Raport końcowy - klasyfikacja ryżu

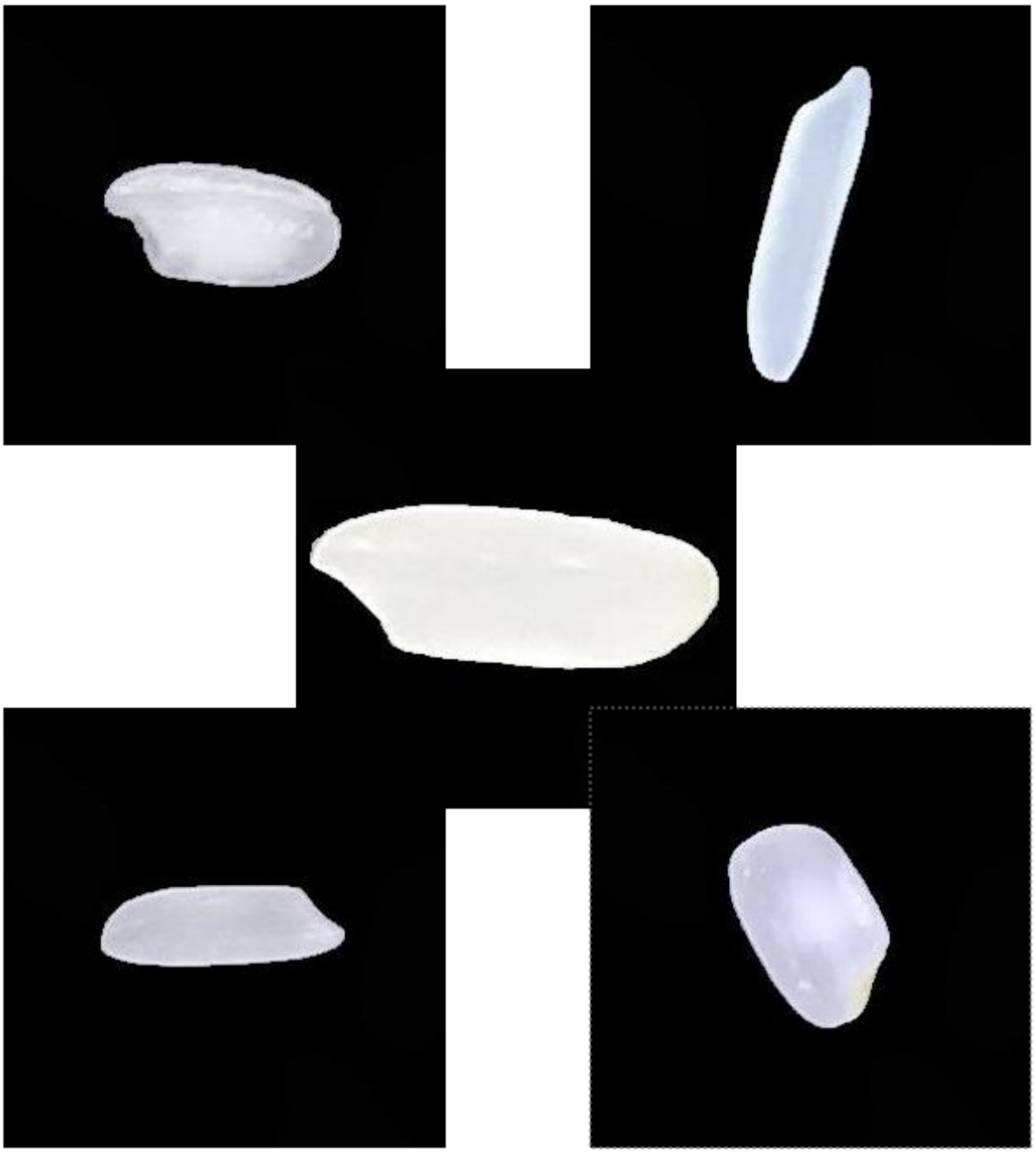
Warsztaty z technik uczenia maszynowego

Maciej Chrabąszcz
Weronika Kamińska
Aleksander Kozłowski
Michał Rosiński

Klasyfikacja na podstawie obrazu

Dane

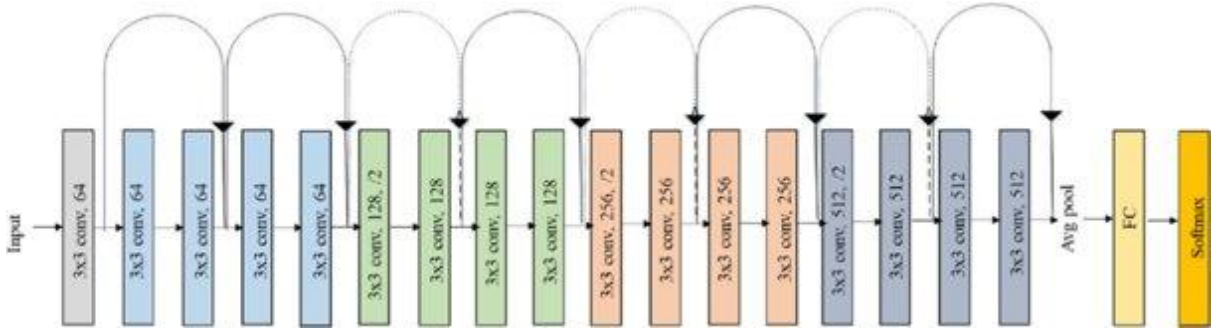
Danymi wejściowymi są zdjęcia ziaren ryżu na czarnym tle. Obrazy są wymiary 224x224 piksele. W naszym końcowym rozwiązaniu zmniejszyliśmy ich wymiar do 96x96 co przyspieszyło trening sieci jednocześnie nie obniżając znacząco wyników. Przykładowe zdjęcia ze zbioru



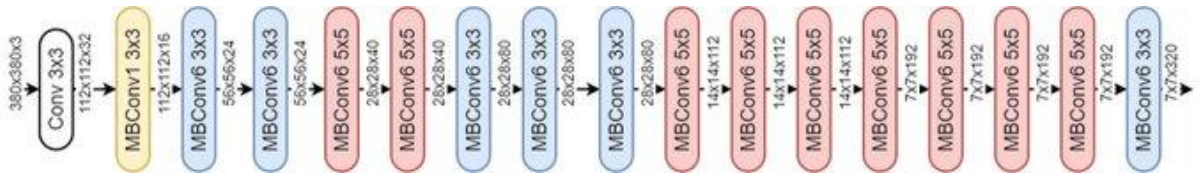
Wykorzystane modele

Jako modele klasyfikujące wykorzystaliśmy sieci konwolucyjne. Skorzystaliśmy z modeli, które dobrze sprawdzają się do klasyfikacji obrazów i są często wykorzystywane w praktyce, *Resnet* i *Efficientnet*. Ze względów na złożoność obliczeniową wykorzystaliśmy najmniejsze architektury (o najmniejszej liczbie wag) dla tych klas modeli. Resnet poza sieciami konwolucyjnymi wykorzystuje *residual connection*. Natomiast Efficientnet dodatkowo korzysta z warstwy *DepthwiseConv2D*, która dla każdego kanału wejściowego stosuje inne filtry konwolucyjne.

