How to Compete in Machine Learning Challenges

on Kaggle and Other platforms

Karim Amer
Co-Founder and Head of AI/ML at VAIS

Biography

- Experience
  - Co-founder and Head of AI/ML at VAIS (Aug, 2020 - current)
  - Research Intern at Siemens Healthineers, USA (Feb, 2019 - Feb, 2020)
  - Research Assistant at Nile University (Oct, 2016 - Jan, 2019)
- Accomplishments
  - 3 US patents (pending)
  - 9 publications (CVPR, ICCV, WACV, ISBI, AGU, Nature (pending))
  - +100 citations

Biography

VAIS
www.vais.ai

- International Machine Learning competitions:
  - **1st** place at Crop Detection from Satellite Imagery - Radiant Earth Foundation (**Zindi**)
  - **2nd** place at Predict Wind Speeds of Tropical Storms - NASA (**DrivenData**)
  - **2nd** place at Embedded Real Time Inference - KU Leuven (**CVPR**)
  - **3rd** place at AWS Informal Settlements in South Africa - SANSA (**Zindi**)
  - **4th** place at Computer Vision for Crop Disease - CGIAR (**Zindi**)
  - **4th** place at MagNet: Model the Geomagnetic Field - NOAA (**DrivenData**)
  - **8th** place at DeepGlobe: Land Cover Classification - Facebook (**CVPR**)
  - **10th** place at Ugandan Air Quality Forecast - AirQo (**Zindi**)
  - **Gold medal** at OpenVaccine - Stanford (**Kaggle**)
  - **Bronze medal** at SETI Breakthrough Listen - E.T. Signal Search - Berkeley (**Kaggle**)
  - **Bronze medal** at Duplicate Ads Detection - Avito (**Kaggle**)

Biography

KarimAmer

Head of AI/ML at Visual and AI Solutions (VAIS)
Cairo, Cairo Governorate, Egypt
Joined 7 years ago · last seen in the past day

Followers 61
Following 1

Competitions Master

Home | Competitions (23) | Datasets | Code (9) | Discussion (24) | Followers (61) | Notifications | Account | Edit Public Profile

Competitions Summary

Competitions Master

Current Rank	Highest Rank
1783 of 171,346	1217

Gold	Silver	Bronze
1	2	2

Competitions: 23
Solo: 11 (48%)
Team: 12 (52%)

Biography

- Competition co-host:
  - UmojaHack Egypt 2021 (**Zindi**).
  - Lacuna - Correct Field Detection Challenge (**Zindi**).

Objective

- What is covered?
  - Introduce DS platforms.
  - Describe practical ML development cycle (concepts and ideas).
  - Focus on Supervised Learning.
- What isn't covered?
  - Details of ML/DL models.
  - Learning path.

Data Science Platforms

- A place where:
  - Entities host their problems as challenges with money prizes for top solutions.
  - Data Scientists are asked to solve these challenges.
- Entities can be:
  - Companies: Google, Facebook, Microsoft, … etc.
  - Universities: Stanford, Berkeley, … etc.
  - Governments: USA, England, … etc.
  - NGOs: Radiant Earth Foundation, CGIAR, … etc.
- Mostly anyone can participate and win:
  - Restrictions might apply to certain countries like North Korea.

How is a competition run?

1. An entity comes to the platform with a business requirements and a dataset.
2. The platform team analyze the requirements and clean the dataset.
3. The platform team design the competition metric and split the dataset into training, public test and private test (the results are known only after the competition ends).
4. Participants compete for a certain time frame.
5. After the competition ends, top solutions are reviewed by the platform team and receive prize money.
6. The entity receives the codes and documentation of top solutions and use them to fulfil the business requirements.

Data Science Platforms

- Kaggle.
- Driven Data.
- Zindi.
- AI Crowd.
- Conference workshops (**hidden gems!**).
  - Check the workshop challenges of the big conferences.

What can we learn from competitions?

VAIS
www.vais.ai

- Gain practical experience.
- Find new state of the art methods.
- Learn from the experts.
- Add more weight to your CV.

