**A high-throughput phenotyping pipeline for quinoa (*Chenopodium quinoa*) panicles using image analysis with convolutional neural networks**

# ABSTRACT

Quinoa is a grain crop with excellent nutritional properties which attracts global attention for its potential contribution to future food security in a changing climate. Grain yield is affected by panicle traits whose phenotypic analysis is time-consuming and error-prone because of their complex architecture, and automated image analysis is an efficient alternative. We designed a panicle phenotyping pipeline implemented in Python using Mask R-CNN for panicle segmentation and classification. We used it to analyze 5,151 images of quinoa panicles collected over three consecutive seasons from a breeding program in the Peruvian highlands. The pipeline follows a stage-wise approach, which first selects the optimal segmentation model and then another model that best classifies panicle shape. The best segmentation model achieved average precision (mAP) score of 83.16 and successfully extracted panicle length, width, area, and RGB values. The classification model achieved 95% prediction accuracy for panicle shape. Our pipeline enables accurate and cost-effective phenotyping of quinoa panicles. Using automated phenotyping based on deep learning, optimal panicle ideotypes can be selected in quinoa breeding.

**Keywords:** deep learning, genetic resources, genbank phenomics, ImageJ, plant breeding, image analysis

# INTRODUCTION

Phenotyping is a key process in plant breeding programs aimed at developing new varieties of crops with higher yields. However, phenotyping poses a bottleneck in the breeding process due to the labor-intensive nature of the task and high cost. Manual phenotyping may result in reduced accuracy and prolonged timelines in breeding programs. Recent advances in phenotyping technologies, such as the use of neural networks for image analysis, revolutionize plant phenotyping by increasing precision, reducing labor intensity, and enabling the identification of new traits relevant to crop breeding [(Araus et al. 2018; Warman and Fowler 2021)](https://www.zotero.org/google-docs/?6YYpPH). This development of technology not only benefits major crops but also minor and orphan crops because the limited resources available for the improvement of such crops can be compensated by low-cost technological approaches for genetic and phenotypic analyses.

Quinoa (*Chenopodium quinoa* Willd.) is an example of an orphan crop. It originates from the Andean region, where it has played a vital role as a staple food for small-scale farmers in the Andean highlands [(Jacobsen et al. 2003; Hellin and Higman 2005)](https://www.zotero.org/google-docs/?broken=3TTa0R). For thousands of years, farmers have contributed to the domestication and selection of quinoa varieties, resulting in a high genetic and phenotypic diversity [(Bazile et al. 2016a; Patiranage et al. 2022)](https://www.zotero.org/google-docs/?2GmyCZ). The grains of quinoa are highly valued for their nutritional properties and serve as a rich source of macronutrients and energy ([Bhargava et al. 2006](https://www.zotero.org/google-docs/?broken=2ATfg1), [Repo-Carrasco et al. 2003; Nowak et al. 2016; Chandra et al. 2018](https://www.zotero.org/google-docs/?broken=zo2N0a)). Furthermore, quinoa is remarkably resilient to abiotic stressors such as drought and salinity [(reviewed by Grenfell-Shaw and Tester 2021)](https://www.zotero.org/google-docs/?broken=fVs6I3). Due to these benefits, quinoa cultivation has started to spread beyond its native area. Nevertheless, despite the increasing number of countries growing quinoa, production volumes are small compared to the Andean region [(Zurita-Silva et al. 2014; Bazile et al. 2016)](https://www.zotero.org/google-docs/?broken=OFU2vN). Although several breeding programs were established [(Böndel and Schmid 2021)](https://www.zotero.org/google-docs/?cxHhxq), modern quinoa varieties are still little improved compared to ancient landraces. The extensive phenotypic diversity of quinoa in its center of origin reveals a substantial degree of genetic diversity, which likely results from local adaptation and farmer selection [(Bazile et al. 2016](https://www.zotero.org/google-docs/?MBpVOB), [Jarvis et al. 2017)](https://www.zotero.org/google-docs/?tezQZF). This diversity manifests itself in various forms of morphological and physiological variation, including panicle shape, seed color, leaf color [(Bioversity International et al. 2013)](https://www.zotero.org/google-docs/?Zcu0r4), and varying degrees of tolerance to both biotic and abiotic stress factors. Recent efforts have aimed to standardize quinoa evaluation and to characterize its diversity of traits for more effective application in breeding programs [(Stanschewski et al. 2021)](https://www.zotero.org/google-docs/?A6PR2F). However, the prevailing evaluation methods still largely rely on labor-intensive and frequently imprecise manual phenotyping using visual scales and cards.

The inflorescence of quinoa is a panicle, on which individual flowers are arranged in groups on secondary axes [(Wrigley et al. 2015)](https://www.zotero.org/google-docs/?0lFEyZ). The dimensions of the panicles, both in width and length influence grain yield, making them target traits for selection in breeding programs [(Benlhabib et al. 2016a; Maliro et al. 2017; Santis et al. 2018)](https://www.zotero.org/google-docs/?NpuggU). We previously observed that the product of panicle length and panicle width was strongly correlated with yield and showed a high heritability (H2=0.81; [Lozano-Isla et al. 2023)](https://www.zotero.org/google-docs/?broken=IxMvgD). Two types of inflorescences are distinguished [(Tapia et al. 1979)](https://www.zotero.org/google-docs/?axuvj9). The ‘glomerulate’ type panicles develop compact primary axillary internodes and elongated internodes of flower clusters, which results in spherical inflorescences. The ‘amarantiform’ type develops elongated inflorescences similar to amaranth species, and the finger-shaped partial inflorescences originate directly from the main axis. Glomerulate inflorescences are considered to be wild type and are dominant over the amarantiform type [(Gandarillas](https://www.zotero.org/google-docs/?x1nZUx)1974), but the genetic basis of this variation is currently unknown. In the description of quinoa traits, a third type of panicle shape was defined as ‘intermediate’ [(Bioversity International et al. 2013)](https://www.zotero.org/google-docs/?3eLMNs). Since grain yield is influenced by the size and type of the quinoa panicle [(Craine et al. 2023)](https://www.zotero.org/google-docs/?ileCWn), efficient phenotyping of these panicle traits is expected to improve the selection of superior genotypes and enhance genetic gain.

Computational image analysis is revolutionizing plant phenotyping because of its precision and throughput. In particular, neural networks emerge as a potent tool for phenotyping and characterizing crop diversity [(Arya et al. 2022)](https://www.zotero.org/google-docs/?9DCtZE) because they frequently outperform traditional image analysis techniques [(Liu and Wang 2020; Xie et al. 2020; Sabouri et al. 2021; Arya et al. 2022; Kang et al. 2023; Yu et al. 2023)](https://www.zotero.org/google-docs/?JVmMFB). The main objectives of image analysis are classification and image segmentation, which enable the measurement of traits such as size and shape. Additionally, it allows for the counting of features like fruits or color extraction [(Ganesh et al. 2019; Zhou et al. 2019; Jia et al. 2020; Lee and Shin 2020)](https://www.zotero.org/google-docs/?TI8jco). Among available algorithms for neural network-based image analysis, Mask R-Convolutional Neural Networks (Mask R-CNN; [He et al. 2018)](https://www.zotero.org/google-docs/?vc67hD), have been particularly successful, because they allow accurate pixel-wise mask prediction for objects. Mask R-CNN has found widespread applications in robotics [(Jia et al. 2020)](https://www.zotero.org/google-docs/?FHEn3C), medicine [(Anantharaman et al. 2018; Chiao et al. 2019)](https://www.zotero.org/google-docs/?8JFfm8), autonomous driving [(Fujiyoshi et al. 2019)](https://www.zotero.org/google-docs/?TC5x04), and plant science [(Ganesh et al. 2019; Machefer et al. 2020; Jia et al. 2020; Kienbaum et al. 2021)](https://www.zotero.org/google-docs/?vqG0qm). Among classification models that identify and categorize objects within images, neural network architectures such as VGG16 [(Simonyan and Zisserman 2015)](https://www.zotero.org/google-docs/?ceeNQd), InceptionV3 [(Szegedy et al. 2016)](https://www.zotero.org/google-docs/?broken=ZBFkRK), and EfficientNetB0 [(Tan and Le 2020)](https://www.zotero.org/google-docs/?bXptwi) achieve state-of-the-art performance on multiple image classification tasks.