Architektura Resnet



Architektura EfficientNet

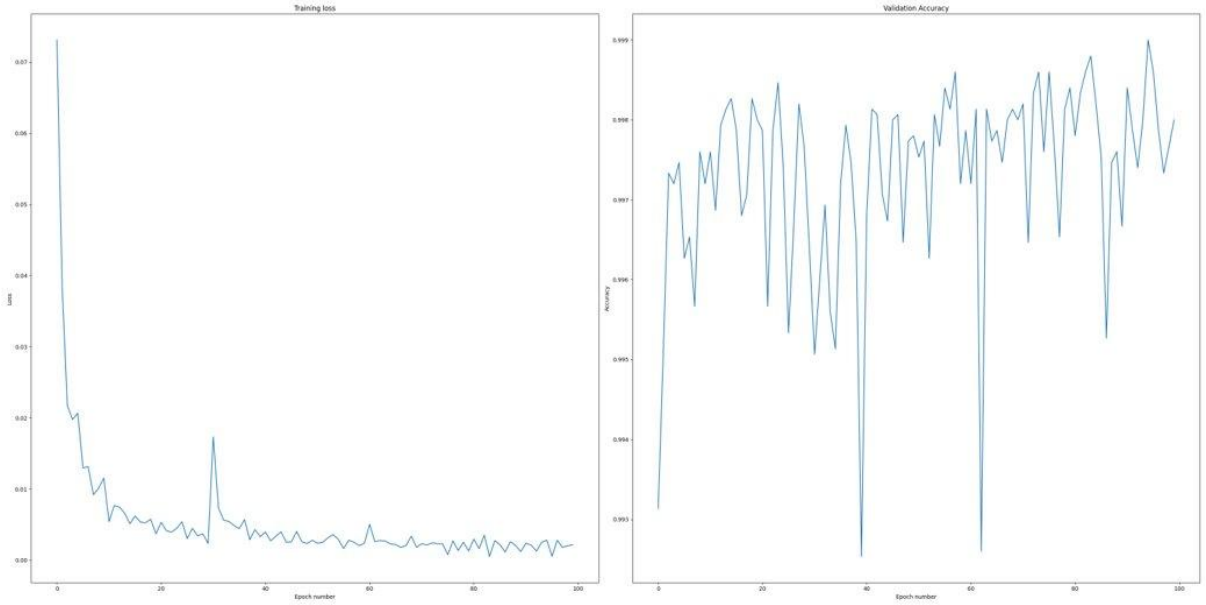


Trening

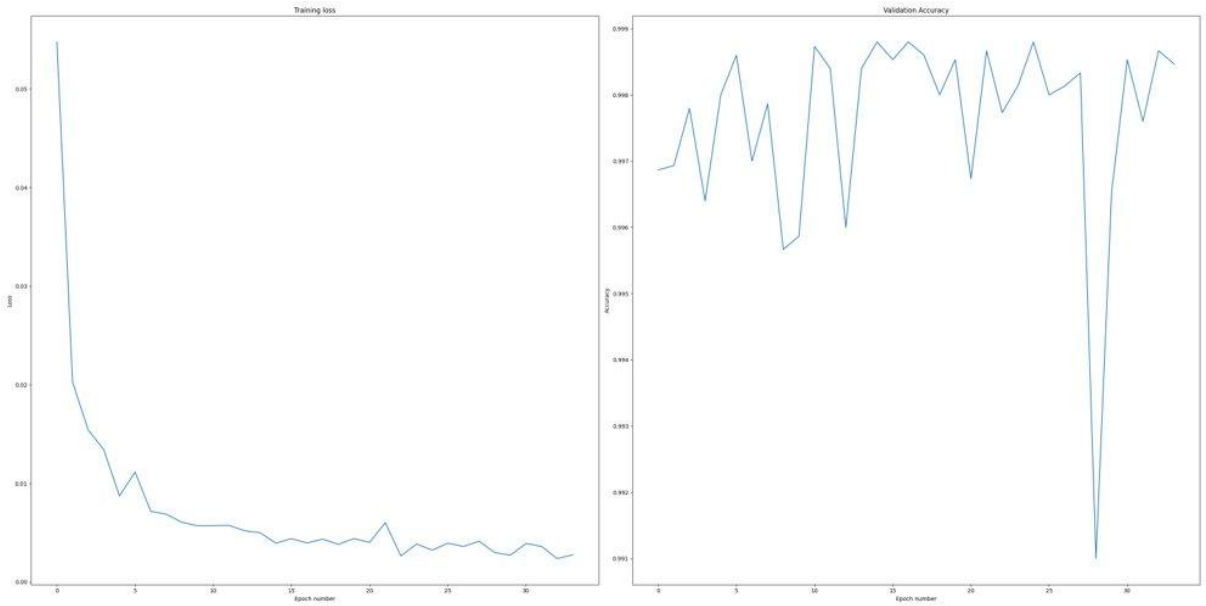
Funkcją straty była *entropia krzyżowa* między przewidzianymi prawdopodobieństwami klas a rzeczywistymi klasami odpowiadającymi obrazom. W celu minimalizacji funkcji straty wykorzystany został algorytm gradientowy *AdamW.* Obrazy były normalizowane w każdym kanale przed przekazaniem do modeli. Oba modele były inicjowane wagami z modeli przetrenowanych na zbiorze *ImageNet*.