Practical ML Cycle

Practical ML Cycle

Data Exploration → Validation strategy → Data Preprocessing → Feature Engineering → Modeling → Results Inspection

Data Exploration

- Goal:
  - Understand dataset dynamics.
- Methodology:
  - Have a deep look on all the data.
  - Research the domain knowledge related for the dataset.
  - Visually inspect the data.
  - Calculate data statistics on numerical features.
  - Explore categorical features.
  - Check if there are missing values.

Validation Strategy

- Goal:
  - Create a local validation set that emulates test data (**very important!!**).
- Methodology:
  - Depends on the available dataset.
  - Use k-fold validation [1] when possible.
  - Examples:
    - Balanced classes or regression (normal cases): random splitting.
    - Imbalanced classes: stratified splitting [2].
    - Dataset has several subsets (patients for example): group splitting [3].
- Note:
  - It is a make or break step!

Data Preprocessing

- Goal:
  - Create a clean version of the dataset to improve the results.
- Methodology:
  - Normalize numerical features.
  - Correct distribution to gaussian distribution.
  - Impute missing values.
  - Transform categorical features to numerical features.
- Note:
  - Most likely not needed for tree-based models.

Feature Engineering

- Goal:
  - Create descriptive features from the raw dataset to make modeling step easier.
- Methodology:
  - Use the info discovered in the data exploration step.
  - Try to come up with strong composition features.
  - Find a good space transformation for your features (for example cartesian -> polar).
- Note:
  - Most likely not needed for deep learning models (**not always!**).

Modeling

- Goal:
  - Train a good model using the generated features (or the raw data).
- Methodology:
  - Always start with a simple and fast to develop model.
  - Increase model complexity:
    - Try adding more layers.
    - Try different layers.
    - Try increasing layers size.
  - Decrease overfitting:
    - Do more augmentation.
    - Try adding pooling layers.
    - Try smoothing predictions by: bagging ensemble, Snapshot ensemble [4], SWA [5], … etc.

Modeling

- Notes:
  - Change one parameter at a time.
  - If Neural Networks isn’t suitable, trust Gradient Boosting Machines.
  - Augmentation and ensemble are the secret recipes.
  - Don’t sleep on semi-supervised learning or self-supervised learning.

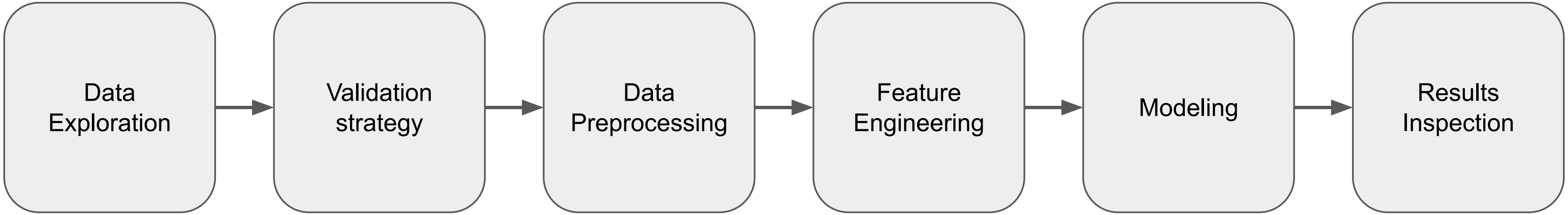
Modeling - The Art of Ensemble

- It is all about diversity!
  - Train different models.
  - Train with different subsets of features.
- Ensemble of 1000 models is (most likely) better than ensemble of 100 models.
- Soft voting is better than hard voting.
- Whenever possible use Ensemble Selection [6].
- Stacking [7] is great if you can make it work but it is difficult.

