SUPPLEMENTARY MATERIALS
to the article E.G. Komyshev, M.A. Genaev, I.D. Busov, M.V. Kozhekin, N.V. Artemenko, A.Y. Glagoleva, V.S. Koval, D.A. Afonnikov "Determination of the melanin and anthocyanin content in barley grains by digital image analysis using machine learning methods"

Supplementary Materials 1 and 2 contain the characteristics of barley accessions used in the work: ID, name, presence of the melanin and anthocyanins, the presence or absence of the hull. Accessions with a distinct blue and purple grain shell colors are additionally mentioned where possible, since such characterization is difficult for samples that accumulate anthocyanins and melanin simultaneously. The anthocyanins color was not considered in the analysis.

Supplementary Material 1. Barley accessions with dark colored grain shells (pigmented) and grains without pigments for training and test dataset

Accession ID	Accession name	Melanin presence	Anthocyanins presence	Hull presence
1	OWB-Rec	No	No	Hulless
2	BLP-h (ICG)	Yes	No	Hulled
3	BLP-n (ICG)	Yes	No	Hulless
4	Bowman (ICG)	No	No	Hulled
5	BA (ICG)	No	Yes (cyan)	Hulless
6	PLP (ICG)	No	Yes (purple)	Hulled
7	OWB18	Yes	Yes	Hulless
8	gBLP (ICG)	Yes	No	Hulled
9	Purple Black (ICG)	Yes	Yes (purple)	Hulled
10	OWB1	Yes	Yes	Hulled
11	OWB5	Yes	Yes	Hulless
12	OWB7	Yes	Yes	Hulled
13	OWB8	No	Yes (cyan)	Hulless
14	OWB17	No	Yes (purple)	Hulless
15	OWB20	No	Yes (purple)	Hulled
16	OWB36	Yes	Yes	Hulled
17	OWB39	No	Yes (purple)	Hulled
18	OWB49	No	Yes (cyan)	Hulled
19	OWB53	No	Yes (purple)	Hulless
20	OWB59	No	Yes (purple)	Hulless
21	OWB62	Yes	Yes	Hulled
22	OWB64	Yes	No	Hulled
23	OWB69	Yes	No	Hulled
24	OWB85	No	Yes (cyan)	Hulled
25	OWB90	No	Yes (purple)	Hulled
26	OWB91	Yes	Yes	Hulless
27	OWB93	No	Yes (cyan)	Hulled
28	k-5144 (VIR)	Yes	No	Hulled
29	k-5445 (VIR)	Yes	No	Hulled
30	k-5447 (VIR)	Yes	No	Hulled
31	k-20040 (VIR)	Yes	No	Hulless
32	k-20048 (VIR)	Yes	No	Hulless
33	k-23441 (VIR)	Yes	No	Hulled
34	k-3502 (VIR)	Yes	No	Hulled
35	k-12265 (VIR)	Yes	No	Hulled
36	k-15151 (VIR)	Yes	No	Hulled
37	k-31075 (VIR)	Yes	No	Hulless
38	k-23378 (VIR)	Yes	No	Hulless
39	k-23306 (VIR)	Yes	No	Hulled
40	k-30675 (VIR)	Yes	No	Hulless
41	k-30676 (VIR)	Yes	No	Hulled

Supplementary Material 1 (end)

