Detecting Volcanoes on Venus

January 2019

The Data

What: 9,734 images of the surface of Venus taken by NASA’s Magellan spacecraft

Source: https://www.kaggle.com/fmena14/volcanoesvenus

Format:

- ❏ CSVs for training and test data with corresponding label spreadsheets
- ❏ Each row in the image data is one 110x110 grayscale image
- ❏ 12,100 columns with a value between 0 and 255

The Task

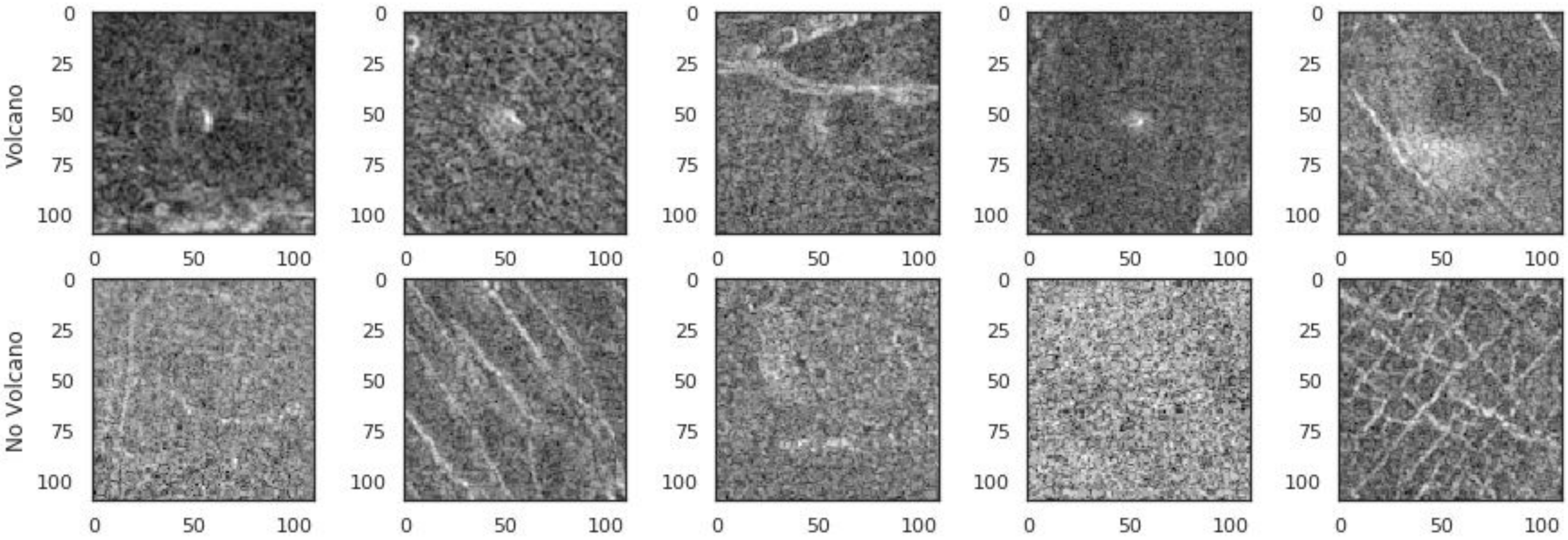
Train a model to correctly identify if a given photo has a volcano in it or not