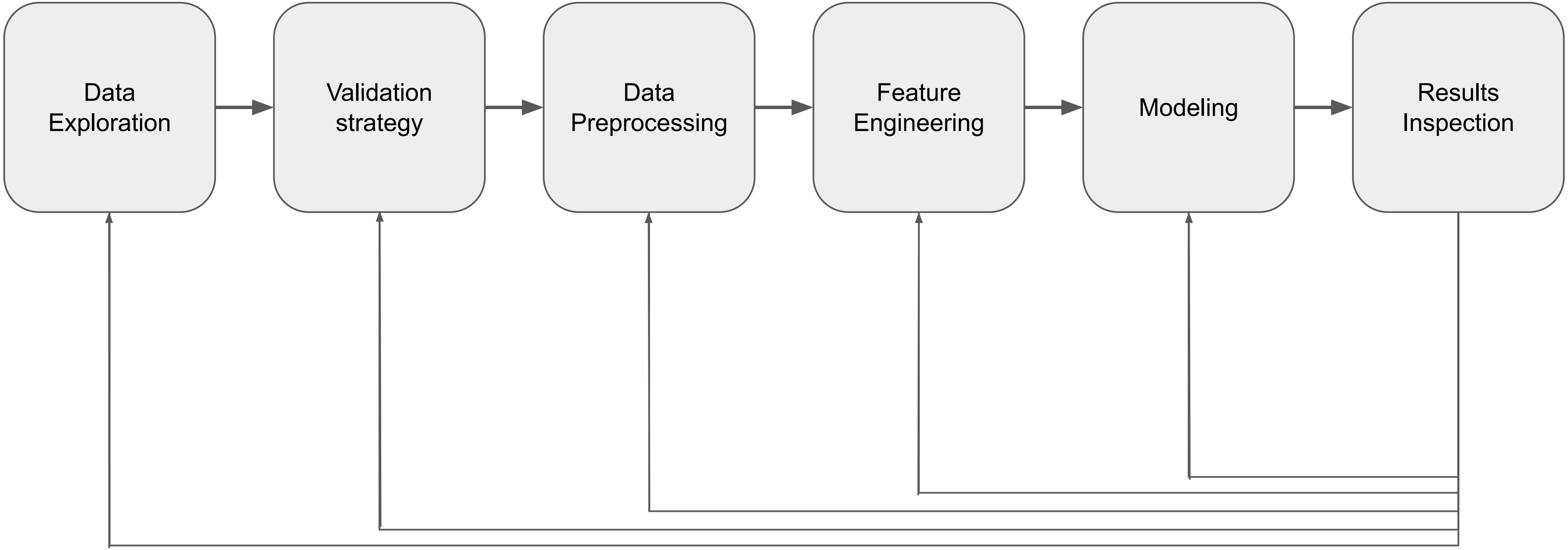
Results Inspection

- Goal:
    - Analyzing the results to find where to improve.
- Methodology:
    - Visually inspect the data with the results (not possible if not images or text).
    - Visualize the results statistics.
    - Heat maps of Neural Networks might be useful [8].

Practical ML Cycle - Expectation



Practical ML Cycle - Reality



Tools

- Pandas
- Numpy
- Matplotlib
- Scikit-learn
- Xgboost (or LightGBM or CatBoost)
- Pytorch (or Keras or TF)

Tips

- Read the competition rules (team size, allowed submissions, external dataset, … etc).
- Organize and document experiments.
- Don’t start from scratch if possible.
- Don’t give up.
- Be creative.
- Learn from the solutions of top competitors.
- Choose competitions suitable to available resources.
- Invest in better resources.

What can't we learn from competitions?

- Problem definition.
- Data collection.
- Data cleaning (to some extent).
- Maintaining annotation quality.

Other Ways to Build Your ML Profile

- Publish scientific papers.
- Replicate other people works.
- Get involved in open source projects.
- Internships.
- Master’s (and Phd) degree.

Crop Detection from Satellite Imagery Challenge

Challenge Overview

- One of the important monitoring tasks for EO systems.
- Classifying planted crop types across any country can help governments in:
    - Monitoring the national agricultural plans
    - Early yield estimation
    - Harvest planning

Challenge Overview

VAIS
www.vais.ai

ICLR Workshop Challenge #2: Radiant Earth Computer Vision for Crop Detection from Satellite Imagery

$5,000 USD

Identify crop type using satellite imagery, and win a trip to present your work at ICLR 2020 in Addis Ababa.

492 data scientists enrolled, 110 on the leaderboard

Agriculture | Computer Vision | Satellite | Unstructured | SDG2

Kenya

3 February 2020—29 March 2020

55 days

Info | Data | Discussions | Leaderboard | Team | Submissions

Get a score

- Description
- Rules
- Evaluation
- Prizes
- Timeline

Agriculture is the driving engine for economic growth in many countries, particularly in the global development community. Therefore, accurate and reliable agricultural data around the world is critical to global resilience and food security. These data are also essential to monitor the progress toward several UN Sustainable Development Goals, including ending poverty, zero hunger, economic growth and more.

