Распознавание генотипа колоса пшеницы методами компьютерного зрения

Паулиш Анна*, **Пронозин А., Комышев Е., Генаев М.**

** e-mail: apaulish99@gmail.com*

Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия

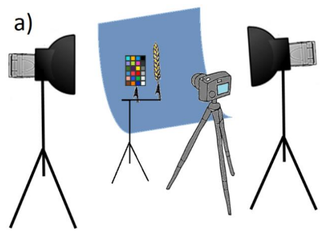
Федеральный исследовательский центр институт цитологии и генетики СО РАН, Новосибирск, Россия

Курчатовский геномный центр федерального исследовательского центра институт цитологии и генетики СО РАН, Новосибирск, Россия

Пшеница и ее сородичи являются важнейшими мировыми продовольственными культурами. Современные виды пшениц произошли от дикого предка в результате нескольких раундов удвоения генома: среди них есть как диплоидные (1x), так тетраплоидные (2x) и гексаплоидные (3x) представители. Мягкая пшеница, например, является гексаплоидом и содержит три генома, A,B,D. Геномный состав (плоидность) служит одним из основных классифицирующих признаков видов пшениц. Геномный состав можно устанавливать молекулярными методами, а также на основе сравнения морфологических характеристик растений. В настоящей работе разработан метод классификации пшеницы по типу плоидности на основе изображения колоса с использованием методов компьютерного зрения. Для анализа использовались 3603 изображения колосьев, 2344 из них относились к гексаплоидным видам, а 1259 к тетраплоидным.

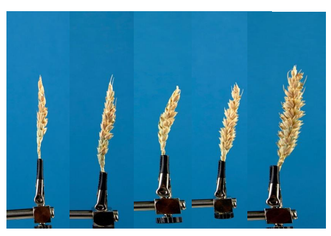
Протоколы получения изображений

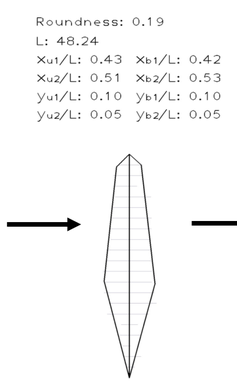
На столе. 1 проекция.



На прищепке. 4 проекции.

Первый подход: с помощью алгоритма [1] были извлечены характеристики колоса, далее, используя методы машинного обучения, были обучены модели логистической регрессии, случайный лес и градиентный бустинг для предсказания типов плоидности





Feature extraction

ML

Classification

Тетраплоид

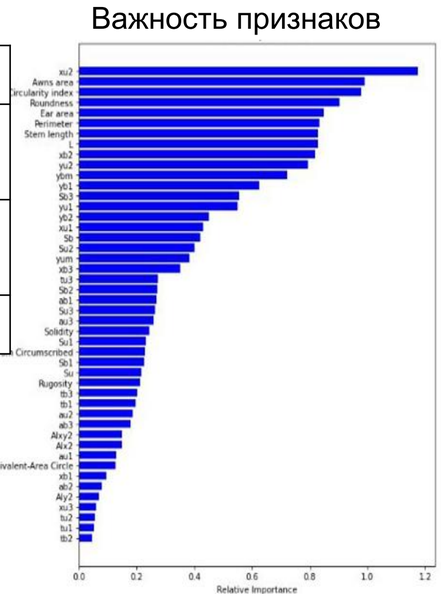
Гексаплоид

Сравниваем точность на одних и тех же данных, которые были предварительно нормированы. В качестве меры точности мы выбрали AUC. Кроме того, ранжировали признаки по значимости.

AUC (Area Under Curve)

Метод	Train	Valid	Holdout
Logistic Regression	0.77	0.70	0.72
Random Forest	1.0	0.83	0.82
Boosting	0.99	0.83	0.85

Важность признаков

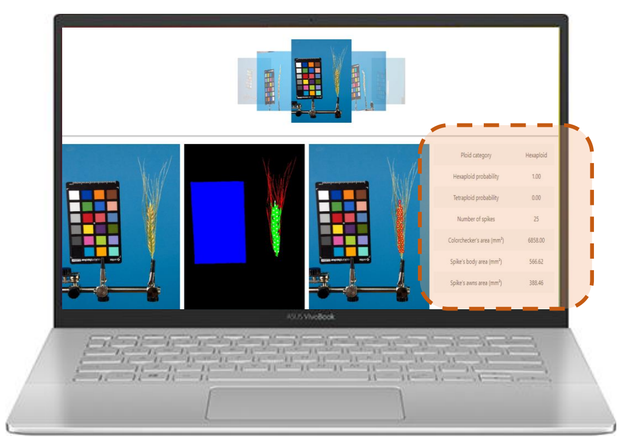


Стандартная нормировка

$$z = \frac{x - \mu}{\sigma}$$

$\mu = Mean$

$\sigma = Standard\ Deviation$



https://spikecv-demo.sysbio.ru

Тетраплоиды

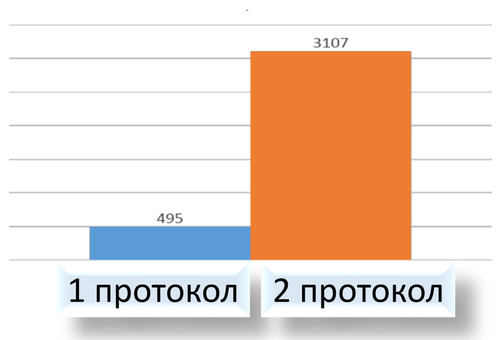
Гексаплоиды

Всего 3603 растений с 644 уникальными посевными номерами.
Растения получены в течении 2015 по 2018 годы в теплицах ИЦиГ СО РАН.