Considering the labor-intensive nature of phenotyping quinoa panicles under field conditions, we created a high-throughput pipeline for extracting phenotypic traits from images of quinoa panicles. This pipeline is based on our previous work ([Kienbaum et al. 2021)](https://www.zotero.org/google-docs/?PGzM9N), which employed Mask R-CNN for the classification and segmentation of maize cob images. In the present study, we adapted and optimized this approach for image classification, specifically for differentiating between glomerulate and amarantiform panicle types. We performed image segmentation to extract trait measurements to estimate quantitative-genetic parameters for yield components related to panicles. Our objective was to create a pipeline that achieves the same or higher level of accuracy as manual phenotyping but with significantly improved throughput, showcasing the effectiveness of state-of-the-art image analysis using Mask R-CNN in characterizing complex panicle traits. First, we describe the development of a deep learning model specifically designed for the classification and segmentation of quinoa panicles using Mask R-CNN. Subsequently, we demonstrate the application of this pipeline in estimating quantitative genetic parameters from a multi-location trial.

# MATERIALS AND METHODS

## Plant material

The plant material used for the experiments was derived from six segregating populations of quinoa (*Chenopodium quinoa* Willd.) originating from single crosses of landraces provided by the germplasm bank of the Universidad Nacional del Altiplano, Puno, Peru [(Lozano-Isla et al. 2023)](https://www.zotero.org/google-docs/?Tumusi). The field trials were conducted in three successive growing seasons from 2016 to 2019 as multi-location trials for line selection (“selection trials”), and variety registration (“registration trials”) trials in the Peruvian Highlands ([Table 1](#tbl-id.r5y16fnovap5)). Images from the 2021-2022 season for production and seed increase (“production trials”) were used for the validation set using scales in the images.

## Panicle images

Pictures were taken during flowering at stage 69 of the BBCH scale [(Sosa-Zuniga et al. 2017)](https://www.zotero.org/google-docs/?RVTuqo). A representative panicle for each experimental unit (breeding lines grown in experimental plots) was selected for photographing under field conditions in front of a blue background with different light conditions ([Fig. 1](#fig-id.6ns3q793e24l), Figure S1). Images were taken with different cameras and resolutions ([Table 1](#tbl-id.r5y16fnovap5)), resulting in an image collection exhibiting a high heterogeneity ([Fig. 1](#fig-id.6ns3q793e24l)a). We excluded images with panicles that showed bird damage, were blurred, or were overly dry from the analysis ([Fig. 1](#fig-id.6ns3q793e24l)b).

## Development of a Mask R-CNN mode for image segmentation

We found it challenging to obtain a single Mask R-CNN model suitable for both panicle segmentation and classification. Mask R-CNN is simple to train and extends Faster R-CNN by adding a branch for predicting an object mask in parallel with the existing branch for bounding box recognition [(Bharati and Pramanik 2020)](https://www.zotero.org/google-docs/?v9oOZn). Consequently, we adopted a stagewise approach, selecting separate models for each task. A pipeline was implemented, connecting the best models of each stage. First, we developed models for image segmentation by testing 16 models and selecting the best one [(He et al. 2018)](https://www.zotero.org/google-docs/?broken=sdiG3s) using the following procedure.

### Sampling and annotation

To ensure the selection of an unbiased training dataset, we created two random samples. The first sample involved selecting 25% from each experiment conducted in different years and locations ([Table 1](#tbl-id.r5y16fnovap5)) making a total of ca. 1,300 images. A random sample of 320 images was taken from this initial pool of images. The sample was divided into training and validation sets with 238 and 82 images respectively. The number of training images was determined based on a previous study of maize cob segmentation ([Kienbaum et al. 2021)](https://www.zotero.org/google-docs/?7qUKEG). The images were annotated using VGG Image Annotator version 2.0.8 [(Dutta and Zisserman 2019)](https://www.zotero.org/google-docs/?FNaCkN). For selection and registration field trials, only the panicle class was annotated ([Table 1](#tbl-id.r5y16fnovap5)). For the production field trial, seven classes were determined for the images: panicle, scale, label, barcode, and each channel from RGB spectra (Figure S2).

### Model training

We utilized the Mask R-CNN framework [(He et al. 2018)](https://www.zotero.org/google-docs/?f4VkHW) and fitted it with the Resnet 101 backend to the seven classes. Model training was conducted on four parallel Tesla K80 GPUs on the BinAC high-performance computing cluster at the University of Tübingen. The Mask R-CNN training parameters were partly selected based on previous work in image analysis for maize cobs [(Kienbaum et al. 2021)](https://www.zotero.org/google-docs/?M5ZDsr), with further variations and improvements implemented as needed. Quinoa panicles exhibit greater detail and diversity in their outline and shape compared to maize cobs. Therefore, we varied parameters such as mask resolution, loss weight, and training configuration (heads.m) to investigate potential differences in the final model performance. This led to training 16 distinct Mask R-CNN models with a learning rate of 1e-4 over 200 epochs, each with different parameters ([Table 2](#tbl-id.vur4v4bpxbe8))*.*

In Mask R-CNN, the mask resolution determines the size of the predicted masks for each object instance. A higher resolution yields a more detailed delineation of object boundaries but increases computational time. We adjusted this parameter from the standard resolution of 28x28 pixels to an enhanced resolution of 56x56. This change required a minor modification in the Mask R-CNN architecture, which involved adding an extra convolutional layer at the end of the mask branch. The aim was to achieve a more precise panicle mask.

Model optimization was based on the loss weight parameter. It ranged from the standard loss weight of 1 for each mask and class (mask01-class01) to an emphasis on mask optimization with a mask loss weight of 10 and a class loss weight of 1 (mask10-class01). We also experimented with classification optimization, adjusting the class loss weight to 10 and mask loss weight to 1 (mask01-class10). The parameter heads.m indicated the training configuration, which either involved training all layers of the Resnet-101 architecture or fine-tuning only the head layers while freezing all other layers \*\*([Table 2](#tbl-id.vur4v4bpxbe8))*.*

The performance of the segmentation model in detecting panicles was assessed using the Intersection over Union (IoU) score, also known as the Jaccard index [(Jaccard 1901)](https://www.zotero.org/google-docs/?ZU4rg8). This metric is widely used in evaluating the performance of object detection. The effectiveness of our trained models was measured using the mean average precision (mAP) over different IoU thresholds ranging from 50 to 95% in 5% increments (AP@[IoU = 0.50:0.95]; [Kienbaum et al. 2021)](https://www.zotero.org/google-docs/?5VKQZc).

### Model development for image classification

In the second stage, three deep learning architectures were implemented and tested for model classification using VGG16, InceptionV3, and EfficientNetB0 to differentiate between the two panicle types. VGG16 has the advantage of accurate identification and performs better on large-scale data sets and complex background recognition tasks [(Wang 2020)](https://www.zotero.org/google-docs/?a3QIal). InceptionV3 is characterized by its multi-level feature extraction and factorization techniques, achieving a balance between accuracy and computational efficiency [(Li et al. 2021)](https://www.zotero.org/google-docs/?oFtTZ7). The efficientNetB0 compound scaling approach offers high performance while maintaining reduced computational complexity, which is useful for real-time applications and resource-constrained environments [(Ramamurthy et al. 2023)](https://www.zotero.org/google-docs/?hQirCz).

