SHRUG-FM: Reliability-Aware Foundation Models for Earth Observation

Kai-Hendrik Cohrs[†1], Maria Gonzalez-Calabuig[†1], Vishal Nedungadi[†2], Zuzanna Osika[†3], Ruben Cartuyvels[4], Steffen Knoblauch[5], Joppe Massant[*6], Shruti Nath[*7], Patrick Ebel[*4], Vasileios Sitokonstantinou[*2]
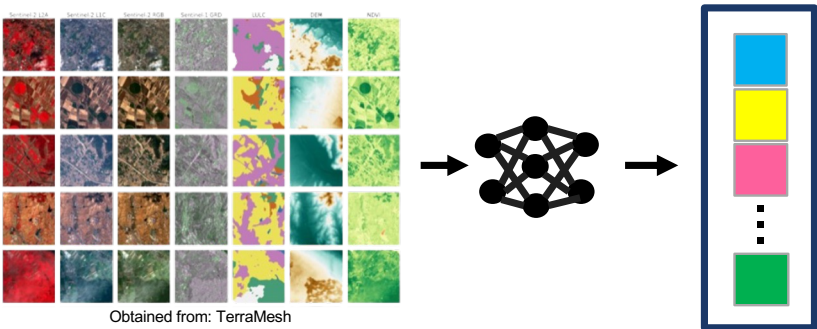
[1]University of Valencia, [2]Wageningen University and Research, [3]Delft University of Technology, [4]European Space Agency, [5]Heidelberg University, [6]Ghent University, [7]University of Oxford, [†]Equally-Contributing First Author
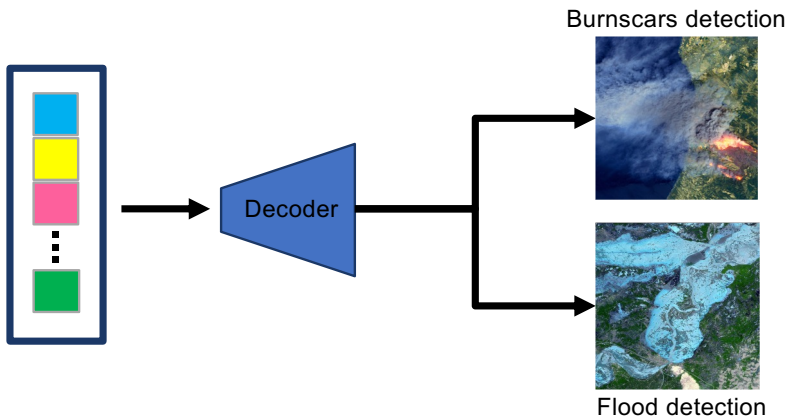
* Supervisors

CHALLENGE AND OPPORTUNITY

Foundational Models (FMs) are trained on a wealth of data to encode information into embedding spaces
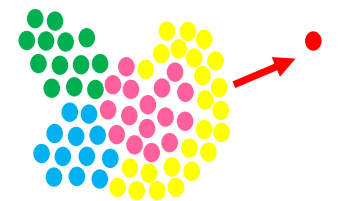

Obtained from: TerraMesh

Embedding spaces efficiently compress data into its salient structures that can be further fine-tuned for downstream tasks
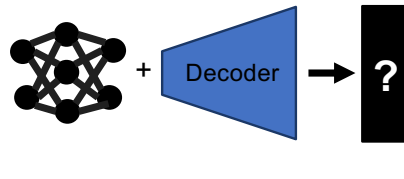

Burnscars detection

Flood detection

But what happens for extreme cases when:

New data is Out-of-Distribution: The downstream uncertainty is high:

Can we make FMs say #IDK?

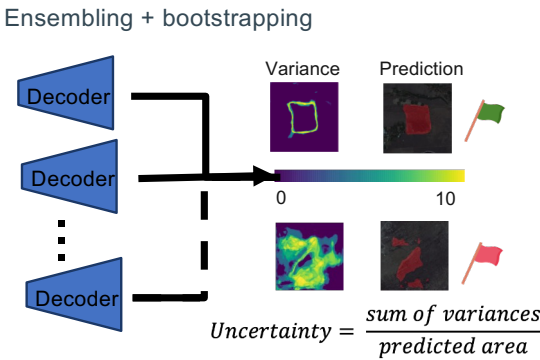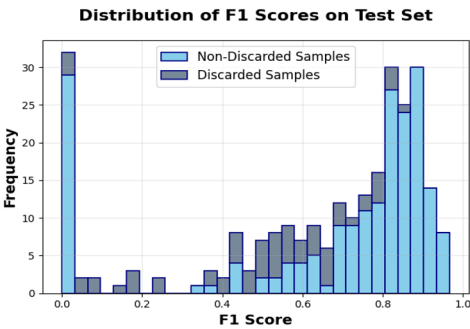## SHRUG-FM

Out-of-Distribution Detection

Pre-training embeddings   Known embedding space clusters

K-Means clustering

New embeddings   Nearest Centroid Distance Deficit informed flagging

Mean F-score

Mean NCDD

Model Uncertainty Quantification

Ensembling + bootstrapping

Variance   Prediction

Decoder

Decoder

Decoder

0   10

$$Uncertainty = \frac{sum\ of\ variances}{predicted\ area}$$

Discard-based evaluation to monitor practical utility of uncertainty estimates

Distribution of F1 Scores on Test Set
- Non-Discarded Samples
- Discarded Samples

Frequency

F1 Score

Combined reliability flags

Data within distribution
+
Low model uncertainty

Data OOD
+
High model uncertainty

SHRUG-FM

Next Steps

Further Analysis:

- Extend and thoroughly evaluate selective prediction
- Extend to other downstream use cases (Floods and Landslides)
- Extend to more FMs (varying pretraining task and architectures)

Proof-of-concept dashboard

Township of Lee, Minnesota, US
Broadleaf Deciduous Forest (Closed)

Very unreliable   Too much variance   Not enough similar data

Model Reliability: 0.37
Low   High

Data Reliability: 0.48
Low   High