Earth observations (EO) provide invaluable data at different spatial and temporal scales and at consistent frequencies. These data can be used to build models for agricultural monitoring, increasing farmers productivity and enhancing the impact of intervention mechanisms.

In contrast with a survey, agricultural maps based on satellite data provide a more accurate

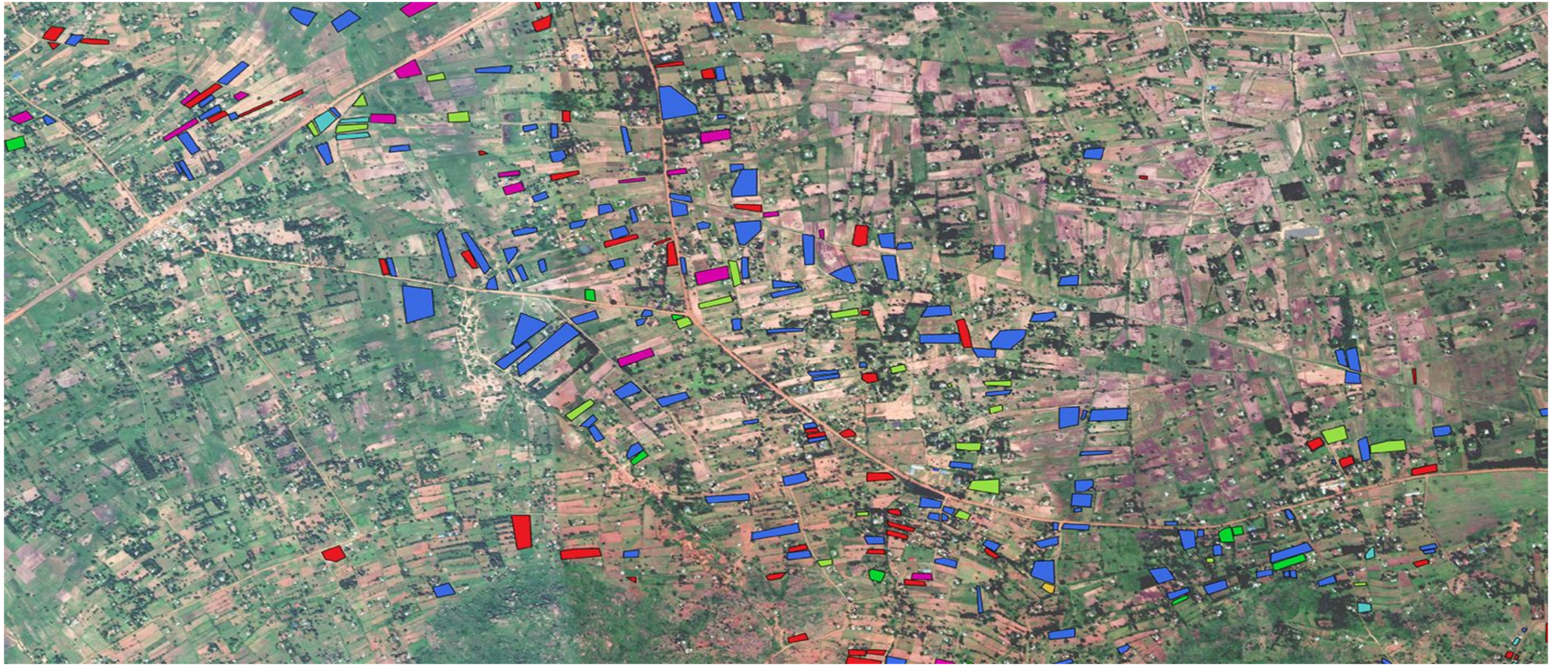
Challenge Overview

- Given a small crop field (farm), classify the planted crop into one of the following:
    - Maize
    - Cassava
    - Common Bean
    - Maize & Common Bean (intercropping)
    - Maize & Cassava (intercropping)
    - Maize & Soybean (intercropping)
    - Cassava & Common Bean (intercropping)
- Metric: Cross Entropy