### Sampling

A total of 320 panicle masks from the segmentation pipeline and pictures from the experiments were randomly selected and divided into two groups: training and validation. Each group contained two panicle classes: glomerulate and amarantiform. The training group included 110 images of the amarantiform and 142 of the glomerulate shape, whereas the validation group included 30 and 36 images of the same respective panicle classes. The imbalance in the number of panicles in the datasets primarily reflects the predominance of glomerulate panicles over amarantiform panicles.

### Model training

To classify images into two classes of panicle shapes, namely amarantiform and glomerulate, we implemented 12 models using various combinations of convolutional neural network architectures, dense layers, and activation functions ([Table 3](#tbl-id.ddv8lvk6bvgs)). Each model was constructed using a specific neural network architecture such as VGG16, InceptionV3, or EfficientNetB0, and included two dense layers. The first layer employed a ReLU (Rectified Linear Unit) activation function and utilized either 128 or 1024 dense layers. The choice to vary the number of neurons was influenced by computational resource limitations and the need for efficient model utilization. The second layer used either a sigmoid or softmax output activation function for image classification [(Maharjan et al. 2020)](https://www.zotero.org/google-docs/?w1TwxQ). We employed a standard image augmentation technique, commonly used to artificially increase the size and diversity of the image dataset, for a more robust model. This involved horizontal flips, random crops, random Gaussian blur, varied contrast, brightness, zoom, translation, and rotation using the *imgaug* library in Python [(Jung 2022)](https://www.zotero.org/google-docs/?I3XYpc). The models were executed under three replications for 200 epochs ([Table 3](#tbl-id.ddv8lvk6bvgs)). The ModelCheckpoint function was used to automatically save the model exhibiting the highest performance, based on the lowest validation loss. The prediction accuracy (%) for the two panicle classes, amarantiform and glomerulate, was evaluated considering the three factors: neural network architectures, dense layers, and activation functions.

## Pipeline for quinoa panicle image analysis

The best segmentation model, selected based on the mAP score, was then used to export the pixel-wise mask of panicles from each image and to extract the phenotypic traits: panicle, scale, label, barcode, and RGB values by channel ([Fig. 2](#fig-id.n3ualh2wrwq)c-d). The panicle masks were submitted to the best classification model to identify the panicle shape as either amarantiform or glomerulate ([Fig. 2](#fig-id.n3ualh2wrwq)e-f).

For the first image dataset, selection, and registration trials, the length, width, area in pixels, and RGB values were extracted. The barcode and image scale were included for the second image dataset, the production trial. Since the images of the first dataset did not contain scales that would allow us to calculate absolute measurements. We calculated two indices related to grain yield [(Lozano-Isla et al. 2023)](https://www.zotero.org/google-docs/?Ec4Hcs) based on panicle length and width (i.e. panicle width/length and length/width) with the aim of determining if there are differences in genetic parameters such as heritability. The pipeline ran on a workstation with 64GB RAM and 8 CPU kernels with a total of 16 threads.

## Deep learning pipeline versus manual annotation

Fifteen panicle images from the selection and registration field trials ([Table 1](#tbl-id.r5y16fnovap5)) were randomly selected and manually measured using ImageJ [(Schneider et al. 2012)](https://www.zotero.org/google-docs/?cvj5Cb) for comparison with the results from the segmentation model. The measures extracted using ImageJ were panicle area, width, and length (all measured in pixels), and mean and standard deviation of each RGB channel (Figure S3). The results were compared using Pearson correlation analysis. To extract the RGB value from the mask, the RGB Measure plugin was installed from <https://imagej.nih.gov/ij/plugins/rgb-measure.html>.

## Quantitative genetic analysis of panicles in multilocation trials

Due to the imbalance in the number of entries between years ([Table 1](#tbl-id.r5y16fnovap5)), the indices were analyzed using a stagewise approach [(Schmidt et al. 2019; Buntaran et al. 2020)](https://www.zotero.org/google-docs/?dc86UO) under multi-location trials with three locations. In stage 1, a linear mixed model with lines as the fixed effect was used for each experiment under the lattice design [(Zystro et al. 2018)](https://www.zotero.org/google-docs/?N045rh). This was done to estimate the Best Linear Unbiased Estimators (BLUEs) using the following model:

where is the response variable of the genotype in the block of the replicate, is the first-stage intercept, is the effect for the ith genotype in the first stage, is the effect of the replicate, is the effect of the incomplete block of the replicate, and is the plot error effect corresponding to .

In stage 2, a linear mixed model for line by environment interaction, where the lines were treated as fixed effects. This was done to calculate the Best Linear Unbiased Predictors (BLUPs) using the following model:

where is the adjusted mean of the genotype in the location obtained in the first stage, is the intercept, is the main effect for the location, is the main effect of the genotype, is the genotype x location interaction effect and is the error of the mean obtained in the first stage.

## Statistical analysis

To conduct statistical analyses and produce graphs, we used the statistical programming language R version 4.4.0 [(R Core Team 2020)](https://www.zotero.org/google-docs/?MGsNq5), while image analysis was performed using Python version 3.7.

The image analysis models were compared with different parameters (i.e., segmentation and classification), and each model with its respective replicas was subjected to a variance analysis with Anova(type = “III”) in the *car* package [(Fox et al. 2023)](https://www.zotero.org/google-docs/?FsGail). The factors that showed significance were subjected to a means comparison analysis using the Tukey comparison test (p<0.05) implemented in the package *emmeans* [(Lenth et al. 2023)](https://www.zotero.org/google-docs/?HhrrDU) and *multcomp* [(Hothorn et al. 2023)](https://www.zotero.org/google-docs/?1I0o3y).

For the multi-environment analyses, broad-sense heritability (i.e., Cullis approach), variance components, Best Linear Unbiased Estimators (BLUEs), and Best Linear Unbiased Predictors (BLUPs) were estimated based on the ‘H2cal()’ function of the ‘inti’ package [(Lozano-Isla 2021)](https://www.zotero.org/google-docs/?xxhkmJ). This function uses a linear mixed model for both random and fixed effects for genotypes, based on the *lme4* package [(Bates et al. 2020)](https://www.zotero.org/google-docs/?IDpQRC). Outlier removal for multi-location trials was based on method 4, Bonferroni-Holm, using re-scaled median absolute deviation for standardizing residuals as described in Bernal-Vasquez et al. [(2016)](https://www.zotero.org/google-docs/?knvdO2) and implemented in the same function.

A Pearson correlation plot was used to compare the trait predictions made with Mask R-CNN and ImageJ and was performed using the *psych* R package [(Revelle 2021)](https://www.zotero.org/google-docs/?qtObKc). The plots were produced with the package *ggplot2* [(Wickham et al. 2023)](https://www.zotero.org/google-docs/?liMTUD) and *ggside* [(Landis 2022)](https://www.zotero.org/google-docs/?CzxJLJ). Code and reproducible data analysis were implemented under Quarto, an open-source scientific and technical publishing system [(Allaire et al. 2022, Supplementary File 1)](https://www.zotero.org/google-docs/?4aG49l).

# RESULTS

## Training models for segmentation and classification

To implement a pipeline for quinoa panicle analysis, a comparative analysis of various segmentation and classification models was conducted. The selection of the best segmentation model involved the assessment of 16 Mask R-CNN models, each tested with a combination of parameters, mask resolution, loss weight, and neural network. The accuracy of the segmentation models was evaluated using the mean average precision (mAP) metric ([Table 2](#tbl-id.vur4v4bpxbe8)). The selection of the best classification model entailed the comparison of 12 classification models that were developed through different combinations of neural network architectures, dense layers, and activation functions. The accuracy of the classification models was assessed based on the predictive performance for two classes (i.e. amarantiform and glomerulate, [Table 3](#tbl-id.ddv8lvk6bvgs)).