Funkcje straty podczas treningu i accuracy na zbiorze walidacyjnym

Resnet



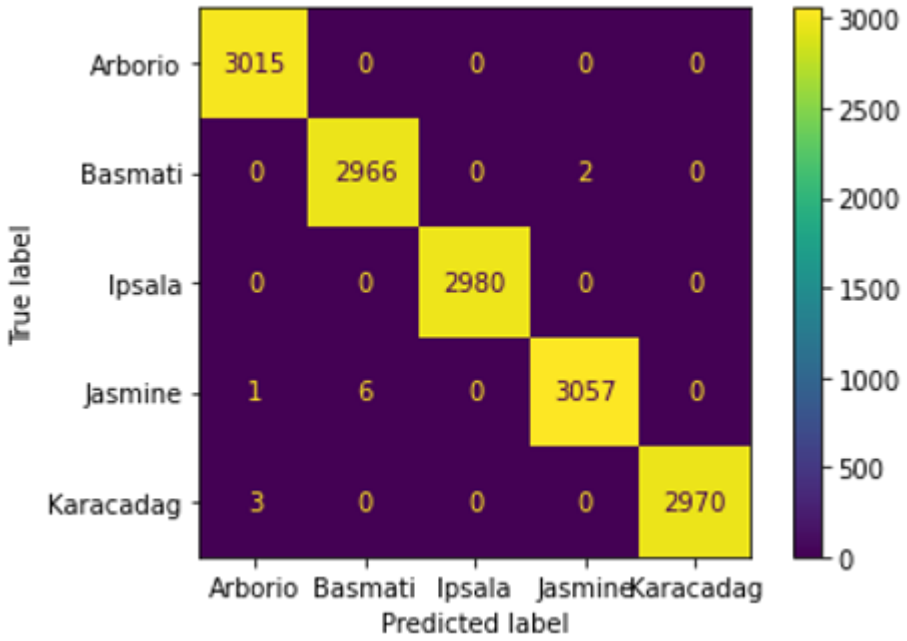
EfficientNet

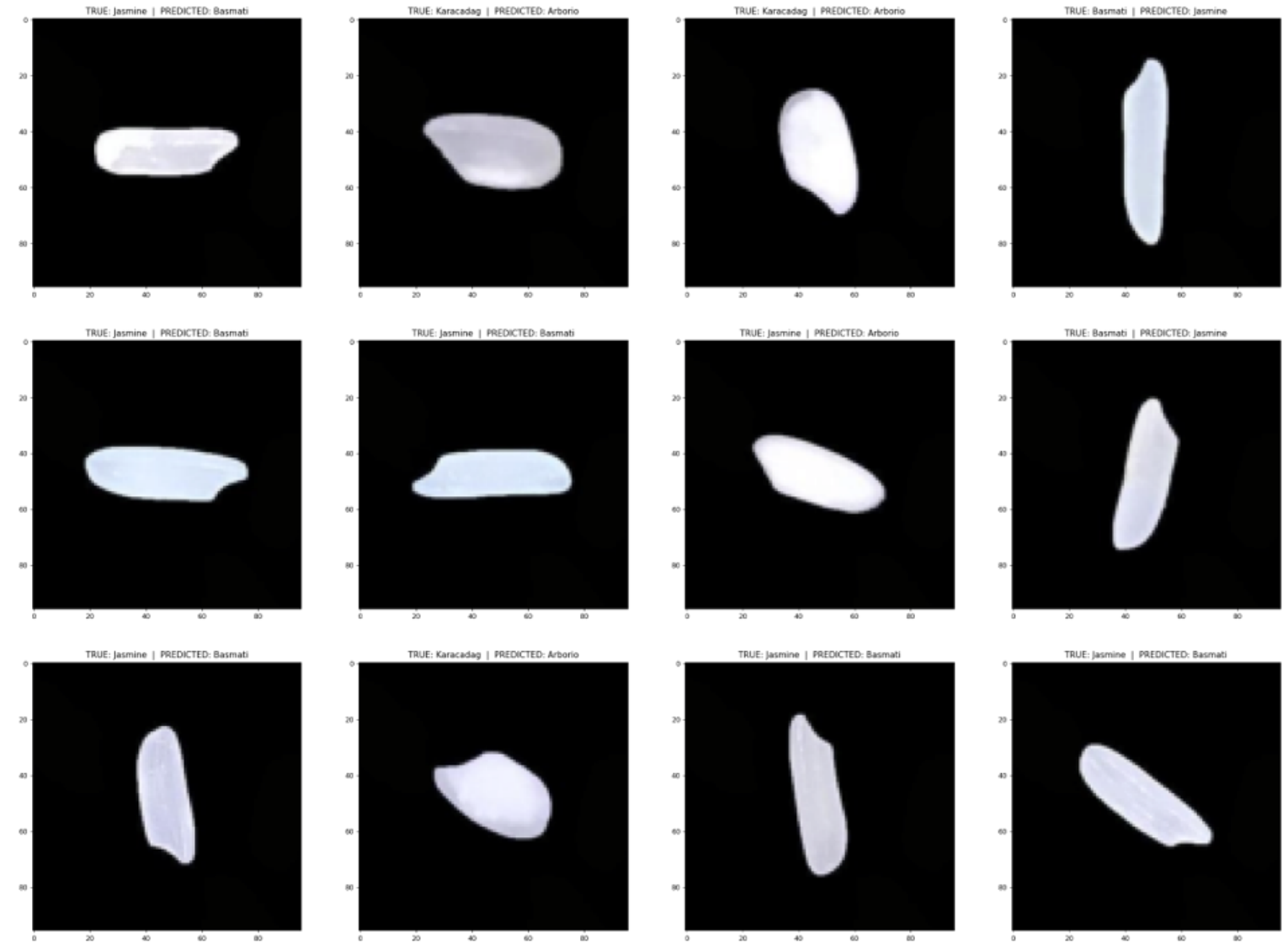


Jak widzimy modele już po pierwszej epoce miały accuracy ponad 0.99. Zatem modele bardzo szybko uczyły się poprawnie rozpoznawać ziarna ryżu. Resnet trenowaliśmy przez 100 epok natomiast Efficientnet przez 20, ponieważ zauważyliśmy stabilizację funkcji straty.