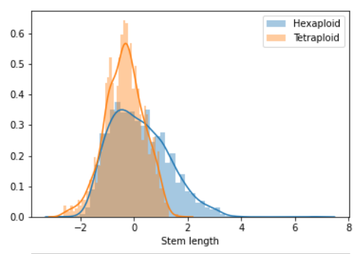
- обучающая выборка (train) – 60%
- валидационная выборка (valid) – 20%
- отложенная выборка (hold out) – 20%

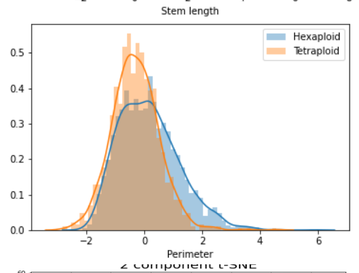
20 видов растений:
- 10 гексаплоидных
- 10 тетраплоидных

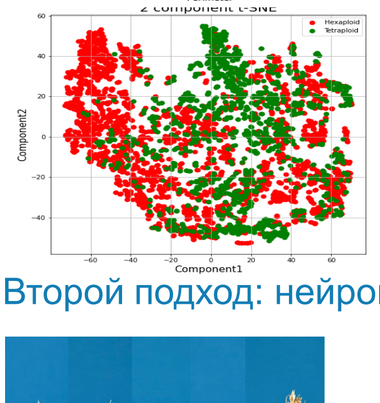
Уникальных генотипов: 496
Уникальных вегетаций: 10



Результаты кластеризации топ 10 признаков методом t-SNE. В целом, большая плоидность дает большие размеры растений и большее разнообразие признаков

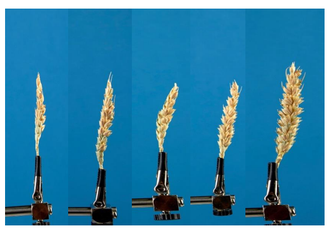




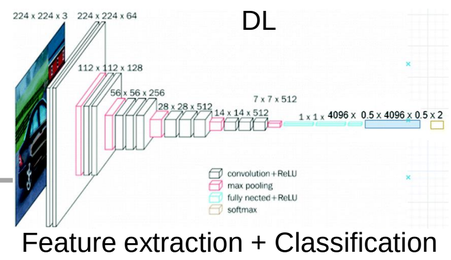


Name	F-statistic	p-value	Disp Hexaploid	Disp Tetraploid
Awns area	0,376	1,000	1,415	3,763
Circularity index	1,188	0,065	0,959	0,807
Roundness	1,828	0,000	1,312	0,718
Perimeter	1,570	0,000	1,080	0,688
Stem length	3,500	0,000	1,320	0,377
xu2	3,928	0,000	1,336	0,340
L	3,500	0,000	1,320	0,377
xb2	4,437	0,000	1,331	0,300
yu2	4,275	0,000	2,491	0,583

Второй подход: нейронные сети



Input



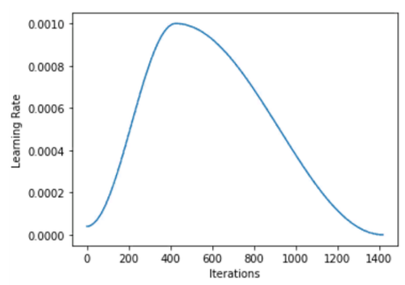
DL

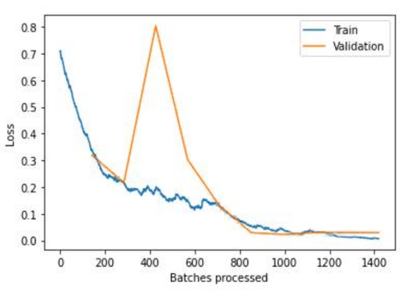
Feature extraction + Classification

Гексаплоид

Output

Framework: FastAI
EfficientNet B0 [2] с весами на ImageNet [3]
Loss: CrossEntropyLoss, *Оптимизатор*: Adam
Batch size: 16, *Epochs*: 10, *Image size*: 224x224
Scheduler: fit_one_cycle (start_lr: 1e-4)
Аугментации: случайные повороты на -20 +20 градусов, изменение яркости, контрастности, насыщенности, зеркальное отображение.
Целевая метрика: “AUC“





Точность на отложенной выборке:	Метод	roc_auc_score	accuracy_score
	EfficientNet	0.99535	0.9878

Список литературы:
1. Genaev et al., Morphometry of the Wheat Spike by Analyzing 2D Images, 2019.
2. Tan M., Le Q. V. Efficientnet: Rethinking model scaling for convolutional neural networks //arXiv preprint arXiv:1905.11946. – 2019.
3. Deng J. et al. Imagenet: A large-scale hierarchical image database //2009 IEEE conference on computer vision and pattern recognition. – Ieee, 2009. – С. 248-255.

Благодарность Николаю Петровичу Гончарову и Афонникову Дмитрию Аркадьевичу за предоставленный биологический материал и помощь в интерпретации результатов.
Работа выполнена при финансовой поддержке Курчатовского геномного центра ИЦиГ СО РАН (Новосибирск, Россия) по договору с Минобрнауки РФ № 075-15-2019-1662