The segmentation models showed statistical significance in the interaction of the loss weight, mask resolution, and neural network (p = 0.01, [Table 2](#tbl-id.vur4v4bpxbe8), Supplementary Table 1). The mAP scores ranged from 0.68 to 0.8, with a coefficient of variation of 3.65%. The model segmentation-09 achieved the highest mAP score of 0.8, while the model segmentation-08 obtained the lowest score at 0.69 ([Table 2](#tbl-id.vur4v4bpxbe8)). The segmentation model selected for implementation in the pipeline was chosen based on the replica that achieved the highest mAP with 0.83 (Supplementary Table 2).

For classification, the prediction accuracy results for the glomerulate and amarantiform classes ranged from 0.85 to 0.92. No significant differences were observed among the factors studied in the classification models ([Table 3](#tbl-id.ddv8lvk6bvgs), Supplementary Table 3). Therefore, we selected model classification-08 with the highest accuracy of 0.95 in the repeated model evaluations for further analyses (Supplementary Table 4).

## Pipeline for quinoa panicle phenotyping

We implemented a Python pipeline to extract phenotypic traits from quinoa panicles, leveraging the best model for segmentation and classification analysis ([Fig. 2](#fig-id.n3ualh2wrwq)). To segment panicle images, we used the model segmentation-09 ([Fig. 2](#fig-id.n3ualh2wrwq)a-d). As certain images featured multiple panicles, we generated individual masks for each panicle. Subsequently, these masks were employed in the classification process. The segmentation model’s pixel-wise masking of each panicle allowed us to extract panicle length, width, area, and mean RGB color values for each channel. The segmentation of panicles was followed by the input of these segmented structures into the classification model. Subsequently, the model categorized the images based on one of two panicle shapes: glomerulate or amarantiform ([Fig. 2](#fig-id.n3ualh2wrwq)e-f). We combined the results from the segmentation and classification pipeline ([Fig. 2](#fig-id.n3ualh2wrwq)g) and used them to calculate quantitative-genetic parameters such as variance components, BLUEs, BLUPs, and heritability. The results generated by this pipeline can be utilized to carry on association studies, such as Genome-Wide Association Studies (GWAS), or Quantitative Trait Locus (QTL) mapping ([Fig. 2](#fig-id.n3ualh2wrwq)h).

## ImageJ versus deep learning pipeline

To assess the performance of the deep learning pipeline, we conducted a regression analysis comparing manual trait extraction using ImageJ software with the results obtained from our model pipeline ([Fig. 3](#fig-id.s67t3260nbfo)).

We selected a total of 60 images, with fifteen images from each experiment ([Table 1](#tbl-id.r5y16fnovap5)), and manually annotated them with ImageJ (Figure S1) for panicle length, width, and area. The regression (R2) for panicle length was 0.93 (r = 0.96), for width 0.92 (r = 0.96), and area 0.99 (r = 0.99) showing a high correlation ([Fig. 3](#fig-id.s67t3260nbfo)a-c). Panicle distribution for width and area shows similar distributions between manual annotation and model pipeline.

The classification model achieved an accuracy of 97.6%. The sensitivity was 96.7%, representing the proportion of panicles correctly identified as panicle-positive by the model ([Fig. 3](#fig-id.s67t3260nbfo)e). The specificity of the model was 100%, indicating the proportion of individuals without panicles correctly identified as panicle-negative ([Fig. 3](#fig-id.s67t3260nbfo)f). The model’s precision, or the proportion of positive predicted values, was 100%.

## Scaled images and deep learning pipeline

To evaluate the efficiency of the pipeline under field conditions, we updated the segmentation model to include a dataset with 106 images with scale and QR codes ([Fig. 4](#fig-kix.n7evacder2cp)). We took photos during the seed production campaign using a 10 cm scale ([Table 1](#tbl-id.r5y16fnovap5), [Fig. 4](#fig-kix.n7evacder2cp)f).

The regression analysis between the manual measurement and the model prediction did not show a correlation for the panicle length (R2 = <0.01, [Fig. 4](#fig-kix.n7evacder2cp)a). In contrast, panicle width showed a significant correlation (r = 0.73, R2 = 0.54, [Fig. 4](#fig-kix.n7evacder2cp)b). A medium correlation was presented for the panicle indices. For width/length (r = 0.6, R2 = 0.36 [Fig. 4](#fig-kix.n7evacder2cp)c) and width/length (r = 0.56, R2 = 0.31, [Fig. 4](#fig-kix.n7evacder2cp)d). The classification model achieved an accuracy of 98.1%. The sensitivity was 100% and the model’s specificity was 94.1% ([Fig. 4](#fig-kix.n7evacder2cp)e).

## Quantitative-genetic analysis from the multi-location trials

To assess the effectiveness of the pipeline within the framework of breeding programs. We computed the quantitative-genetic parameters by utilizing the extracted phenotypic values of quinoa panicles through the pipeline ([Fig. 2](#fig-id.n3ualh2wrwq)). We conducted a stagewise analysis under a multi-location trial for the registration trials ([Table 1](#tbl-id.r5y16fnovap5)). In the first stage, we calculated the adjusted means. In the second stage, we performed a model selection to calculate the BLUPs, choosing the model with the lowest AIC (Supplementary Table 5). In the absence of scales in the images from the registration trials, we proceeded to estimate the broad-sense heritability for two indices, specifically derived from the length and width measurements of the panicles ([Fig. 5](#fig-id.e5f491agsax)).

The predominant panicle shape was glomerulate ([Fig. 5](#fig-id.e5f491agsax)a). Its frequency was 82% among F7 genotypes (Illpa site), while in the F8 generation, its frequency was 84% (Camacani) and 63 % (Illpa). ANOVA shows a significant interaction between panicle shape and its length-to-width ratio (p < 0.001). Glomerulate panicles were found to be longer than they were in width, whereas in the case of amaranthiform panicles, a tendency to be wider and less long ([Fig. 5](#fig-id.e5f491agsax)b).

For the quantitative genetic parameters, the highest heritability was observed for the ratio of panicle width to length (H² = 0.61), while panicle length to width showed a heritability of 0.53 ([Fig. 5](#fig-id.e5f491agsax)c). The evaluated traits revealed significant genetic variance, with values of 48.8 and 58.3 for the ratios of length to width and width to length, respectively. The panicle length-to-width ratio exhibited a better normal distribution compared to the width-to-length ratio ([Fig. 5](#fig-id.e5f491agsax)d).

Panicle shape is a bimodal trait (i.e., amaranthiform and glomerulate), and it does not exhibit a normal distribution ([Fig. 5](#fig-id.e5f491agsax)d). To compare the heritability using linear and generalized mixed models, we analyzed the results with a linear mixed model (lmer) and a generalized linear mixed model with a binomial distribution (glmer; Supplementary Table 8). For lmer, the standard heritability was 0.66, while for the glmer model it was 0.69. However, as we cannot calculate the error variance for the glmer model, this results in a 4% difference between the outcomes (Supplementary Table 6). Based on this result, we decided to proceed with the analysis using lmer.

Using the mixed linear model, we observed high repeatability for panicle shape in the F7 season in Illpa when experimental units had replicates (Supplementary Table 7). In the second phase of the analysis, using the dataset from the experiments in Illpa and Camacani for the F8 season, we found that heritability decreased to 0.24 when experimental units did not have replicates ([Fig. 5](#fig-id.e5f491agsax)c, Supplementary Table 7). However, the standard heritability for panicle type is 0.63 (H2.s, Supplementary Table 8).

