.65	53.62
	EffB0-FN-B	*EffB0-FN-A*	83.20	39.53	80.65	53.05
	Res18-FN-A	ImageNet	81.92	37.62	81.45	51.46
	Res18-FN-B	*Res18-FN-A*	80.78	36.25	83.45	50.55
Healthy Unhealthy FDK	*EffB0-HUF-S*	None	41.53	62.32	41.53	34.34
	EffB0-HUF-A	**ImageNet**	**67.11**	**72.58**	**67.11**	**68.30**
	EffB0-HUF-B	*EffB0-HUF-A*	66.35	72.55	66.16	67.45
	Res18-HUF-A	ImageNet	65.21	71.63	65.21	66.51
	Res18-HUF-B	*Res18-HUF-A*	60.70	70.64	60.70	62.11



(a) *EffB0-HU-B*

(b) *EffB0-FN-S*

(c) *EffB0-HUF-A*

Figure 4: Confusion matrices for the best-performing models on the SS_{te} dataset for our three distinct tasks.

ployed Cohen's Kappa [5] score as a statistical assessment to quantify the inter-annotator agreement. Cohen's Kappa score is formally defined as $K = \frac{P_o - P_e}{1-P_e}$ where P_o is the observed agreement ratio and P_e represents the expected agreement when both annotators assign labels randomly. We calculated the Cohen's Kappa score for every pair of raters for each experiment.

5. Discussion

Fusarium-damaged kernels (FDKs) are detrimental to wheat quality and productivity, resulting from Fusarium head blight infection. Given their significant impact on crop productivity and quality, FDK content is of great interest to wheat breeders and agronomists. In this study, we developed a high-throughput single-kernel screening tool for FDK assessment using automated image acquisition and analysis. Our approach utilizes a semi-supervised learning method on the WheatSeedBelt dataset, which consists of over 40,000 high-resolution RGB images capturing top and side views of individual wheat kernels from 268 genetically-distinct varieties. This dataset is unique in its scale and composition and potentially provides a unique re-

Table 3: Cohen's Kappa scores of the binary Healthy-Unhealthy (left), the binary FDK-nonFDK (middle), and of the 3-class Healthy-Unhealthy-FDK (right) tasks performed on the SS_{ra} dataset.

	MO	**LW**	**SP**
MO	1.0	0.669	0.453
LW		1.0	0.455
SP			1.0

	MO	**LW**	**SP**
MO	1.0	0.669	0.591
LW		1.0	0.633
SP			1.0

	MO	**LW**	**SP**
MO	1.0	0.643	0.393
LW		1.0	0.390
SP			1.0

source for deep model development and benchmarking.

We implemented an image-processing pipeline to efficiently extract the WheatSeedBelt informative features for analysis and modeling. Additionally, we conducted three deep-learning experiments that leveraged the dataset to distinguish healthy, unhealthy but non-FDK, and FDK of wheat. For each experiment, we conducted pretraining and semi-supervised fine-tuning phases to develop EfficientNet-B0 [41] and ResNet18 [13] models on the top-view wheat kernels, SS, extracted operating the image-processing pipeline. Our approach is considered semi-supervised [48] as it uses only a few wheat kernel packets, which includes 63% of our test set, SS_{te}, and only 3% of the whole dataset, SS. We fine-tuned the pretrained models on the pseudo-labeled unannotated subset SS_{un}, which is considered a semi-supervised refinement approach [21].

EfficientNet outperformed ResNet in general, but both models demonstrate promising performance during training, delivering the WheatSeedBelt's practical efficacy for deep learning model development. Fine-tuning the pretrained models did not improve performance, except for the *EffB0-HU-B* model, which exhibited a 2% increase in F1 score compared to the *EffB0-HU-A* model. This highlights that the accuracy of the labels assigned to each sample is critical for the success of deep models on pseudo-labeled datasets, therefore inaccurate labeling can mislead a model that already performs well on the test set. In addition to the substantial model accuracies (Table 2), we observed promising performance levels for both binary tasks (Figures 4a and 4b). Further, the $F1$ score of 68.30 achieved by the model *EffB0-HUF-A*, as illustrated in Figure 4c, indicates the difficulty of the 3-class classification task of identifying Unhealthy samples from Healthy and FDK images, which is supported by the rater reliability study (Table 3).