❑ Baseline models

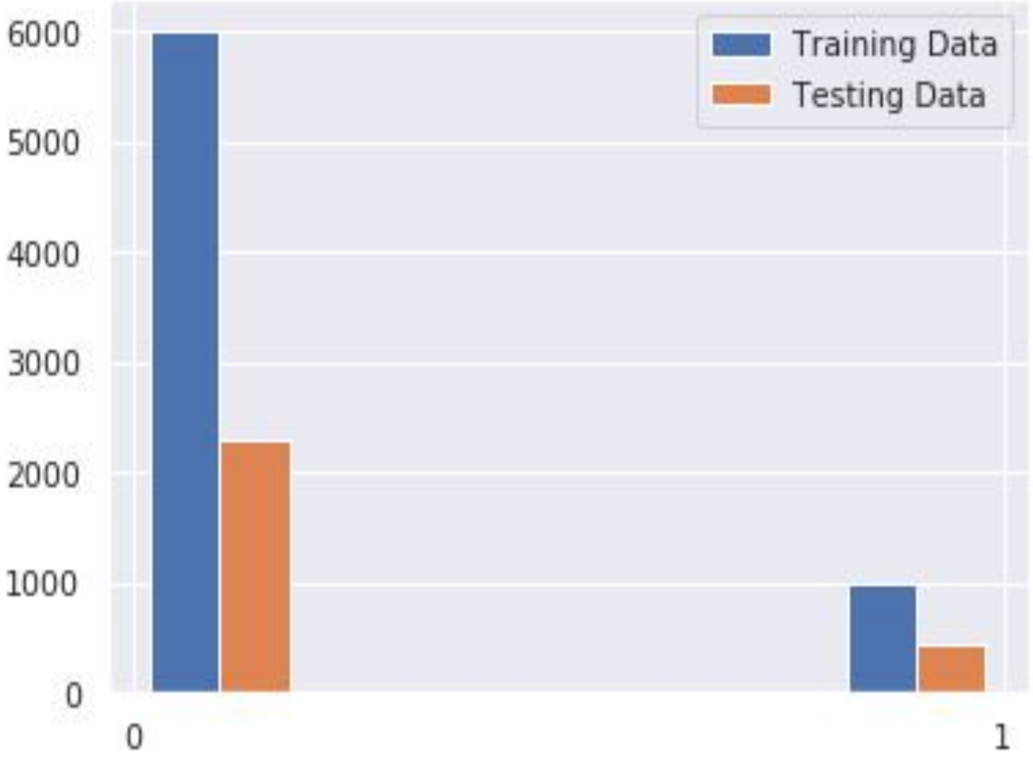
❑ Reduce dimensionality and re-train baseline models

❑ Neural networks

Samples



Distribution of the Data



Training Total Images:	7000
Training Volcanoes:	1000
Training Baseline:	**0.8571**
Testing Total Images	2734
Testing Volcanoes:	434
Testing Baseline:	**0.8413**

Baseline Models

- ❑ Random Forest
- ❑ Gradient Boosting
- ❑ XGBoost
- ❑ Lasso Regression
- ❑ Ridge Regression
- ❑ Logistic Regression
- ❑ Support Vector Machines