# DISCUSSION

Panicle traits are important components for increasing yield in quinoa. However, field-based panicle phenotyping is time-consuming and labor-intensive, limiting its use in breeding programs and large-scale genetic studies such as QTL mapping or GWAS. To address this limitation, we aimed to develop a pipeline for high-throughput phenotyping of quinoa panicles using deep learning-based image analysis. We built a pipeline that selects the best models using a stagewise approach. In the first stage, we tested 16 different models to identify the most suitable one for segmenting panicle images. In the second stage, we implemented 12 classification models to categorize images into two panicle shapes: amaranthiform and glomerulate. To verify the accuracy of the pipeline, we compared the output of tested models with manual evaluations using ImageJ. The pipeline output demonstrated a high correlation with manual evaluations at a pixel-based scale for segmentation and a high prediction accuracy for quinoa panicle shape. We applied the complete pipeline to calculate quantitative-genetic parameters in multi-location trials, exemplifying its implementation for breeding programs.

## Pipeline to panicle image analysis in quinoa

Despite training various Mask R-CNN models with different parameters, our pipeline was not able to identify a single model that performed best in both classification and segmentation tasks simultaneously. Therefore, we decided to implement a stagewise approach selecting a separate model for each task for panicle phenotyping.

In a previous study on maize cob image analysis [(Kienbaum et al. 2021)](https://www.zotero.org/google-docs/?broken=yRotfR), a single Mask R-CNN model reliably predicted the classes and masks of maize cobs and ruler elements. The features of maize cobs and rulers were simpler and more distinguishable than the different quinoa panicles with the presence of leaves, similar colors between structures (i.e., stem, leaves, and panicle), and in some cases, the panicles not exhibiting compact density. The need to choose two models probably relies on the difficulty of optimizing both for the challenging classification between the panicle shapes and for accurately segmenting the detailed panicle traits with the masks. We selected a segmentation model that had a score of 83.16 mAP with two classes, a similar result to what Kienbaum et al., (2021) found with two classes at 87.7 mAP in maize (Kienbaum et al. 2021), and with two classes at 89.85 mAP in strawberries [(Yu et al. 2019)](https://www.zotero.org/google-docs/?broken=9RK4g5).

Effective detection using a simple image processing method can be challenging under field conditions due to various environmental such as background color, image position, shadow, and image rotation. [Lee and Shin (2020)](https://www.zotero.org/google-docs/?jYD0VN) managed to detect potato shapes under field conditions and calculated parameters based on pixels using Mask R-CNN. We adopted a similar approach in this study to detect and analyze quinoa panicles under field conditions, showing that Mask R-CNN is a suitable method to reduce the tedious and costly labor associated with traditional approaches. Zhou et al. ([2019](https://www.zotero.org/google-docs/?broken=4oBFnw)) analyzed 1,064 panicles from 272 genotypes of sorghum and extracted the area, volume, and panicle length and width, showing a high correlation with manual annotation. We obtained similar results with our approach when we compared it with manual annotation ([Fig. 3](#fig-id.s67t3260nbfo)).

Training a model for image segmentation, compared to image classification, requires annotation of both class and object location information in images for model training, which can be a time-consuming task. In our dataset, many photos did not contain complete panicles, because they were taken under field conditions by students (and not experienced photographers), suggesting the need for adjustments in image acquisition protocols and better training of personnel taking photographs. Additionally, the images from selection and registration trials did not have any scale, so we had to rely on panicle indices based on relative rather than trait values for our analysis. For panicle classification, we did not include intermediate panicle shapes, which are challenging to differentiate, even with visual scoring [(Craine et al. 2023)](https://www.zotero.org/google-docs/?broken=qdD8f8).

## Quantitative-genetic analysis of multi-location trials

The primary objective of our pipeline was to facilitate high-quality phenotyping of quinoa panicles, enabling reliable estimation of quantitative-genetic parameters such as variance components and heritability. This information facilitates the prediction of responses to selection in breeding programs ([Visscher et al. 2008)](https://www.zotero.org/google-docs/?broken=5k2JkI).

To minimize measurement errors and improve estimates, it is necessary to investigate a large number of genotypes. For these reasons, automatic and semi-automatic methods have been proposed for analyzing panicles in rice and sorghum ([Zhou et al. 2019; Kong and Chen 2021](https://www.zotero.org/google-docs/?broken=9B5U5d)). In both cases, the panicles were removed from the plants and analyzed under controlled conditions, leading to increased labor and cost. Our pipeline robustly analyzes images taken during crop development in a non-destructive way under variable field conditions, differing for example for light, angle, shadows, background color, and photographic devices, all of which create variation in image quality.

Phenotyping and quantifying panicle traits are crucial as they correlate with the yield of quinoa. Several studies have found a positive correlation between grain yield and both panicle length and panicle width [(Benlhabib et al. 2016b; Maliro et al. 2017; Santis et al. 2018)](https://www.zotero.org/google-docs/?YBAqhL). We used the ratio between panicle Width and Length, which exhibits a 0.86 heritability at maturity (Lozano et al. 2022). Efficient estimation of quantitative-genetic parameters for panicle traits could enable more efficient selection and increase the selection gain for yield in future breeding populations.

According to quinoa descriptors, there are three different types of quinoa panicles: amarantiform, intermediate, and glomerulate ([Bioversity International, 2013)](https://www.zotero.org/google-docs/?broken=DQfUPk). Intermediate panicle types are challenging to distinguish, even for the human eye. For this reason, we classified the panicles into the two most distinct types, glomerulate and amarantiform. Future analysis also might include panicle density, described as lax versus compact, as it could be a valuable trait for breeding programs and is subject to environmental effects ([Manjarres-Hernández et al. 2021](https://www.zotero.org/google-docs/?broken=wGvoWb)). The selection of amarantiform panicles can promote increased airflow throughout the quinoa inflorescences leading to a reduced incidence of pests and diseases. The integration of panicle shape and density has the potential to enhance the panicle ideotype in quinoa.

A significant limitation of our analysis is the absence of a scale in the images taken in the selection and registration field trials. This absence prevented us from using absolute measurements, such as panicle length and width in centimeters, for quantitative genetic analysis. Instead, we had to depend on a relative index from the ratio between panicle length and width. We included a few images from the seed production trial and updated the pipeline with images that include barcode and scale. The information on the pictures was extracted from the barcodes and the scales to extract the panicle traits in centimeters. This implementation significantly streamlines image processing and subsequent analysis. The updated model using scales was not able to predict the real length with accuracy, maybe due to two main reasons, (i) the few numbers of images included in the model training, and (ii) the scale was not distributed around the picture frame during the picture acquisition. Despite this, our pipeline is now equipped for large-scale analysis of panicle images that include panicles, barcodes, and scales.

However, taking the panicle images correctly in the field could be time-consuming than scoring the two shapes visually than taking images and running the pipeline. Nevertheless, parameters such as area, color, and leaflet in the panicles could be difficult to score visually. Color is typically considered a qualitative trait, the ability to extract RGB values could prove useful for genetic analysis, as demonstrated for seed color in peanuts by Zhang et al. ([2020](https://www.zotero.org/google-docs/?broken=0pQ0tY)). The use of high-throughput phenotyping using deep learning opens new opportunities for the characterization and evaluation of the diversity in large numbers of panicle images in quinoa. Implementation of this approach in plant breeding programs can increase phenotyping efficiency and improve the selection of future quinoa varieties with better panicle traits.