Given the extensively discussed challenges in achieving high-performance models using a limited expert-annotated dataset for FDK detection, and considering the complexity of the task, we believe further experiments with a larger number of wheat kernel packets, a balanced distribution of samples across classes in the training set, and the exploration of alternative deep learning models might improve model performance in the 3-class classification task. Additionally, the large-scale unannotated WheatSeedBelt dataset could be useful for self-supervised learning, with generative and contrastive learning models potentially improving FDK detection. Machine-learning models could also be developed using the FS dataset to detect kernel health.

We also conducted an inter-rater reliability study to assess the labeling process difficulty of the WheatSeedBelt dataset. To do so, we recruited three experts as raters to label a portion of the SS_{te} set, specifically SS_{ra}, which constituted approximately 10% of the set. To determine the agreement between the raters, we calculated Cohen's Kappa score for all three labeling tasks. Overall, the results indicate a low level of agreement among the raters, illustrating the difficulty of the labeling task, even for human experts in the field (Tables 3). Nevertheless, the findings of the rater reliability study indicated that the differentiation between FDK and non-FDK kernels was reasonably straightforward compared to discrimination between Healthy and Unhealthy samples. Meanwhile, in terms of model performance comparison, it can be inferred that discrimination between Healthy and Unhealthy (with FDK) samples is characterized by a more pronounced disparity in ease, as compared to the performance of models attempting to classify FDK and nonFDKs.

6. Conclusion

In this study, we created a novel and comprehensive dataset consisting of a diverse range of wheat varieties and devised an image processing pipeline to clean and process the raw images and extract informative features that could be utilized for machine and deep learning models. Through this dataset, we developed several semi-supervised models to recognize Healthy, non-FDK unhealthy, and Fusarium damaged wheat kernels. While our models demonstrated promising results for binary classification tasks such as Healthy-Unhealthy and FDK-nonFDK, our 3-class classification task exposed the complexity of distinguishing non-FDK unhealthy samples from healthy and FDK samples. Furthermore, the difficulty of the labeling process was confirmed by our inter-rater reliability analysis, highlighting the difficulty of this classification task, even for human experts. This will require future work to improve classification models using our publicly-available WheatSeedBelt dataset and investigating alternative imaging modalities for FDK classification such as X-ray and multi-spectral imaging.

References

[1] Timo Ahonen, Abdenour Hadid, and Matti Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(12):2037–2041, 2006.

[2] Alexander Buslaev, Vladimir I Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A Kalinin. Albumentations: Fast and flexible image augmentations. *Information*, 11(2):125, 2020.

[3] Olivier Chapelle, Patrick Haffner, and Vladimir N Vapnik. Support vector machines for histogram-based image classification. *IEEE Transactions on Neural Networks*, 10(5):1055–1064, 1999.

[4] Clyde M Christensen and HH Kaufmann. Deterioration of stored grains by fungi. *Annual Review of Phytopathology*, 3(1):69–84, 1965.

[5] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

[6] Etienne David, Mario Serouart, Daniel Smith, Simon Madec, Kaaviya Velumani, Shouyang Liu, Xu Wang, Francisco Pinto, Shahameh Shafiee, Izzat SA Tahir, et al. Global wheat head detection 2021: An improved dataset for benchmarking wheat head detection methods. *Plant Phenomics*, 2021.

[7] Fernando Roberti De Siqueira, William Robson Schwartz, and Helio Pedrini. Multi-scale gray level co-occurrence matrices for texture description. *Neurocomputing*, 120:336–345, 2013.

[8] Edward R Dougherty. An introduction to morphological image processing. In *SPIE*. Optical Engineering Press, 1992.