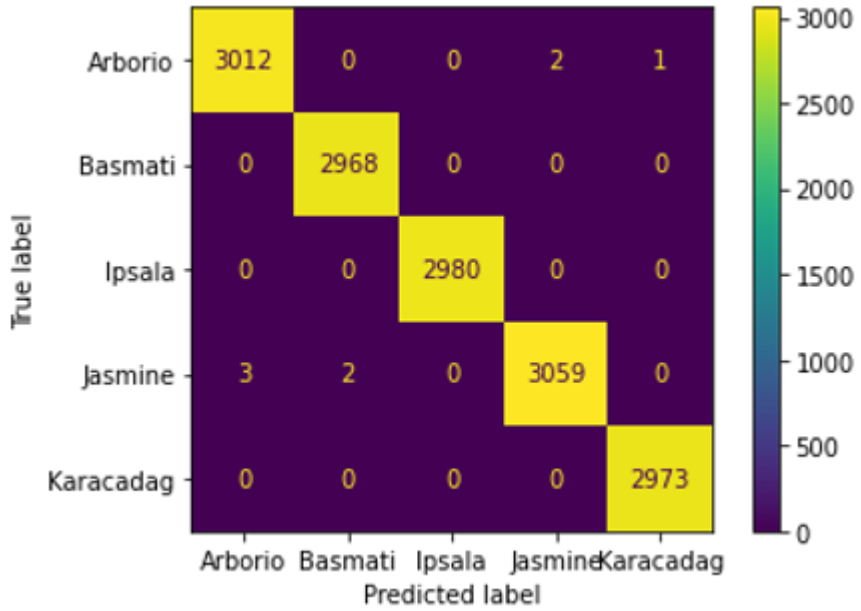
Ewaluacja

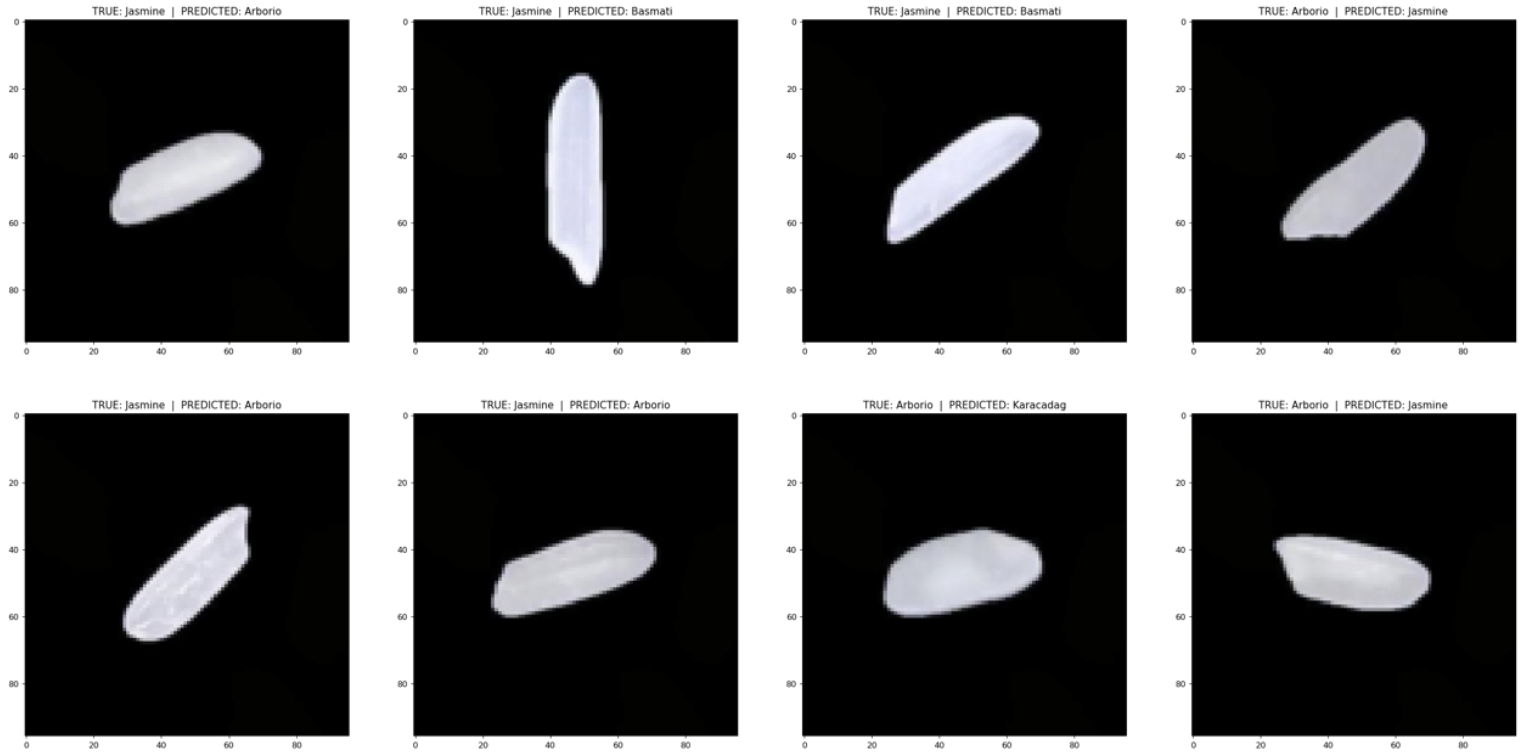
Resnet





EfficientNet





Jak widać modele nie popełniają często błędów. Po przeanalizowaniu obrazów, dla których modele dokonywały błędnej klasyfikacji, stwierdziliśmy, że te ziarna rzeczywiście są podobne do ziaren z przypisanej przez model klasy. Zauważyliśmy, że oba modele mylą ziarna *Basmati* i *Jasmine*, które potrafią być do siebie bardzo podobne.