Dataset Exploration

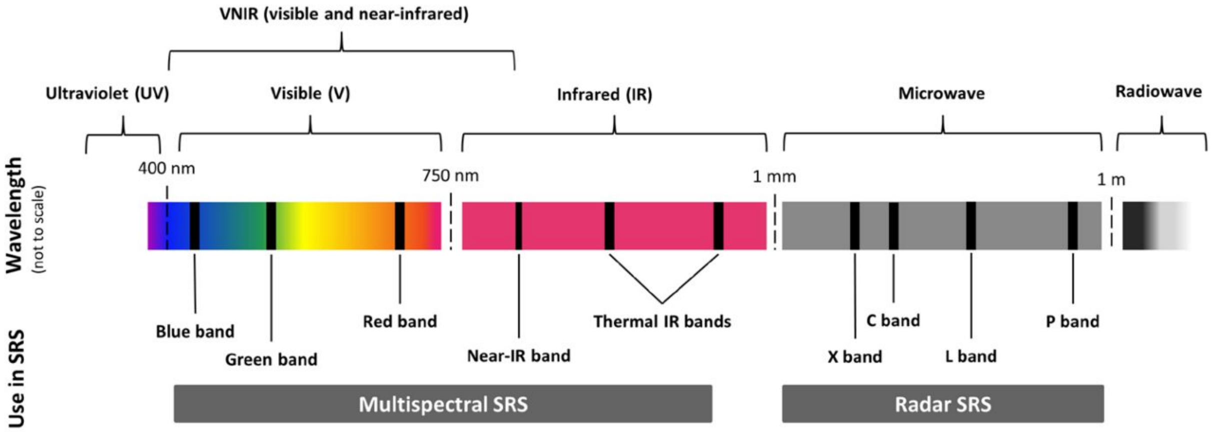
- Time series of high resolution satellite images of an agricultural area in west Kenya acquired in 13 different days within 5 months.
- Each image has
  - Size of 4032 X 6070 pixels.
  - 13 spectral bands.
- Number of annotated crop fields in the area is 4688.
  - 3286 for training.
  - 1402 for testing.

Dataset Exploration



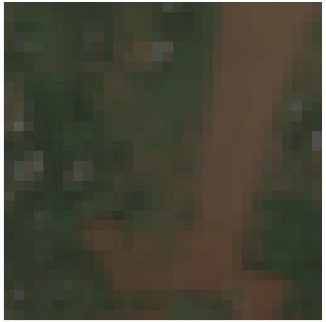
Dataset Exploration

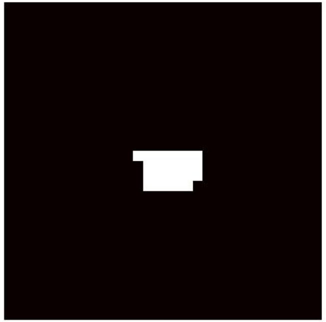
- Provided spectral bands:
  - RGB
  - Visual and Near Infrared
  - Ultra-Blue
  - Short Wave Infrared
  - Cloud probability layer