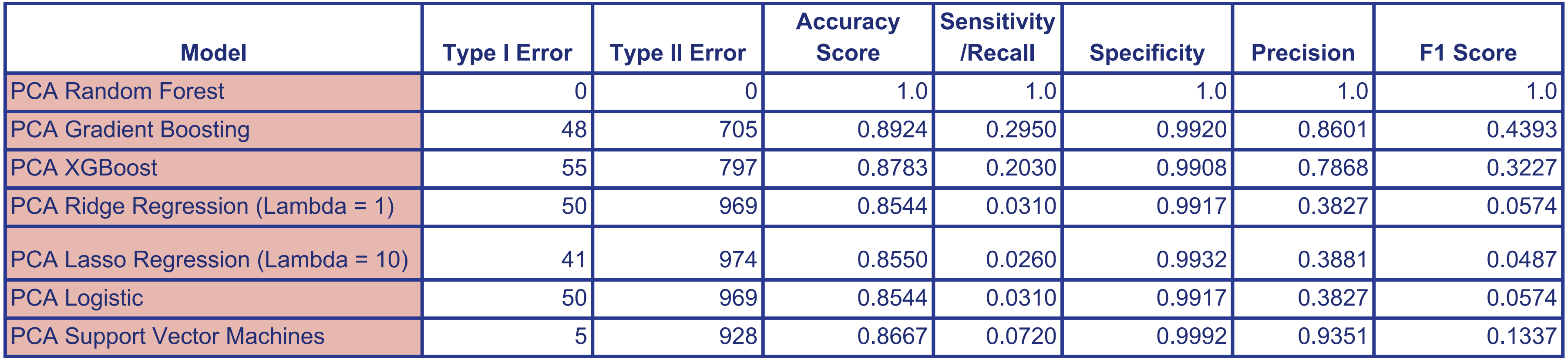
Baseline Models on Training

Model	Type I Error	Type II Error	Accuracy Score	Sensitivity/Recall	Specificity	Precision	F1 Score
Random Forest	0	0	1.0	1.0	1.0	1.0	1.0
Gradient Boosting	9	232	0.9656	0.7680	0.9985	0.9884	0.8644
XGBoost	9	208	0.9690	0.7920	0.9985	0.9888	0.8795
Ridge (Lambda = 0.01)	0	0	1.0	1.0	1.0	1.0	1.0
Lasso (Lambda = 1.0)	17	191	0.9703	0.8090	0.9972	0.9794	0.9794
Logistic	0	13	0.9981	0.9870	1.0	1.0	0.9935
Support Vector Machines	0	0	1.0	1.0	1.0	1.0	1.0

Baseline Models on Testing

Model	Type I Error	Type II Error	Accuracy Score	Sensitivity/Recall	Specificity	Precision	F1 Score
Random Forest	17	233	0.9086	0.4631	0.9926	0.9220	0.6166
Gradient Boosting	22	187	0.9236	0.5691	0.9904	0.9182	0.7027
XGBoost	22	180	0.9261	0.5853	0.9904	0.9203	0.7155
Ridge (Lambda = 0.01)	51	149	0.9268	0.6567	0.9778	0.8482	0.7403
Lasso (Lambda = 1.0)	42	154	0.9283	0.6452	0.9817	0.8696	0.7407
Logistic	46	159	0.9250	0.6336	0.9800	0.8567	0.7285
Support Vector Machines	79	147	0.9173	0.6613	0.9657	0.7842	0.7175

Baseline Models with PCA on Training



Model	Type I Error	Type II Error	Accuracy Score	Sensitivity /Recall	Specificity	Precision	F1 Score
PCA Random Forest	0	0	1.0	1.0	1.0	1.0	1.0
PCA Gradient Boosting	48	705	0.8924	0.2950	0.9920	0.8601	0.4393
PCA XGBoost	55	797	0.8783	0.2030	0.9908	0.7868	0.3227
PCA Ridge Regression (Lambda = 1)	50	969	0.8544	0.0310	0.9917	0.3827	0.0574
PCA Lasso Regression (Lambda = 10)	41	974	0.8550	0.0260	0.9932	0.3881	0.0487
PCA Logistic	50	969	0.8544	0.0310	0.9917	0.3827	0.0574
PCA Support Vector Machines	5	928	0.8667	0.0720	0.9992	0.9351	0.1337

Baseline Models with PCA on Testing

Model	Type I Error	Type II Error	Accuracy Score	Sensitivity/Recall	Specificity	Precision	F1 Score
PCA Random Forest	49	349	0.8544	0.1959	0.9787	0.6343	0.2993
PCA Gradient Boosting	74	349	0.8453	0.1959	0.9678	0.5346	0.2867
PCA XGBoost	37	370	0.8511	0.1475	0.9839	0.6337	0.2393
PCA Ridge Regression (Lambda = 1)	29	419	0.8361	0.0346	0.9874	0.3409	0.0628
PCA Lasso Regression (Lambda = 10)	22	421	0.8380	0.0300	0.9904	0.3714	0.0554
PCA Logistic	29	419	0.8361	0.0346	0.9874	0.3409	0.0628
PCA Support Vector Machines	6	414	0.8464	0.0461	0.9974	0.7692	0.0870

Neural Networks

- ❑ Multi Layer Perceptron
- ❑ Several Convolutional Neural Networks
- ❑ Image Data Generator