In conclusion, we developed a pipeline to be directly applicable to any set of panicle images without requiring prior model training. It is user-friendly, easy to execute, and does not demand expensive computational resources such as GPUs. For this reason, it is highly suitable for quinoa breeders with limited resources.

# REFERENCES

[Allaire JJ, Teague C, Scheidegger C, et al (2022) Quarto: open-source scientific and technical publishing system built on Pandoc](https://www.zotero.org/google-docs/?a3Kwz3)

[Anantharaman R, Velazquez M, Lee Y (2018) Utilizing Mask R-CNN for Detection and Segmentation of Oral Diseases. In: 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp 2197–2204](https://www.zotero.org/google-docs/?a3Kwz3)

[Araus JL, Kefauver SC, Zaman-Allah M, et al (2018) Translating High-Throughput Phenotyping into Genetic Gain. Trends Plant Sci 23:451–466. https://doi.org/10.1016/j.tplants.2018.02.001](https://www.zotero.org/google-docs/?a3Kwz3)

[Arya S, Sandhu KS, Singh J, kumar S (2022) Deep learning: as the new frontier in high-throughput plant phenotyping. Euphytica 218:47. https://doi.org/10.1007/s10681-022-02992-3](https://www.zotero.org/google-docs/?a3Kwz3)

[Bates D, Maechler M, Bolker B, Walker S (2020) lme4: Linear mixed-effects models using “eigen” and S4](https://www.zotero.org/google-docs/?a3Kwz3)

[Bazile D, Jacobsen S-E, Verniau A (2016a) The Global Expansion of Quinoa: Trends and Limits. Front Plant Sci 7:. https://doi.org/10.3389/fpls.2016.00622](https://www.zotero.org/google-docs/?a3Kwz3)

[Bazile D, Pulvento C, Verniau A, et al (2016b) Worldwide Evaluations of Quinoa: Preliminary Results from Post International Year of Quinoa FAO Projects in Nine Countries. Front Plant Sci 7:. https://doi.org/10.3389/fpls.2016.00850](https://www.zotero.org/google-docs/?a3Kwz3)

[Benlhabib O, Boujartani N, Maughan PJ, et al (2016a) Elevated Genetic Diversity in an F2:6 Population of Quinoa (*Chenopodium quinoa*) Developed through an Inter-ecotype Cross. Front Plant Sci 7:. https://doi.org/10.3389/fpls.2016.01222](https://www.zotero.org/google-docs/?a3Kwz3)

[Benlhabib O, Boujartani N, Maughan PJ, et al (2016b) Elevated Genetic Diversity in an F2:6 Population of Quinoa (Chenopodium quinoa) Developed through an Inter-ecotype Cross. Front Plant Sci 7:](https://www.zotero.org/google-docs/?a3Kwz3)

[Bernal-Vasquez A-M, Utz H-F, Piepho H-P (2016) Outlier detection methods for generalized lattices: a case study on the transition from ANOVA to REML. Theor Appl Genet 129:787–804. https://doi.org/10.1007/s00122-016-2666-6](https://www.zotero.org/google-docs/?a3Kwz3)

[Bharati P, Pramanik A (2020) Deep Learning Techniques—R-CNN to Mask R-CNN: A Survey. In: Das AK, Nayak J, Naik B, et al. (eds) Computational Intelligence in Pattern Recognition. Springer, Singapore, pp 657–668](https://www.zotero.org/google-docs/?a3Kwz3)

[Bioversity International, Fundación para la Promoción e Investigación de Productos Andinos, Instituto Nacional de Innovación Agropecuaria y Forestal, et al (2013) Descriptors for quinoa (*Chenopodium quinoa* Willd) and wild relatives. Bioversity International](https://www.zotero.org/google-docs/?a3Kwz3)

[Böndel KB, Schmid KJ (2021) Quinoa Diversity and Its Implications for Breeding. In: Schmöckel SM (ed) The Quinoa Genome. Springer International Publishing, Cham, pp 107–118](https://www.zotero.org/google-docs/?a3Kwz3)

[Buntaran H, Piepho H-P, Schmidt P, et al (2020) Cross-validation of stagewise mixed-model analysis of Swedish variety trials with winter wheat and spring barley. Crop Sci 60:2221–2240. https://doi.org/10.1002/csc2.20177](Buntaran%20H,%20Piepho%20H-P,%20Schmidt%20P,%20et%20al%20(2020)%20Cross-validation%20of%20stagewise%20mixed-model%20analysis%20of%20Swedish%20variety%20trials%20with%20winter%20wheat%20and%20spring%20barley.%20Crop%20Sci%2060:2221–2240.%20https://doi.org/10.1002/csc2.20177)

[Chiao J-Y, Chen K-Y, Liao KY-K, et al (2019) Detection and classification the breast tumors using mask R-CNN on sonograms. Medicine (Baltimore) 98:e15200. https://doi.org/10.1097/MD.0000000000015200](https://www.zotero.org/google-docs/?a3Kwz3)

[Craine EB, Davies A, Packer D, et al (2023) A comprehensive characterization of agronomic and end-use quality phenotypes across a quinoa world core collection. Front Plant Sci 14:1101547. https://doi.org/10.3389/fpls.2023.1101547](https://www.zotero.org/google-docs/?a3Kwz3)

[Dutta A, Zisserman A (2019) The VIA Annotation Software for Images, Audio and Video. In: Proceedings of the 27th ACM International Conference on Multimedia. Association for Computing Machinery, New York, NY, USA, pp 2276–2279](https://www.zotero.org/google-docs/?a3Kwz3)

[Fox J, Weisberg S, Price B, et al (2023) car: Companion to Applied Regression](https://www.zotero.org/google-docs/?a3Kwz3)

[Fujiyoshi H, Hirakawa T, Yamashita T (2019) Deep learning-based image recognition for autonomous driving. IATSS Res 43:244–252. https://doi.org/10.1016/j.iatssr.2019.11.008](https://www.zotero.org/google-docs/?a3Kwz3)

[Gandarillas H (1974) Genética y origen de la quinua. Instituto Nacional del Trigo, La Paz, Bolivia](https://www.zotero.org/google-docs/?a3Kwz3)

[Ganesh P, Volle K, Burks TF, Mehta SS (2019) Deep Orange: Mask R-CNN based Orange Detection and Segmentation. IFAC-Pap 52:70–75. https://doi.org/10.1016/j.ifacol.2019.12.499](https://www.zotero.org/google-docs/?a3Kwz3)

[He K, Gkioxari G, Dollár P, Girshick R (2018) Mask R-CNN. ArXiv170306870 Cs](https://www.zotero.org/google-docs/?a3Kwz3)

[Hothorn T, Bretz F, Westfall P, et al (2023) multcomp: Simultaneous Inference in General Parametric Models](https://www.zotero.org/google-docs/?a3Kwz3)

[Jaccard P (1901) Etude comparative de la distribution florale dans une portion des Alpes et des Jura. Bull Soc Vaudoise Sci Nat 37:547–579](https://www.zotero.org/google-docs/?a3Kwz3)

[Jarvis DE, Ho YS, Lightfoot DJ, et al (2017) The genome of *Chenopodium quinoa*. Nature 542:307–312. https://doi.org/10.1038/nature21370](https://www.zotero.org/google-docs/?a3Kwz3)

[Jia W, Tian Y, Luo R, et al (2020) Detection and segmentation of overlapped fruits based on optimized mask R-CNN application in apple harvesting robot. Comput Electron Agric 172:105380. https://doi.org/10.1016/j.compag.2020.105380](https://www.zotero.org/google-docs/?a3Kwz3)

[Jung A (2022) imgaug: Image augmentation library for deep neural networks](https://www.zotero.org/google-docs/?a3Kwz3)