Dataset Exploration

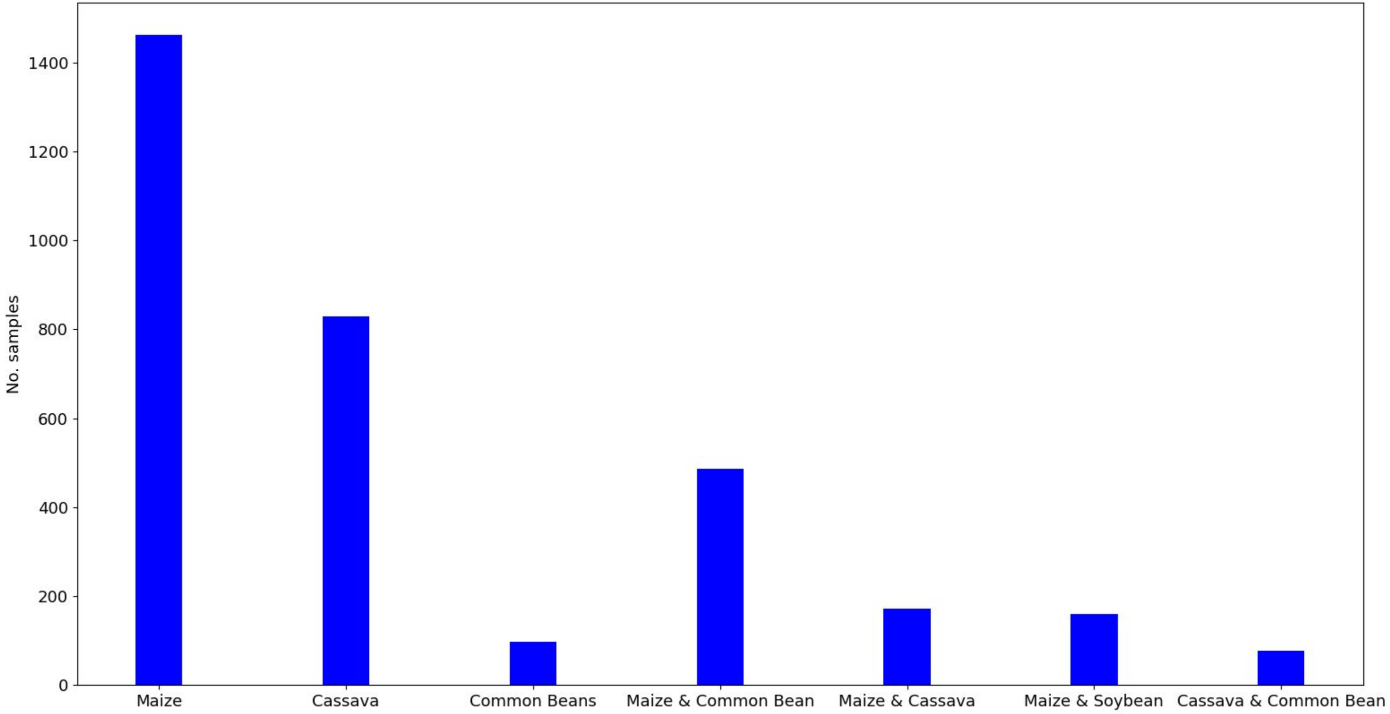
- Sample field visualization (RGB only)



