Deep Learning Based Mineral Dust Detection and Feature Selection

Ping Hou[1], Peng Wu[2]

Research Assistant: Pei Guo[3]
Faculty Mentor: Aryya Gangopadhyay[3]

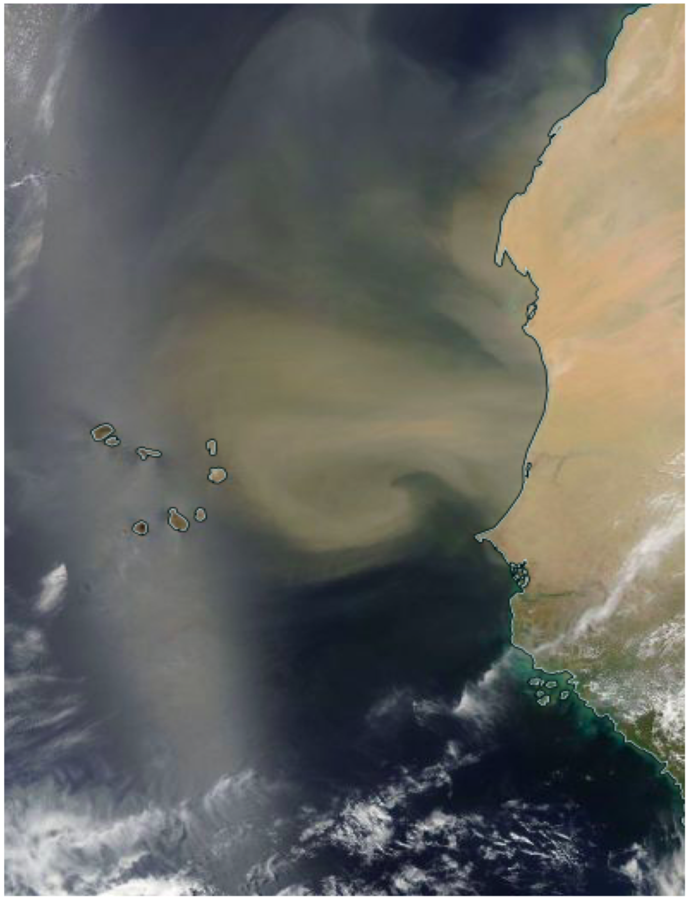
[1] School for Environment and Sustainability, University of Michigan
[2] Department of Hydrology and Atmospheric Sciences, University of Arizona
[3] Department of Information Systems, UMBC

Acknowledgements: NSF, UMBC, HPCF

UMBC
AN HONORS UNIVERSITY IN MARYLAND

CyberTraining: Big Data + High-Performance Computing + Atmospheric Sciences
University of Maryland, Baltimore County, Spring 2019

Background - Mineral Dust Aerosol



NASA Worldview, May 9 2007:
https://go.nasa.gov/2IlNV7r

- Defined as soil particles in the air
- Adversely affect air quality and human health
- Change temperature structure in the atmosphere

Background - Physical methods

Algorithm	Aerosol type	Threshold 1	Threshold 2	Others
Deep blue	Dust	AAI > 10.0	DSDI ≥ 0.0	—
	Thin smoke	AAI ≥ 5.0	DSDI ≤ –3.0	—
	Thick smoke	AAI ≥ 9.0	DSDI ≤ –2.0	$0.2 < R_{0.41} < 0.4$
IR-visible	Thin dust	$BT_{10.8} - BT_{12.0} \leq -0.2$	$BT_{3.7} - BT_{10.8} \geq 15$	$R_{1.38} < 0.035$, MNDVI < 0.8, RAT2 > 0.005
	Thin dust	—	$BT_{3.7} - BT_{10.8} \geq 20$	—
	Thick dust	$BT_{10.8} - BT_{12.0} \leq -0.2$	$BT_{3.7} - BT_{10.8} \geq 20$	$R_{1.38} < 0.035$, MNDVI < 0.2
Dust RGB	Dust	$BT_{12.0} - BT_{10.8} > 0$	$BT_{10.8} - BT_{8.7} < 0.5$ in western CONUS–Mexico	$BT_{10.8} > 273$
			$BT_{10.8} - BT_{8.7} < 4$ in North Africa and Arabian Peninsula	