[Kang F, Li J, Wang C, Wang F (2023) A Lightweight Neural Network-Based Method for Identifying Early-Blight and Late-Blight Leaves of Potato. Appl Sci 13:1487. https://doi.org/10.3390/app13031487](https://www.zotero.org/google-docs/?a3Kwz3)

[Kienbaum L, Correa Abondano M, Blas R, Schmid K (2021) DeepCob: precise and high-throughput analysis of maize cob geometry using deep learning with an application in genebank phenomics. Plant Methods 17:91. https://doi.org/10.1186/s13007-021-00787-6](https://www.zotero.org/google-docs/?a3Kwz3)

[Landis J (2022) ggside: Side Grammar Graphics](https://www.zotero.org/google-docs/?a3Kwz3)

[Lee H-S, Shin B-S (2020) Potato Detection and Segmentation Based on Mask R-CNN. J Biosyst Eng 45:233–238. https://doi.org/10.1007/s42853-020-00063-w](https://www.zotero.org/google-docs/?a3Kwz3)

[Lenth RV, Bolker B, Buerkner P, et al (2023) emmeans: Estimated Marginal Means, aka Least-Squares Means](https://www.zotero.org/google-docs/?a3Kwz3)

[Li Y, Feng X, Liu Y, Han X (2021) Apple quality identification and classification by image processing based on convolutional neural networks. Sci Rep 11:16618. https://doi.org/10.1038/s41598-021-96103-2](https://www.zotero.org/google-docs/?a3Kwz3)

[Liu J, Wang X (2020) Tomato Diseases and Pests Detection Based on Improved Yolo V3 Convolutional Neural Network. Front Plant Sci 11:](https://www.zotero.org/google-docs/?a3Kwz3)

[Lozano-Isla F (2021) inti: Tools and Statistical Procedures in Plant Science](https://www.zotero.org/google-docs/?a3Kwz3)

[Lozano-Isla F, Apaza J-D, Mujica Sanchez A, et al (2023) Enhancing quinoa cultivation in the Andean highlands of Peru: a breeding strategy for improved yield and early maturity adaptation to climate change using traditional cultivars. Euphytica 219:26. https://doi.org/10.1007/s10681-023-03155-8](https://www.zotero.org/google-docs/?a3Kwz3)

[Machefer M, Lemarchand F, Bonnefond V, et al (2020) Mask R-CNN Refitting Strategy for Plant Counting and Sizing in UAV Imagery. Remote Sens 12:3015. https://doi.org/10.3390/rs12183015](https://www.zotero.org/google-docs/?a3Kwz3)

[Maharjan S, Alsadoon A, Prasad PWC, et al (2020) A novel enhanced softmax loss function for brain tumour detection using deep learning. J Neurosci Methods 330:108520. https://doi.org/10.1016/j.jneumeth.2019.108520](https://www.zotero.org/google-docs/?a3Kwz3)

[Maliro MFA, Guwela VF, Nyaika J, Murphy KM (2017) Preliminary Studies of the Performance of Quinoa (*Chenopodium quinoa* Willd.) Genotypes under Irrigated and Rainfed Conditions of Central Malawi. Front Plant Sci 8:. https://doi.org/10.3389/fpls.2017.00227](https://www.zotero.org/google-docs/?a3Kwz3)

[Patiranage DS, Rey E, Emrani N, et al (2022) Genome-wide association study in quinoa reveals selection pattern typical for crops with a short breeding history. eLife 11:e66873. https://doi.org/10.7554/eLife.66873](https://www.zotero.org/google-docs/?a3Kwz3)

[R Core Team (2020) R: A language and environment for statistical computing. Vienna, Austria](https://www.zotero.org/google-docs/?a3Kwz3)

[Ramamurthy K, Thekkath RD, Batra S, Chattopadhyay S (2023) A novel deep learning architecture for disease classification in Arabica coffee plants. Concurr Comput Pract Exp 35:e7625. https://doi.org/10.1002/cpe.7625](https://www.zotero.org/google-docs/?a3Kwz3)

[Revelle W (2021) psych: Procedures for Psychological, Psychometric, and Personality Research](https://www.zotero.org/google-docs/?a3Kwz3)

[Sabouri H, Sajadi SJ, Jafarzadeh MR, et al (2021) Image processing and prediction of leaf area in cereals: A comparison of artificial neural networks, an adaptive neuro-fuzzy inference system, and regression methods. Crop Sci 61:1013–1029. https://doi.org/10.1002/csc2.20373](https://www.zotero.org/google-docs/?a3Kwz3)

[Santis GD, Ronga D, Caradonia F, et al (2018) Evaluation of two groups of quinoa (*Chenopodium quinoa* Willd.) accessions with different seed colours for adaptation to the Mediterranean environment. Crop Pasture Sci 69:1264–1275. https://doi.org/10.1071/CP18143](https://www.zotero.org/google-docs/?a3Kwz3)

[Schmidt P, Hartung J, Rath J, Piepho H-P (2019) Estimating Broad-Sense Heritability with Unbalanced Data from Agricultural Cultivar Trials. Crop Sci 59:525–536. https://doi.org/10.2135/cropsci2018.06.0376](https://www.zotero.org/google-docs/?a3Kwz3)

[Schneider CA, Rasband WS, Eliceiri KW (2012) NIH Image to ImageJ: 25 years of image analysis. Nat Methods 9:671–675. https://doi.org/10.1038/nmeth.2089](https://www.zotero.org/google-docs/?a3Kwz3)

[Simonyan K, Zisserman A (2015) Very Deep Convolutional Networks for Large-Scale Image Recognition](https://www.zotero.org/google-docs/?a3Kwz3)

[Sosa-Zuniga V, Brito V, Fuentes F, Steinfort U (2017) Phenological growth stages of quinoa (*Chenopodium quinoa*) based on the BBCH scale. Ann Appl Biol 171:117–124. https://doi.org/10.1111/aab.12358](https://www.zotero.org/google-docs/?a3Kwz3)

[Stanschewski CS, Rey E, Fiene G, et al (2021) Quinoa Phenotyping Methodologies: An International Consensus. Plants 10:1759. https://doi.org/10.3390/plants10091759](https://www.zotero.org/google-docs/?a3Kwz3)

[Tan M, Le QV (2020) EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks](https://www.zotero.org/google-docs/?a3Kwz3)

[Tapia M, Gandarillas H, Alandia S, et al (1979) La quinua y la kañiwa: cultivos andinos. Centro Internacional de Investigaciones para el Desarrollo (CIID), Instituto Interamericano de Ciencias Agricolas (IICA), Bogotá](https://www.zotero.org/google-docs/?a3Kwz3)

[Wang H (2020) Garbage Recognition and Classification System Based on Convolutional Neural Network VGG16. In: 2020 3rd International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE). pp 252–255](https://www.zotero.org/google-docs/?a3Kwz3)

[Warman C, Fowler JE (2021) Deep learning-based high-throughput phenotyping can drive future discoveries in plant reproductive biology. Plant Reprod 34:81–89. https://doi.org/10.1007/s00497-021-00407-2](https://www.zotero.org/google-docs/?a3Kwz3)

[Wickham H, Chang W, Henry L, et al (2023) ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics](https://www.zotero.org/google-docs/?a3Kwz3)

[Wrigley CW, Corke H, Seetharaman K, Faubion J (2015) Encyclopedia of Food Grains. Academic Press](https://www.zotero.org/google-docs/?a3Kwz3)

[Xie X, Ma Y, Liu B, et al (2020) A Deep-Learning-Based Real-Time Detector for Grape Leaf Diseases Using Improved Convolutional Neural Networks. Front Plant Sci 11:](https://www.zotero.org/google-docs/?a3Kwz3)