Własne zdjęcia

Postanowiliśmy przetestować czy nasz model poradzi sobie, gdy podamy mu zdjęcia zrobione przez nas telefonem. Zrobiliśmy zdjęcia ziarna ryżu basmati na dwóch różnych tłach.

Dokonaliśmy jeszcze modyfikacji zdjęć w celu upodobnienia powyższych zdjęć do zbioru danych treningowych.
Resnet uznał 1. zdjęcia za ziarno ryżu jaśminowego, a 2. zdjęcia jako ryż basmati.
Efficientnet uznał oba zdjęcia za ziarna ryżu jaśminowego.
Błąd może wynikać z rozpoznanego przez nas problemu podobieństwa tych ziaren ryżu, a także z różnicy między zdjęciami ze zbioru treningowego a fotografią robioną telefonem.

Klasyfikacja na podstawie danych tabelarycznych

Dane

Danymi wejściowymi są stabelaryzowane zmienne, w tym:

- 106 objaśniających zmiennych ilościowych opisujących różne własności ziarenek ryżu, takie jak długość, szerokość, powierzchnia, twardość, kurtoza itd.
- 1 objaśniana zmienna jakościowa przypisująca obserwacji typ ryżu: Basmati, Arborio, Jasmine, Ipsala, Karacadag.

Poniższa tabela przedstawia fragment naszego zbioru danych.

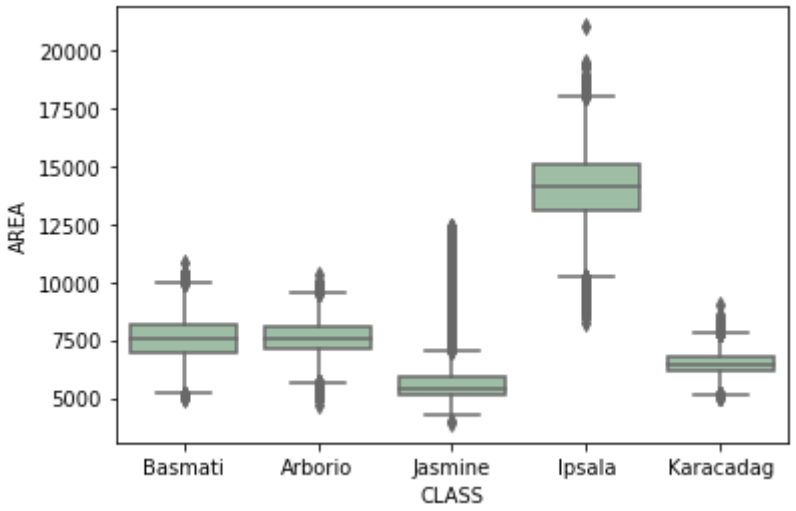
	AREA	PERIMETER	MAJOR_AXIS	MINOR_AXIS	ECCENTRICITY	EQDIASQ	SOLIDITY	CONVEX_AREA	EXTENT	ASPECT_RATIO	...	ALLdaub4L	ALLdaub4a	ALLdaub4b
0	7805	437.915	209.8215	48.0221	0.9735	99.6877	0.9775	7985	0.3547	4.3693	...	113.9924	65.0610	59.5989
1	7503	340.757	138.3361	69.8417	0.8632	97.7400	0.9660	7767	0.6637	1.9807	...	105.7055	64.3685	62.2084
2	5124	314.617	141.9803	46.5784	0.9447	80.7718	0.9721	5271	0.4760	3.0482	...	109.7155	62.6423	58.7439
3	7990	437.085	201.4386	51.2245	0.9671	100.8622	0.9659	8272	0.6274	3.9325	...	116.5405	64.9069	60.2562
4	7433	342.893	140.3350	68.3927	0.8732	97.2830	0.9831	7561	0.6006	2.0519	...	107.7502	64.7071	61.3549

Oryginalny zbiór danych zawierał 75 000 rekordów, w tym 22 obserwacje miały braki w danych - usunęliśmy je na samym początku.

Wstępna eksploracja danych

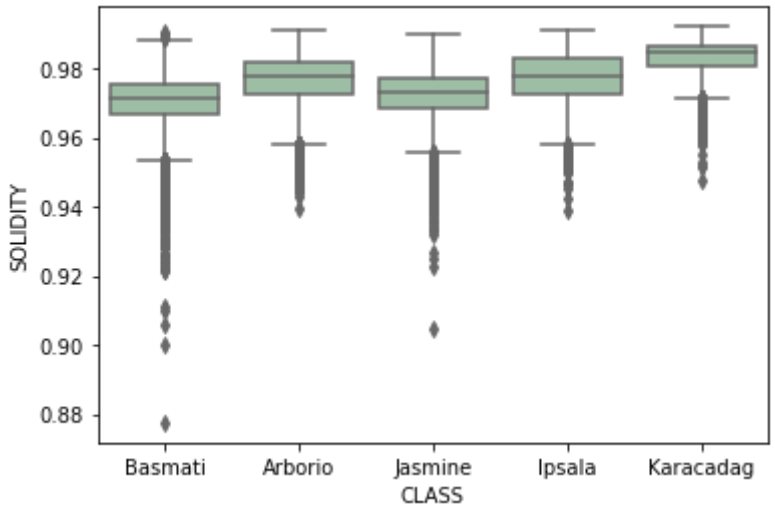
W ramach eksploracji danych sprawdziliśmy jak różnią się od siebie rodzaje ryżu w zależności od 2 wybranych zmiennych. Dla każdej z nich stworzyliśmy boxplot.