Zhang et al. 2018

Data: Input features

Feature index	Name	Center (microns)
1	M01	0.415
2	M02	0.445
3	M03	0.49
4	M04	0.555
5	M05	0.673
6	M06	0.746
7	M07	0.865
8	M08	1.24
9	M09	1.378
10	M10	1.61

Feature index	Name	Center (microns)
11	M11	2.25
12	M12	3.7
13	M13	4.05
14	M14	8.55
15	M15	10.763
16	M16	12.013

Background – previous project in 2018

Table 4.1: Performance comparison among different learning methods:dust detection along CALIPSO track

Method	Accuracy
Random Forest	79.8%
Logistic regression	83.9%
ANN	64.7%
SVM	65.8%
Stacking classifiers(RF, LR, ANN,SVM)	75.6%

- ANN: 1 hidden layer with 5 nodes
- (MODIS) Data length: two day

Background – previous project in 2018

Table 4.2: Performance comparison using different number of variables: dust detection along CALIPSO track

Models	Accuracy
July 15,2007 data: 70% for training, 30% testing	
Physical algorithm	0.554
All band variables	0.924
Selected 16 band variables based on machine learning	0.929
Selected 16 band variables + 4 variables based on physical algorithm	0.931
Selected 16 band variables + 4 sensor angle variables	0.925
July 15,2007 data for training , June 22,2009 data for testing	
Physical algorithm	0.423
All band variables	0.832
Selected 16 variables based on machine learning	0.820
Selected 16 variables + 4 variables based on physical algorithm	0.835
Selected 16 band variables + 4 sensor angle variables	0.809

Problem definition

- Train a deep learning model with a larger dataset (One month in current project)

- Feature selection: find most important features for mineral dust detection
  - Select a best subset of features that have similar or better performance compared with using all features;
  - Identify important features and explain reasons.

Methods

- Train a deep learning model with a larger dataset
  - Find an appropriate structure of the deep learning model (e.g., number of hidden layers, number of neurons in each layer) - trial and error

- Find most important features for mineral dust detection
  - Shuffling procedure
  - Genetic algorithm

Satellite Data: VIIRS and CALIPSO

- VIIRS:
  - Passive sensor onboard Suomi NPP satellite
  - Images the entirety of the earth every 16 days
  - 16 moderate-resolution bands (750 m)

- CALIPSO
  - Satellite with active Lidar sensor
  - Provides robust information of dust identification

Data: Input features

Feature index	Name	Center (microns)
1	M01	0.415
2	M02	0.445
3	M03	0.49
4	M04	0.555
5	M05	0.673
6	M06	0.746
7	M07	0.865
8	M08	1.24
9	M09	1.378
10	M10	1.61

Feature index	Name	Center (microns)
11	M11	2.25
12	M12	3.7
13	M13	4.05
14	M14	8.55
15	M15	10.763
16	M16	12.013
17	solar azimuth angle	
18	solar zenith angle	
19	view azimuth angle	
20	view zenith angle	

Data: Outcomes (labels)

CALIPSO aerosol types from 532-nm lidar ratio:

dust : 40 ± 20 sr

polluted dust: 55 ± 22 sr

Dust_index (from “Pixel_Label”, COD, and AOD)