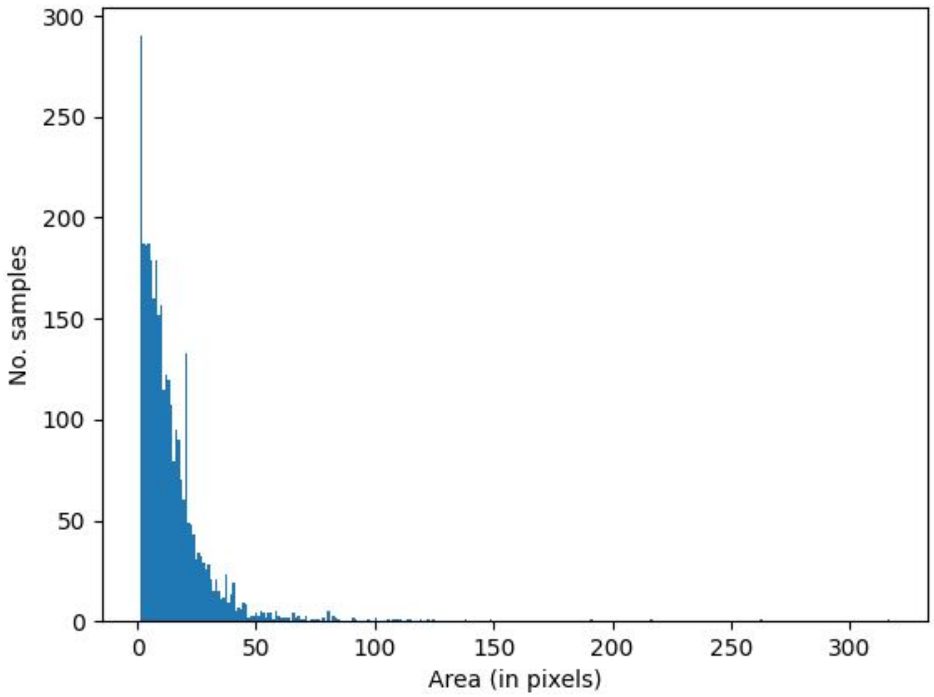
Data Exploration

- Class frequency



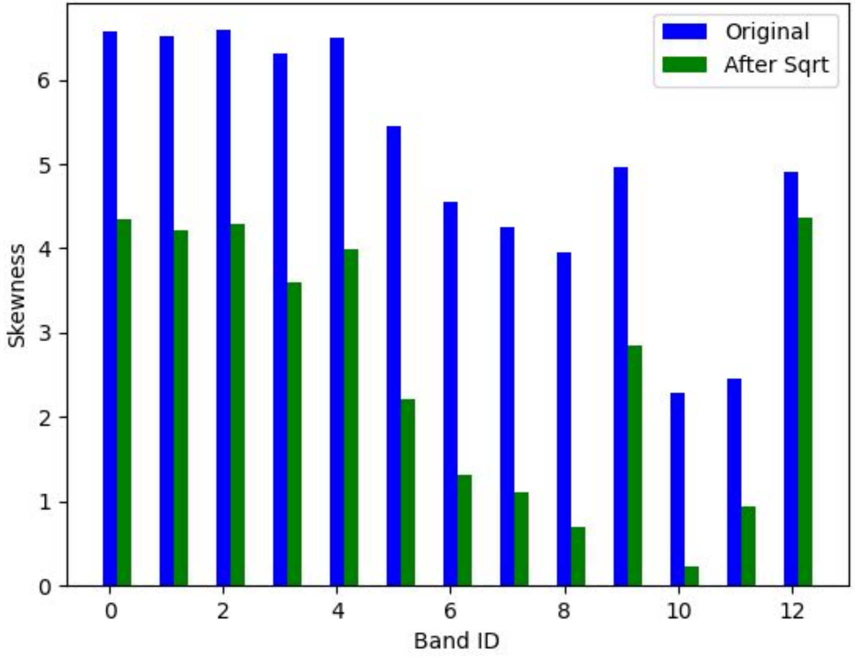
Data Exploration

- Field area distribution



Data Exploration

- Band skewness



Data Exploration

- Challenges:
  - Small dataset.
  - High dimensionality (spatio-temporal data).
  - Unbalanced classes.
  - A lot of crop fields is only couple of pixels.
  - No details to be seen with the naked eye.

Local Validation Strategy

- Initial experiment: 1 split with 75% training, 25% validation.
- Submission experiment: 10 splits with 85% training, 15% validation.
- Splits are stratified.
  - Stratification produces similar distribution between training and validation.
- Why stratification rather than random splitting?
  - Competition metric is cross entropy which is highly sensitive to class distribution.

Related Work

- 3D Convolutional Neural Networks
  - Ji, S., Zhang, Z., Zhang, C., Wei, S., Lu, M., & Duan, Y. (2020). Learning discriminative spatiotemporal features for precise crop classification from multi-temporal satellite images. International Journal of Remote Sensing, 41(8), 3162-3174.
- Random Forest
  - Viskovic, L., Kosovic, I. N., & Mastelic, T. (2019, September). Crop classification using multi-spectral and multitemporal satellite imagery with machine learning. In 2019 International Conference on Software, Telecommunications and Computer Networks (SoftCOM) (pp. 1-5). IEEE.
  - Ok, A. O., Akar, O., & Gungor, O. (2012). Evaluation of random forest method for agricultural crop classification. European Journal of Remote Sensing, 45(1), 421-432.

Data Preprocessing

- Normalization
    - Square root (to decrease skewness).
    - Standard scaling (transform to zero mean and unit std).