[9] Nilay Ganatra and Atul Patel. Deep learning methods and applications for precision agriculture. *Machine Learning for Predictive Analysis: Proceedings of ICTIS 2020*, pages 515–527, 2021.

[10] Bo Gong, Daji Ergu, Ying Cai, and Bo Ma. Real-time detection for wheat head applying deep neural network. *Sensors*, 21(1):191, 2020.

[11] Keith Halcro, Kaitlin McNabb, Ashley Lockinger, Didier Socquet-Juglard, Kirstin E Bett, and Scott D Noble. The belt and phenoseed platforms: shape and colour phenotyping of seed samples. *Plant Methods*, 16(1):1–13, 2020.

[12] Robert M Haralick, Karthikeyan Shanmugam, and Its' Hak Dinstein. Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, (6):610–621, 1973.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, Las Vegas, NV, USA, 27-30 June 2016.

[14] David Hughes, Marcel Salathé, et al. An open access repository of images on plant health to enable the development of mobile disease diagnostics. *arXiv preprint arXiv:1511.08060*, 2015.

[15] Gilberto Igrejas and Gérard Branlard. The importance of wheat. *Wheat Quality for Improving Processing and Human Health*, pages 1–7, 2020.

[16] Karta Kaske Kalsa, Bhadriraju Subramanyam, Girma Demissie, Admasu Fanta Worku, and Nigus Gabbiye Habtu. Major insect pests and their associated losses in quantity and quality of farm-stored wheat seed. *Ethiopian Journal of Agricultural Sciences*, 29(2):71–82, 2019.

[17] Pankaj Kumar Kashyap, Sushil Kumar, Ankita Jaiswal, Mukesh Prasad, and Amir H Gandomi. Towards precision agriculture: Iot-enabled intelligent irrigation systems using deep learning neural network. *IEEE Sensors Journal*, 21(16):17479–17491, 2021.

[18] Saeed Khaki, Nima Safaei, Hieu Pham, and Lizhi Wang. Wheatnet: A lightweight convolutional neural network for high-throughput image-based wheat head detection and counting. *Neuro Computing*, 489:78–89, 2022.

[19] Gurdev S Khush. What it will take to feed 5.0 billion rice consumers in 2030. *Plant Molecular Biology*, 59:1–6, 2005.

[20] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[21] Dong-Hyun Lee et al. Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on Challenges in Representation Learning, ICML*, volume 3, page 896. Atlanta, 2013.

[22] Jingbo Li, Changchun Li, Shuaipeng Fei, Chunyan Ma, Weinan Chen, Fan Ding, Yilin Wang, Yacong Li, Jinjin Shi, and Zhen Xiao. Wheat ear recognition based on retinanet and transfer learning. *Sensors*, 21(14):4845, 2021.

[23] Sara Mardanisamani and Mark Eramian. Segmentation of vegetation and microplots in aerial agriculture images: A survey. *The Plant Phenome Journal*, 5(1):e20042, 2022.

[24] Maria Eloisa Mignoni, Aislan Honorato, Rafael Kunst, Rodrigo Righi, and Angélica Massuquetti. Soybean images dataset for caterpillar and diabrotica speciosa pest detection and classification. *Data in Brief*, 40:107756, 2022.

[25] M Nadimi, JM Brown, J Morrison, and J Paliwal. Examination of wheat kernels for the presence of fusarium damage and mycotoxins using near-infrared hyperspectral imaging. *Measurement: Food*, 4:100011, 2021.

[26] Keyhan Najafian, Alireza Ghanbari, Mahdi Sabet Kish, Mark Eramian, Gholam Hassan Shirdel, Ian Stavness, Lingling Jin, and Farhad Maleki. Semi-self-supervised learning for semantic segmentation in images with dense patterns. *Plant Phenomics*, 2022.

[27] Keyhan Najafian, Alireza Ghanbari, Ian Stavness, Lingling Jin, Gholam Hassan Shirdel, and Farhad Maleki. A semi-self-supervised learning approach for wheat head detection using extremely small number of labeled samples. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1342–1351, Montreal, BC, Canada, 11-17 October 2021.