Accession ID	Accession name	Melanin presence	Anthocyanins presence	Hull presence
Wh_1	OWB2	No	No	Hulless
Wh_2	OWB13	No	No	Hulled
Wh_3	OWB26	No	No	Hulled
Wh_4	OWB45	No	No	Hulless
Wh_5	OWB82	No	No	Hulless
Wh_6	k-13273 (VIR)	No	No	Hulless
Wh_7	k-17527 (VIR)	No	No	Hulled
Wh_8	k-17805 (VIR)	No	No	Hulled
Wh_9	k-26321 (VIR)	No	No	Hulled
Wh_10	Stepnoe (ICG)	No	No	Hulled
Wh_11	Grace (ICG)	No	No	Hulled
Wh_12	Podarok Sibiri (ICG)	No	No	Hulled
Wh_13	Moskovskiy 121 (ICG)	No	No	Hulled
Wh_14	Svetik (ICG)	No	No	Hulled
Wh_15	Raid (ICG)	No	No	Hulled
Wh_16	Zolotnik (ICG)	No	No	Hulled
Wh_17	Lipen (ICG)	No	No	Hulled
Wh_18	Viner (ICG)	No	No	Hulled
Wh_19	Tanay (ICG)	No	No	Hulled
Wh_20	Golden Promise (ICG)	No	No	Hulled
Wh_21	AL (ICG)	No	No	Hulled
Wh_22	Keystone (ICG)	No	No	Hulled
Wh_23	Biome (ICG)	No	No	Hulled
Wh_24	Lun’ (ICG)	No	No	Hulled
Wh_25	Omskiy golozernyi 2 (ICG)	No	No	Hulless
Wh_26	Signal (ICG)	No	No	Hulled
Wh_27	Novosibirsky 80 (ICG)	No	No	Hulled
Wh_28	Kolchan (ICG)	No	No	Hulled
Wh_29	Archekas (ICG)	No	No	Hulless
Wh_30	Sobolek (ICG)	No	No	Hulled
Wh_31	L-421 (ICG)	No	No	Hulled
Wh_32	Morex (ICG)	No	No	Hulled
Wh_33	Belogorsky (ICG)	No	No	Hulled
Wh_34	Dagestansky (ICG)	No	No	Hulled
Wh_35	Krasnoyarsky 1 (ICG)	No	No	Hulled
Wh_36	Talan (ICG)	No	No	Hulled
Wh_37	Alei (ICG)	No	No	Hulled
Wh_38	Impulse (ICG)	No	No	Hulled

Supplementary Material 2. Barley accessions with different combinations of pigments in the grain shell for the holdout dataset

Accession ID	Accession name	Melanin presence	Anthocyanins presence	Hull presence
60	k-3282 (VIR)	Yes	No	Hulless
67	k-17384 (VIR)	Yes	Yes	Hulled
68	k-17554 (VIR)	Yes	Yes	Hulless
69	k-18830 (VIR)	Yes	No	Hulled
75	k-20024 (VIR)	Yes	No	Hulled
76	k-20028 (VIR)	No	Yes (cyan)	Hulless
85	k-20079 (VIR)	No	Yes (cyan)	Hulless
105	k-3477 (VIR)	Yes	Yes	Hulless
110	k-15160 (VIR)	No	Yes	Hulled
118	k-19068 (VIR)	Yes	Yes	Hulled
122	k-25872 (VIR)	Yes	Yes	Hulless
124	k-26750 (VIR)	Yes	No	Hulled
161	k-26311 (VIR)	Yes	No	Hulled
162	k-11356 (VIR)	Yes	No	Hulled
174	k-22591 (VIR)	No	Yes (purple)	Hulless
181	k-26105 (VIR)	Yes	Yes	Hulled
185	k-22603 (VIR)	No	No	Hulled
196	k-3302 (VIR)	Yes	Yes	Hulled
197	k-13241 (VIR)	Yes	No	Hulled
200	k-176 (VIR)	Yes	No	Hulled
201	k-1067 (VIR)	Yes	Yes	Hulled
Wh_31	L-421 (ICG)	No	No	Hulled
Wh_32	Morex (ICG)	No	No	Hulled
Wh_33	Belogorsky (ICG)	No	No	Hulled
Wh_34	Dagestansky (ICG)	No	No	Hulled
Wh_35	Krasnoyarsky (ICG)	No	No	Hulled
Wh_36	Talan (ICG)	No	No	Hulled
Wh_37	Alei (ICG)	No	No	Hulled
Wh_38	Impulse	No	No	Hulled

Supplementary Material 3. Barley grain imaging process using the proposed protocol

Supplementary Material 4. Example image of a sample with manually applied markings of grains (red) and Petri dish boundaries (green). The marking of the grain area (red) was used to train the segmentation model.

Supplementary Material 5. Data stratification

The following stratification was used for the image datasets: all images of one accession were used only in the same subsample (training, validation or test). To balance the distribution of multi-classes in each subsample, an iterative stratification algorithm was used, implemented in the iterative_train_test_split() method of the model_selection python module of the skmultilearn library (Szymański, Kajdanowicz, 2017). Data on the division of accessions into subsamples are presented in two tables below.