[Yu D, Zha Y, Sun Z, et al (2023) Deep convolutional neural networks for estimating maize above-ground biomass using multi-source UAV images: a comparison with traditional machine learning algorithms. Precis Agric 24:92–113. https://doi.org/10.1007/s11119-022-09932-0](https://www.zotero.org/google-docs/?a3Kwz3)

[Zhou Y, Srinivasan S, Mirnezami SV, et al (2019) Semiautomated Feature Extraction from RGB Images for Sorghum Panicle Architecture GWAS. Plant Physiol 179:24–37. https://doi.org/10.1104/pp.18.00974](https://www.zotero.org/google-docs/?a3Kwz3)

[Zystro J, Colley M, Dawson J (2018) Alternative Experimental Designs for Plant Breeding. In: Plant Breeding Reviews. John Wiley & Sons, Ltd, pp 87–117](https://www.zotero.org/google-docs/?a3Kwz3)

|  |
| --- |
| Fig. 1: Different types of heterogeneity in the image dataset of quinoa panicles. (a) Types of variation included in the analysis: Variation of lightning conditions, background, and panicle frame in the pictures. (b) Types of variation excluded in the analysis: Bird-damaged, blurred images, and overdry panicles were excluded from the dataset to avoid bias during training and image analysis. |

|  |
| --- |
| Fig. 2: Quinoa panicle image analysis pipeline using deep learning for segmentation and classification (a) Pictures from the field experiments (b) Image annotation using via software for segmentation model training (c) Model training for image segmentation (d) Best model for segmentation and trait extraction (e) Model training for image classification (f) Best model for classification (g) Merge of phenotypic data extracted from images (h) Data analysis and further application. The asterisk denotes analyses that were not conducted in the present study. |

|  |
| --- |
| Fig. 3: Comparison between ImageJ and deep learning pipeline. (a-b) Regression analysis for panicle length, width, and area. (e) ROC curve for average true and false positive rate. ROC curve for sensitivity and specificity. Fifteen images were taken from the selection and registration trials (n = 60). |

|  |
| --- |
| Fig. 4: Model prediction ability with images with scale under field conditions. (a-d) Regression analysis for panicle traits and indices. (e) ROC curve for sensitivity and specificity. (f) Panicle picture under field conditions with scale and QR code. The pictures were taken during the seed production trials (n = 106). |

|  |
| --- |
| Fig. 5: Quantitative-genetic analysis from quinoa panicle traits under multi-location trials. (a) Panicle shape distribution across generation and location. (b) Panicle relation between the shape and the ratio length and width. (c) Variance partition and heritability from panicle traits in quinoa. (d) Trait distribution after second stage analysis for panicle shape and indices. Values are represented by the BLUPs (n = 548). |

Table 1: Images from multi-environmental trials from 2016 to 2019 employed for image analysis using high-throughput phenotyping for quinoa panicles. Experiments were conducted under different experimental designs (i.e. lattice and randomized complete blocks - RCBD) with presence of unbalanced data. Devices and image resolution were different for each experiment. The images from seasons 2021-2022 were used for model validation with the use of a scale in the images.

| **Trial** | **Season** | **Genotypes** | **Generation** | **Location** | **Exp. design** | **Device** | **Resolution** | **Pictures** | **Scale** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Selection | 2016-2017 | 1200 | F6 | Camacani, Puno | RCBD | Nikon D5101 | 2736 x 3648 | 3862 | no |
| Registration | 2017-2018 | 600 | F7 | Illpa, Puno | Lattice 10x10 | ZTE Blade A610 | 2448 x 3264 | 1240 | no |
| Registration | 2018-2019 | 25 | F8 | Camacani, Puno | Lattice 5x5 | Samsung SM-T285M | 1440 x 2560 | 25 | no |
| Registration | 2018-2019 | 25 | F8 | Illpa, Puno | Lattice 5x5 | Samsung SM-T285M | 1440 x 2560 | 25 | no |
| Production | 2021-2022 | 57 | F10 | Camacani, Puno | RCBD | Motorola G30 | 3456 x 4608 | 108 | yes |
|  | | | | | | | | | |

Table 2: Model performance for 16 segmentation models using Mask R-CNN for extracting the phenotypic information using different mask resolution, loss weight, and neural network parameters. Values represented by Least-squares means (ap595) and standard error (ste). Significance (sig) was estimated based on Tukey (p<0.05). Each model was run with five replications.

| **Model** | **loss weight** | **mask resolution** | **heads.m** | **ap595** | **ste** | **sig** |
| --- | --- | --- | --- | --- | --- | --- |
| segmentation-1 | mask1class10 | 28x28 | all | 0.763 | 0.012 | a |
| segmentation-2 | mask20class1 | 28x28 | all | 0.754 | 0.012 | a |
| segmentation-3 | mask10class1 | 28x28 | all | 0.747 | 0.012 | a |
| segmentation-4 | mask1class1 | 28x28 | all | 0.738 | 0.012 | a |
| segmentation-5 | mask1class1 | 28x28 | heads | 0.734 | 0.012 | a |
| segmentation-6 | mask20class1 | 28x28 | heads | 0.727 | 0.012 | ab |
| segmentation-7 | mask10class1 | 28x28 | heads | 0.725 | 0.012 | ab |
| segmentation-8 | mask1class10 | 28x28 | heads | 0.686 | 0.012 | b |
| segmentation-9 | mask10class1 | 56x56 | all | 0.801 | 0.012 | a |
| segmentation-10 | mask20class1 | 56x56 | all | 0.8 | 0.012 | a |
| segmentation-11 | mask1class1 | 56x56 | all | 0.791 | 0.012 | a |
| segmentation-12 | mask1class10 | 56x56 | all | 0.782 | 0.012 | a |
| segmentation-13 | mask1class1 | 56x56 | heads | 0.758 | 0.012 | a |
| segmentation-14 | mask1class10 | 56x56 | heads | 0.751 | 0.012 | a |
| segmentation-15 | mask10class1 | 56x56 | heads | 0.737 | 0.012 | ab |
| segmentation-16 | mask20class1 | 56x56 | heads | 0.704 | 0.012 | b |

Table 3: Model performance for 12 classification models using three different architectures, dense layers, and activation function configurations for binary image classification for quinoa panicles in amarantiform and glomerulate panicle shapes. Values represented by Least-squares means (accuracy) and standard error (ste). Significance (sig) was estimated based on Tukey (p<0.05). Each model was run with five replications.

| **Model** | **architecture** | **dense layers** | **activation function** | **accuracy** | **ste** | **sig** |
| --- | --- | --- | --- | --- | --- | --- |
| classification-01 | InceptionV3 | 128 | sigmoid | 0.917 | 0.013 | a |
| classification-02 | InceptionV3 | 128 | softmax | 0.871 | 0.013 | a |
| classification-03 | InceptionV3 | 1024 | sigmoid | 0.909 | 0.015 | a |
| classification-04 | InceptionV3 | 1024 | softmax | 0.924 | 0.015 | a |
| classification-05 | VGG16 | 128 | sigmoid | 0.909 | 0.015 | a |
| classification-06 | VGG16 | 128 | softmax | 0.909 | 0.015 | a |
| classification-07 | VGG16 | 1024 | sigmoid | 0.901 | 0.019 | a |
| classification-08 | VGG16 | 1024 | softmax | 0.909 | 0.015 | a |
| classification-09 | EfficientNetB0 | 128 | sigmoid | 0.909 | 0.015 | a |
| classification-10 | EfficientNetB0 | 128 | softmax | 0.901 | 0.019 | a |
| classification-11 | EfficientNetB0 | 1024 | sigmoid | 0.901 | 0.019 | a |
| classification-12 | EfficientNetB0 | 1024 | softmax | 0.854 | 0.015 | a |