- **AREA - powierzchnia**



Widać, że powierzchnia dla ryżów Jasmine, Ipsala i Karacadag znacząco różni się od siebie, natomiast najprawdopodobniej zmienna ta nie pozwoli nam odróżnić Basmati od Arborio.

- **SOLIDITY - trwałość**



W przypadku trwałości osiągamy bardzo podobne wyniki dla wszystkich z badanych rodzajów ryżu. Można się spodziewać, że zmienna ta nie będzie dobrze rozróżniała rodzajów ryżu.

Wykorzystane modele

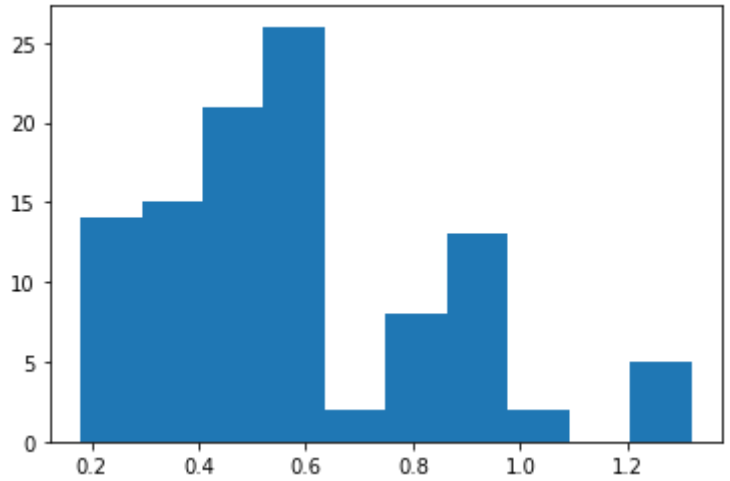
W projekcie skorzystaliśmy z kilku dobrze znanych typów modeli, które wymieniliśmy poniżej:

- KNN
- REGRESJA LOGISTYCZNA
- LASY LOSOWE
- DRZEWA DECYZYJNE
- NAIWNY BAYES

Feature selection

Na początku postanowiliśmy zredukować liczbę zmiennych objaśnianych, jednocześnie oceniając czy nie obniży to za bardzo jakości przyszłych modeli.
Skorzystaliśmy w tym celu z funkcji MIC (mutual info classifier), przypisującej każdej ze zmiennych pewną wartość - im większa wartość, tym lepsza powinna okazać się ta zmienna w klasyfikacji odmian ryżu.

Poniższy histogram przedstawia rozkład wartości funkcji MIC dla zmiennych objaśniających:



1. Najpierw postanowiliśmy wybrać tylko te zmienne, które osiągnęły wartość większą niż 0.5: takich zmiennych było **63**.
2. Następnie postawiliśmy ograniczenie dolne 0.8: takich zmiennych było **26**.
3. Na końcu przyjęliśmy ograniczenie 0.95: takich zmiennych było **8**.

Ostatecznie przyjęliśmy 8 zmiennych wskazanych przez najbardziej restrykcyjne ograniczenie:
AREA, MAJOR_AXIS, MINOR_AXIS, EXTENT, ASPECT_RATIO, ROUNDNESS, COMPACTNESS, SHAPEFACTOR_2.

W dalszej części projektu tworzyliśmy różne modele (wspomniane przez nas wcześniej) osobno dla oryginalnej i zredukowanej liczby zmiennych. Sprawdzaliśmy jak zmieniało się accuracy i czy zredukowana do 8 liczba zmiennych wciąż dawała zadowalające wyniki.

Drzewa decyzyjne

Otrzymaliśmy następujące wartości accuracy dla różnych liczb zmiennych:

- dla 106 zmiennych: **0.996267**,
- dla 63 zmiennych: **0.9968**,
- dla 26 zmiennych: **0.99293**,
- dla 8 zmiennych: **0.9916**.

Zatem redukcja liczby zmiennych nie pogorszyła nam modelu w istotny sposób.

Lasy losowe

Otrzymaliśmy następujące wartości accuracy dla różnych liczb zmiennych:

- dla 106 zmiennych: **0.999**,
- dla 8 zmiennych: **0.995**.

Zatem redukcja liczby zmiennych nie pogorszyła nam modelu w istotny sposób.