Distribution of barley accessions in the training, validation, test datasets

Subsample	Barley accessions included in the dataset
Training	32, 24, Wh_10, 28, Wh_25, 35, Wh_29, 41, Wh_07, 21, 17, Wh_26, Wh_06, Wh_17, 27, Wh_14, Wh_32, Wh_24, Wh_04, Wh_21, 34, 05, 04, 15, Wh_36, Wh_08, 01, Wh_03, 14, Wh_02, 22, 40, 39, 30, 02, 07, 26, 20, 38, Wh_28, 36, Wh_35, Wh_33, Wh_09, 10, 25, Wh_34
Validation	09, Wh_27, 29, 19, Wh_20, Wh_05, 37, 16, Wh_22, Wh_19, Wh_23, Wh_38, 18, 03, 08, Wh_18
Test	Wh_37, 06, 23, Wh_11, Wh_01, 11, Wh_15, Wh_13, 13, 31, 12, 33, Wh_16, Wh_31, Wh_30, Wh_12

Distribution of barley accessions in the training, validation, test and holdout datasets by the presence of pigments

Training	Validation	Test	Holdout	Parameter
47 (60 %)	16 (20 %)	16 (20 %)	29	Number of accessions
26 (63 %)	8 (20 %)	7 (17 %)	20	Number of accessions with pigment
21 (55 %)	8 (21 %)	9 (24 %)	9	Number of accessions without pigments
17 (61 %)	6 (21 %)	5 (18 %)	16	Number of accessions with melanin
12 (60 %)	4 (20 %)	4 (20 %)	12	Number of accessions with anthocyanins
5 (56 %)	2 (22 %)	2 (22 %)	8	Number of accessions with melanins and anthocyanins

Supplementary Material 6

Extraction of color descriptors based on the image grain region

To describe the color characteristics of grains, we used color representation in the form of four models: RGB, HSV, Lab, and YCrCb (Gowda, Yuan, 2018; Komyshev et al., 2020). Each of them represents color by the intensity of three components. The component values of one space can be obtained by transformations of the component values of the other. Color descriptors were calculated independently for each of these components.

The first type of descriptors: average component intensity values for the pixels of an individual grain. For this, the mean color of the grain pixels was initially calculated. Next, the standard deviation of the component was calculated and pixels whose component value differing by more than 3 standard deviations from the mean were discarded. For the remaining pixels, the mean component value was again calculated and used as color feature for further analysis.

The next type of color features is the global color histogram. It is estimated by partitioning the three-dimensional color space into three-dimensional bins whose size is the same and is determined by the number of bins per component, d. This results in a total number of d^3 bins. Each bin is defined by indices ($i, j, k = 1, \ldots, d$) corresponding to the three components and intensity intervals on each component within ($i\times b_i$, $(i+1)\times b_i$; $j\times b_j$, $(j+1)\times b_j$; $k\times b_k$, $(k+1)\times b_k$), where b_c is the bin size on component c. We used two types of histograms with $d = 4$ (total 64 descriptors) and $d = 8$ (total 256 descriptors). The value of each bin corresponds to the number of grain pixels falling within the intensity ranges by component.

Descriptor of dominant colors

A dominant color descriptor provides a compact description of the representative colors in an image or image region (Cieplinski, 2001).

A descriptor is defined as:

$$F = \{\{c_i, p_i, v_i\}, s\}, (i = 1, \ldots, N),$$

where N is the number of dominant colors. Each dominant color c_i, is a vector of corresponding color space components (e. g., 3D vector of RGB space) obtained by clustering all pixels of the considered image region in their color space and then sampling a representative (dominant) color of the cluster. For this purpose, the average color of the cluster was used. The value of p_i, normalized to values between 0 and 1, is the fraction of pixels in the image or image region corresponding to color cluster c_i. The color variance v_i, describes the color variability of the color cluster for the corresponding representative (dominant) color. The optional spatial coherence s is a number representing the overall spatial homogeneity of the dominant colors in the image.