1: dust w/o cloud

2: cloud w/o dust

3: dust w/ cloud

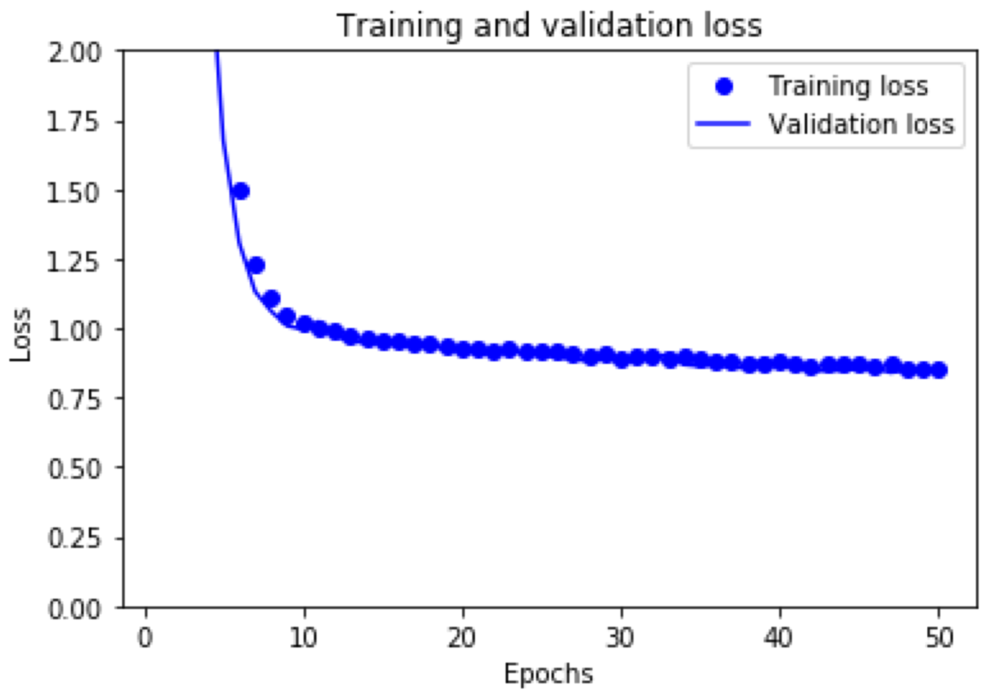
0: other

Results: deep learning model

5 hidden layers with 512 neurons in each layer

```
model = Sequential()
model.add(Dense(512, activation='relu', input_shape=(20,),kernel_regularizer=regularizers.l2(0.01)))
model.add(Dense(512, activation='relu',kernel_regularizer=regularizers.l2(0.01)))
model.add(Dense(512, activation='relu',kernel_regularizer=regularizers.l2(0.01)))
model.add(Dense(512, activation='relu',kernel_regularizer=regularizers.l2(0.01)))
model.add(Dense(512, activation='relu',kernel_regularizer=regularizers.l2(0.01)))
model.add(Dense(4, activation='softmax'))

model.compile(loss='categorical_crossentropy',
              optimizer='adam',
              metrics=['accuracy'])
```