Naiwny Bayes

Poniżej zamieściliśmy raport klasyfikacji dla 106 i 8 zmiennych:

- dla 8 zmiennych:

	precision	recall	f1-score	support
Arborio	0.47	0.30	0.37	1478
Basmati	0.52	0.61	0.56	1509
Ipsala	0.98	0.99	0.98	1543
Jasmine	0.90	0.68	0.77	1480
Karacadag	0.60	0.85	0.70	1490
accuracy			0.69	7500
macro avg	0.69	0.69	0.68	7500
weighted avg	0.70	0.69	0.68	7500

- dla 106 zmiennych:

	precision	recall	f1-score	support
Arborio	0.78	0.53	0.63	1503
Basmati	0.65	0.73	0.69	1548
Ipsala	0.98	0.99	0.99	1492
Jasmine	0.91	0.68	0.78	1483
Karacadag	0.60	0.87	0.71	1474
accuracy			0.76	7500
macro avg	0.78	0.76	0.76	7500
weighted avg	0.78	0.76	0.76	7500

Jak widać, naiwny Bayes przyniósł dużo gorsze rezultaty. Jest tutaj też widoczna różnica w jakości modelu pomiędzy oryginalnym zbiorem danych, a zredukowanym do 8 zmiennych.

KNN

Poniżej zamieściliśmy raport klasyfikacji dla 106 i 8 zmiennych:

- dla 8 zmiennych:

	precision	recall	f1-score	support
Arborio	0.88	0.89	0.88	1491
Basmati	0.87	0.81	0.84	1527
Ipsala	1.00	1.00	1.00	1434
Jasmine	0.91	0.96	0.94	1557
Karacadag	0.86	0.86	0.86	1491
accuracy			0.90	7500
macro avg	0.90	0.90	0.90	7500
weighted avg	0.90	0.90	0.90	7500

- dla 106 zmiennych:

	precision	recall	f1-score	support
Arborio	0.92	0.94	0.93	1503
Basmati	0.89	0.85	0.87	1548
Ipsala	1.00	1.00	1.00	1492
Jasmine	0.93	0.97	0.95	1483
Karacadag	0.88	0.86	0.87	1474
accuracy			0.92	7500
macro avg	0.92	0.92	0.92	7500
weighted avg	0.92	0.92	0.92	7500

Jak widzimy, KNN sprawuje się znacznie lepiej niż naiwny Bayes, ale wciąż gorzej niż lasy losowe. W przypadku KNN różnice pomiędzy 106 a 8 zmiennymi są mniejsze.

Regresja logistyczna

Poniżej zamieściliśmy raport klasyfikacji dla 106 i 8 zmiennych:

- dla 8 zmiennych:

	precision	recall	f1-score	support
Arborio	0.77	0.86	0.81	1491
Basmati	0.75	0.80	0.78	1527
Ipsala	1.00	1.00	1.00	1434
Jasmine	0.75	0.65	0.69	1557
Karacadag	0.87	0.83	0.85	1491
accuracy			0.82	7500
macro avg	0.83	0.83	0.83	7500
weighted avg	0.83	0.82	0.82	7500

- dla 106 zmiennych:

	precision	recall	f1-score	support
Arborio	0.66	0.86	0.75	1503
Basmati	0.73	0.77	0.75	1548
Ipsala	0.97	0.99	0.98	1492
Jasmine	0.64	0.44	0.52	1483
Karacadag	0.85	0.81	0.83	1474
accuracy			0.77	7500
macro avg	0.77	0.77	0.77	7500
weighted avg	0.77	0.77	0.77	7500

Jak widzimy, regresja logistyczna osiąga nieco lepsze rezultaty niż naiwny Bayes, ale gorsze niż KNN.

PCA

Na końcu postanowiliśmy przy pomocy PCA zweryfikować liczbę potrzebnych składowych głównych do osiągnięcia dobrych rezultatów. Z przeprowadzonej analizy płyną poniższe wnioski.
Chcieliśmy wybrać tyle kierunków, by uzyskać przynajmniej 60% wariancji.
Dla 8 zmiennych osiągnęliśmy **0.889874867739401** wariancji.
Zatem w danych mamy bardzo dużo zmiennych, które niezbyt dobrze wyjaśniają typ ryżu, a jest kilka zmiennych, które robią to bardzo dobrze.
Czyli 8 składowych głównych wystarcza do klasyfikacji na wysokim poziomie.
Ponadto 14 składowych wyjaśnia łącznie ponad **95%** wariancji, co daje klasyfikację na bardzo wysokim poziomie.

WNIOSKI

Okazuje się zredukowanie zbioru danych ze 106 zmiennych objaśniających do 8 było dobrą decyzją. Nie spowodowało to znaczącego pogorszenia jakości większości z rozpatrywanych modeli. Może świadczyć to o tym, że tylko te 8 zmiennych bardzo istotnie odróżnia typy ryżu od siebie, a wiele cech ryżu jest podobnych dla wszystkich ich rodzajów.
Najlepsze rezultaty odnieśliśmy przy pomocy lasów losowych, które pozwoliły doskonale przewidywać typ ryżu na podstawie 8 zmiennych.