CNN-based pigment composition classification models

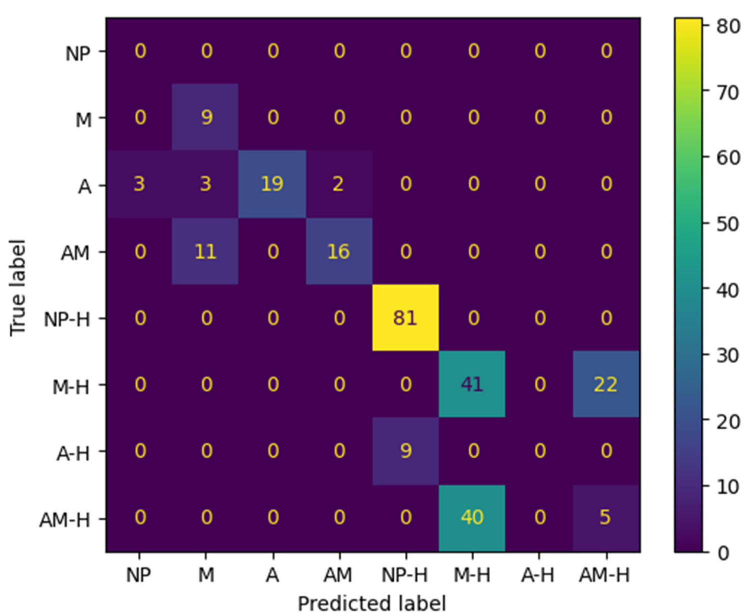
In our work, we used transfer learning: the initial values of the network weights dataset were obtained by training the network on the ImageNet dataset. The models were trained in full depth, and there were no frozen layers. To increase the training sample size and enhance generalization, image augmentations were used, by algorithms implemented in the Albumentations library (Buslaev et al., 2020). Transformations to images from the training sample were applied with a probability of 0.5 and included: vertical and horizontal reflections using HorizontalFlip() and VerticalFlip() methods; brightness and contrast changes using RandomBrightnessContrast(brightness_limit=0.2, contrast_limit=0.2) method.

Supplementary Material 7. Architectural details and training parameters of the segmentation model and grain pigment composition classification models

Task	Model designation	Encoder architecture	Loss function	Metric	Number of training epochs	Batch size	Initial learning rate
Grain region segmentation	U-Net	ResNet-18	Dice	IoU	10	4	lr = 5e–5
Grain region classification	ResNet-18	ResNet-18	BCEWithLogitsLoss + Sigmoid	Accuracy	10	16	lr = 5e–5
Image segmentation and classification	U-Net+ClassHead	EfficientNetB0	BCEWithLogitsLoss, DiceLoss + Sigmoid	IoU (segmentation), F-score и Accuracy (classification)	10	2	lr = 1e–4
2-channel image segmentation	U-Net+ClassSegment	ResNet34	DiceLoss + Sigmoid	Modernized IoU	10	4	lr = 1e–4

Supplementary Material 8. Error matrix for determining the pigmentation type of grain images

A, AM, M, NP denote the precence of anthocyanins, anthocyanins and melanins, melanins and the absence of both pigments, respectively. H denotes the hulled grains.



References

Buslaev A., Iglovikov V.I., Khvedchenya E., Parinov A., Druzhinin M., Kalinin A.A. Albumentations: fast and flexible image augmentations. *Information*. 2020;11(2):125. DOI 10.3390/info11020125

Cieplinski L. MPEG-7 color descriptors and their applications. In: International Conference on Computer Analysis of Images and Patterns. Springer, Berlin, Heidelberg, 2001;11-20

Gowda S.N., Yuan C. ColorNet: Investigating the importance of color spaces for image classification. In: Asian Conference on Computer Vision, Springer, Cham., 2018;581-596

Komyshev E.G., Genaev M.A., Afonnikov D.A. Analysis of color and texture characteristics of cereals on digital images. *Vavilov Journal of Genetics and Breeding*. 2020;24(4):340-347. DOI 10.18699/VJ20.626

Szymański P., Kajdanowicz T. A scikit-based Python environment for performing multi-label classification. *arXiv preprint arXiv*. 2017;1702.01460