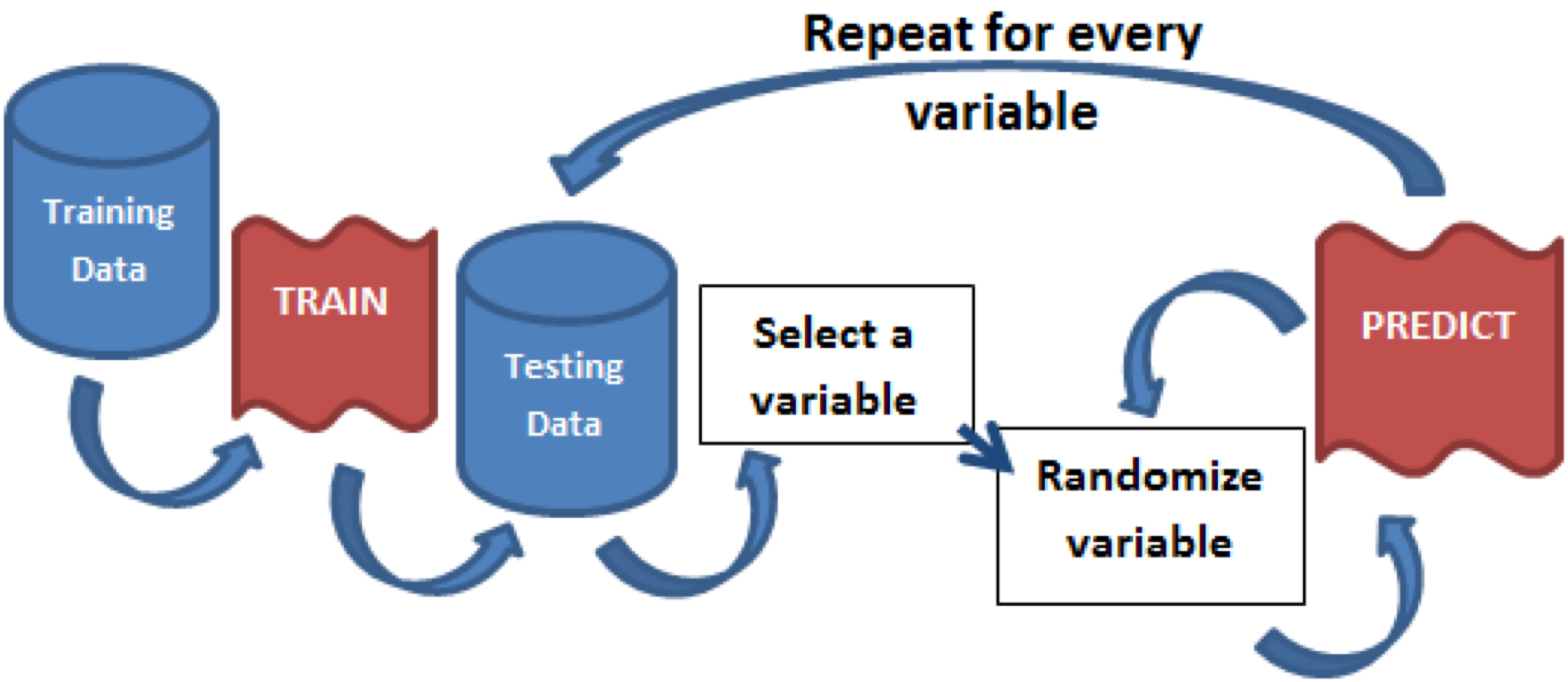


Accuracy: 71.1%

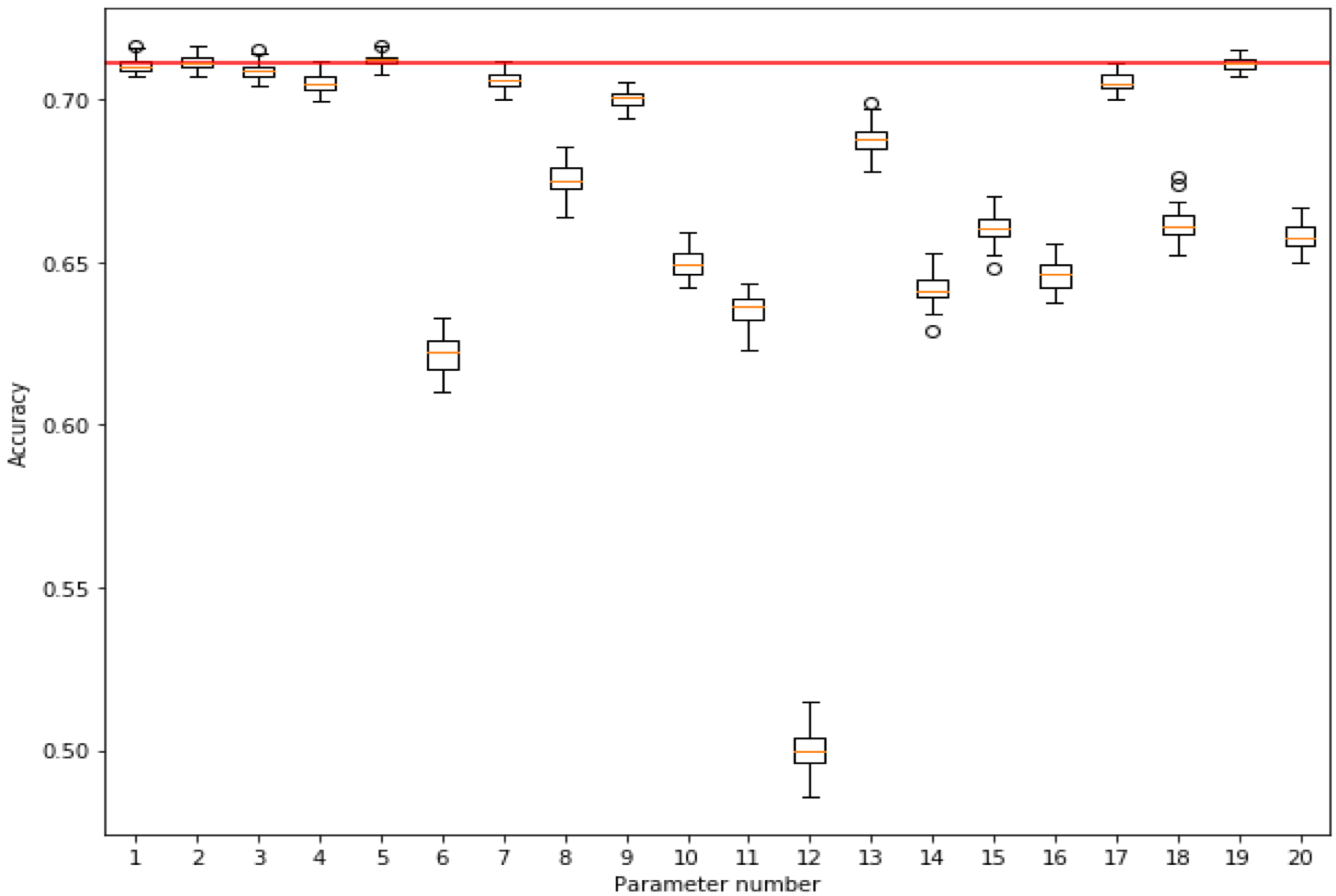
Results: confusion matrix

		Predicted labels			
		dust w/o cloud	**cloud w/o dust**	**dust w/ cloud**	**other**
True labels	**dust w/o cloud**	865	67	286	34
	cloud w/o dust	145	170	1	3
	dust w/ cloud	91	2	1066	6
	other	101	18	113	32