[28] Duc M Nguyen, Mohammad Nazeri, Amirreza Payandeh, Aniket Datar, and Xuesu Xiao. Toward human-like social robot navigation: A large-scale, multi-modal, social human navigation dataset. *arXiv preprint arXiv:2303.14880*, 2023.

[29] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729, Bhubaneswar, India, 16-19 December 2008. IEEE.

[30] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979.

[31] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[32] G Polder, GWAM Van Der Heijden, C Waalwijk, and IT Young. Detection of fusarium in single wheat kernels using spectral imaging. *Seed Science and Technology*, 33(3):655–668, 2005.

[33] Hafiz Tayyab Rauf, Basharat Ali Saleem, M Ikram Ullah Lali, Muhammad Attique Khan, Muhammad Sharif, and Syed Ahmad Chan Bukhari. A citrus fruits and leaves dataset for detection and classification of citrus diseases through machine learning. *Data in Brief*, 26:104340, 2019.

[34] Shivangana Rawat, Akshay L Chandra, Sai Vikas Desai, Vineeth N Balasubramanian, Seishi Ninomiya, and Wei Guo. How useful is image-based active learning for plant organ segmentation? *Plant Phenomics*, 2022, 2022.

[35] Ewa Ropelewska and Piotr Zapotoczny. Classification of fusarium-infected and healthy wheat kernels based on features from hyperspectral images and flatbed scanner images: A comparative analysis. *European Food Research and Technology*, 244:1453–1462, 2018.

[36] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[37] Muhammad A Shahin and Stephen J Symons. Detection of fusarium damaged kernels in canada western red spring wheat using visible/near-infrared hyperspectral imaging and principal component analysis. *Computers and Electronics in Agriculture*, 75(1):107–112, 2011.

[38] Yue Shi, Wenjiang Huang, Juhua Luo, Linsheng Huang, and Xianfeng Zhou. Detection and discrimination of pests and diseases in winter wheat based on spectral indices and kernel discriminant analysis. *Computers and Electronics in Agriculture*, 141:171–180, 2017.

[39] Bekele Shiferaw, Melinda Smale, Hans-Joachim Braun, Etienne Duveiller, Mathew Reynolds, and Geoffrey Muricho. Crops that feed the world 10. past successes and future challenges to the role played by wheat in global food security. *Food Security*, 5:291–317, 2013.

[40] Pierre Soille. Opening and closing. In *Morphological Image Analysis*, pages 105–137. Springer, 2004.

[41] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114, Long Beach, CA, USA, 10-15 June 2019. PMLR.

[42] Stefan Van der Walt, Johannes L Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D Warner, Neil Yager, Emmanuelle Gouillart, and Tony Yu. scikit-image: image processing in python. *PeerJ*, 2:e453, 2014.

[43] D Wang, FE Dowell, and DS Chung. Assessment of heat-damaged wheat kernels using near-infrared spectroscopy. In *2001 ASAE Annual Meeting*, page 1, California, USA, July 30-August 1 2001. American Society of Agricultural and Biological Engineers.

[44] Zijun Zhang. Improved adam optimizer for deep neural networks. In *2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS)*, pages 1–2, Banff, AB, Canada, 04-06 June 2018. IEEE.

[45] Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in Neural Information Processing Systems*, 31, 2018.

[46] Yang-Yang Zheng, Jian-Lei Kong, Xue-Bo Jin, Xiao-Yi Wang, Ting-Li Su, and Min Zuo. Cropdeep: The crop vision dataset for deep-learning-based classification and detection in precision agriculture. *Sensors*, 19(5):1058, 2019.

[47] Lei Zhou, Chu Zhang, Mohamed Farag Taha, Xinhua Wei, Yong He, Zhengjun Qiu, and Yufei Liu. Wheat kernel variety identification based on a large near infrared spectral dataset and a novel deep learning-based feature selection method. *Frontiers in Plant Science*, 11:575810, 2020.

[48] Xiaojin Jerry Zhu. Semi-supervised learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2005.