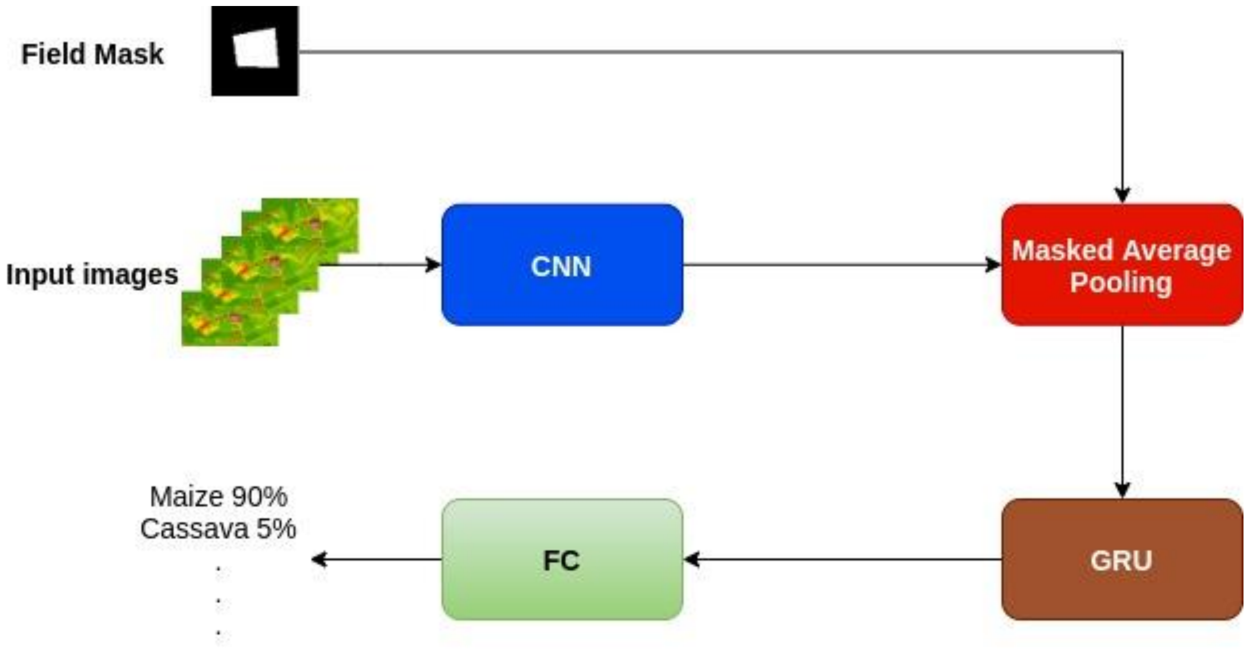
Data Preprocessing

- Transform original Images into small patches:
  1. Calculate the center of each crop field.
  2. Input patch: crop a 32X32 patch around the center so each patch has size (T, C, H, W) where:
     - T: number of time steps = 13
     - C: number of spectral bands = 13
     - H: height = 32
     - W: width = 32
  3. Input field mask: crop a 32X32 binary mask around the same center where field pixels are ones and others are zeros. The size of each field mask is (1, 1, H, W).

Feature Engineering

- Remove one short-wave infrared band (B11, 1610 nm).
- Add 3 vegetation indices.
    - NDVI
    - NDWI
    - B-NDVI
- Total number of spectral bands become ***15***.

Model Architecture



Masked Average Pooling

$$output = \frac{\sum_{H}\sum_{W} input * mask}{\sum_{H}\sum_{W} mask}$$

Data Augmentation

- Spatial augmentations: rotation, flipping and random cropping.
- Mixup [9]: weighted summation of input patch and a random patch cropped from any satellite image.
- Time augmentation: randomly drop one time step.

Ensemble

- Bagging ensemble of **10** models of the same architecture
  - Each trained on a different subset (85%) of the training data.
- Each model is trained using Snapshot ensemble [4]
  - Train the model with cyclical scheduler for **6** cycles.
  - Create ensemble of model snapshots taken at the end of each cycle.
- Total number of models in the ensemble is **60**.

[4] Huang, Gao, et al. "Snapshot ensembles: Train 1, get m for free." arXiv preprint arXiv:1704.00109 (2017).

Results

SCORE	RANK		SUBMISSIONS	SUBMITTED
		This is the final leaderboard. The competitions is officially closed and will not accept any more submissions. Congratulations to all that participated.		
1.102264609	1	KarimAmer oh, hi!	31	~1 month ago
1.168877091	2	youngtard	154	~1 month ago
1.174099923	3	team Be_CarEFuL	91	~2 months ago
1.176934328	4	team Threshold	116	~2 months ago
1.177508763	5	overfitting_PLB Axa Mansard(Nigeria)	114	~1 month ago

References

1. https://machinelearningmastery.com/k-fold-cross-validation/
2. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html
3. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GroupKFold.html
4. https://arxiv.org/abs/1704.00109
5. https://arxiv.org/abs/1803.05407
6. https://www.cs.cornell.edu/~alexn/papers/shotgun.icml04.revised.rev2.pdf
7. https://neptune.ai/blog/ensemble-learning-guide
8. https://jacobgil.github.io/deeplearning/class-activation-maps
9. https://arxiv.org/pdf/1710.09412.pdf