Feature selection by shuffling procedure

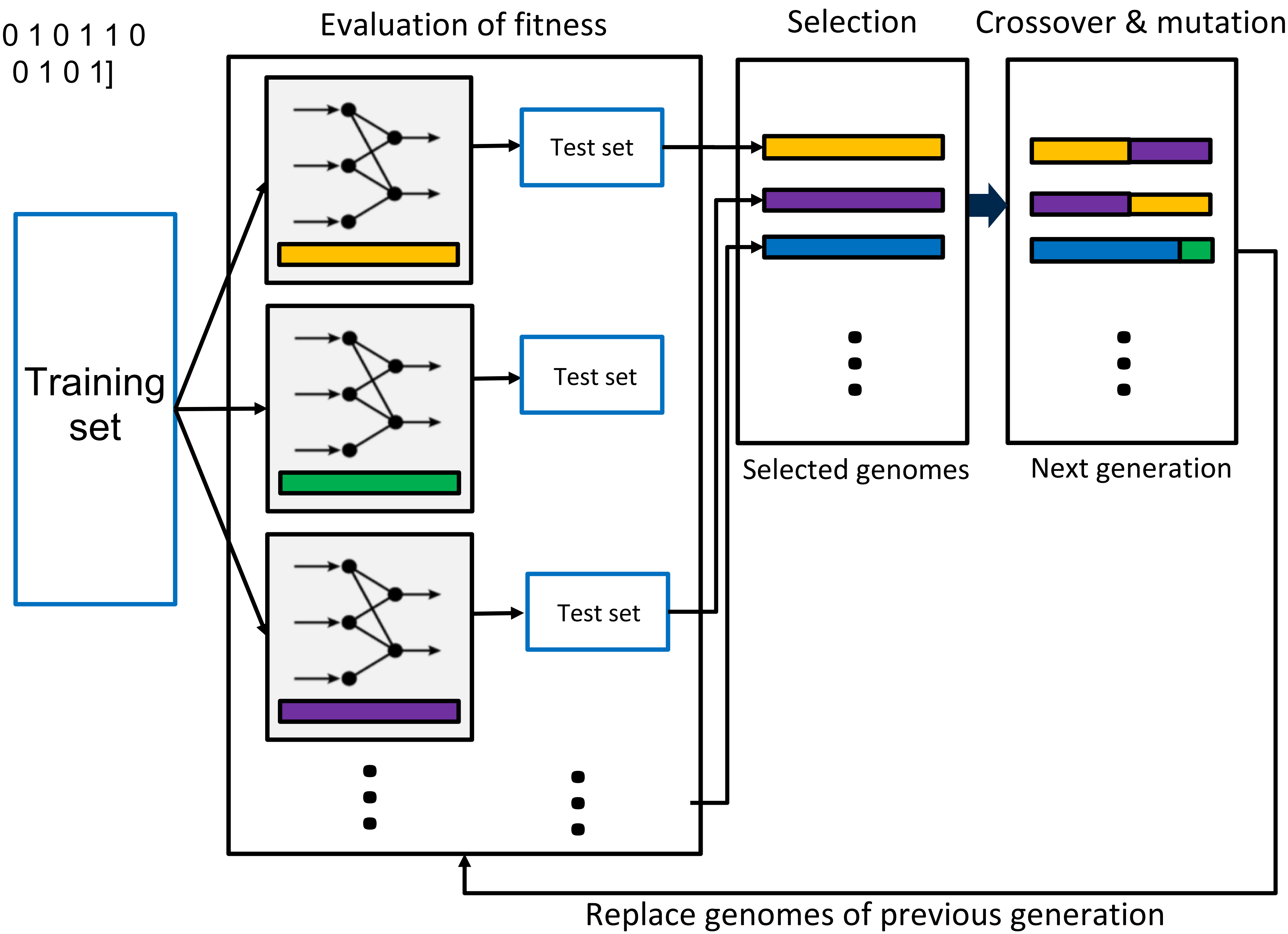


Results of shuffling procedure



Feature selection by genetic algorithm

Genome: [1 1 0 1 0 1 1 0 1 0 1 0 0 1 0 1 0 1 0 1]



Results of genetic algorithm

Population	Number of generation	Selected features	Best test accuracy
8	4	2, 3, 6, 8, 9,11,12,14,15,17,18, 19	67.8%
16	4	2, 4, 8, 11, 13, 14, 15, 17, 18, 19, 20	68.1%
32	4	1,2,3,4,5,9,10,12,13,14,16,17,18,19,20	70.3%
32	8	1, 3, 6, 8, 9, 10, 11, 17, 18, 19, 20	70.1%
64	8	1, 5, 7, 9, 12, 15, 16, 17, 18, 19, 20	71.5%

Conclusion

- A deep learning model was trained and used to classify dust and cloud using VIIRS data. The developed model have a prediction accuracy of 71%.

- Using a genetic algorithm, we find a subset of features that have a comparable accuracy.