Neural Network Models on Training

Model	Type I Error	Type II Error	Accuracy Score	Sensitivity/Recall	Specificity	Precision	F1 Score
Multi Layer Perceptron	0	1000	0.8571	0	1.0	0	0
CNN I	25	101	0.9820	0.8990	0.9958	0.9729	0.9345
CNN II with ImageDataGenerator	0	1000	0.8571	0	1.0	0	0
CNN III	**14**	**116**	**0.9814**	**0.8840**	**0.9977**	**0.9844**	**0.9315**
CNN IV	81	334	0.9407	0.6660	0.9865	0.8916	0.7624

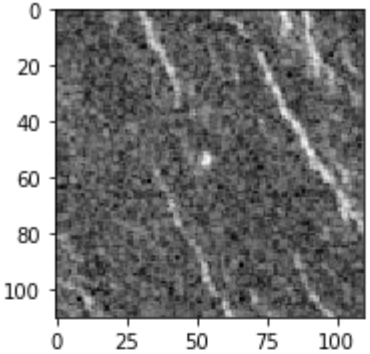
Neural Network Models on Testing

Model	Type I Error	Type II Error	Accuracy Score	Sensitivity/Recall	Specificity	Precision	F1 Score
Multi Layer Perceptron	0	434	0.8413	0	1.0	0	0
CNN I	22	71	0.9660	0.8364	0.9904	0.9429	0.8864
CNN II with ImageDataGenerator	0	434	0.8413	0	1.0	0.0	0
CNN III	**11**	**70**	**0.9704**	**0.8387**	**0.9952**	**0.9707**	**0.8999**
CNN IV	35	158	0.9294	0.6359	0.9848	0.8875	0.7409

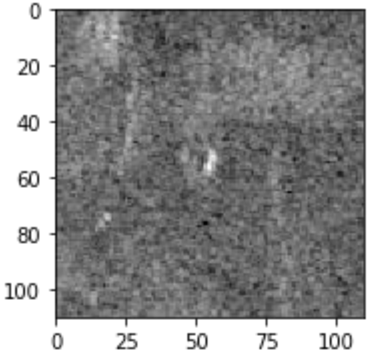
Elements of the Best NN

- ❑ Four convolutional layers
- ❑ Two 32 and two 64-perceptron wide blocks
- ❑ Dropout of 0.1 after the third layer
- ❑ Adam Optimizer
- ❑ 15 epochs
- ❑ ~120 seconds per epoch

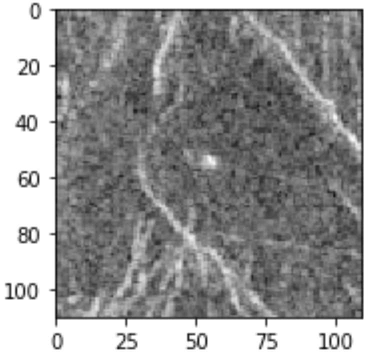
Analyzing Results from the Best NN



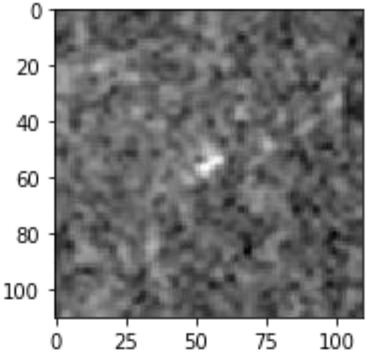
```
1    0.999972
Name: 329, dtype: float64
```



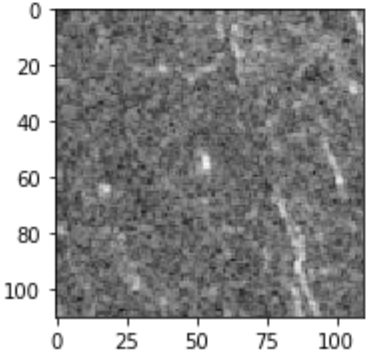
```
1    0.999973
Name: 543, dtype: float64
```



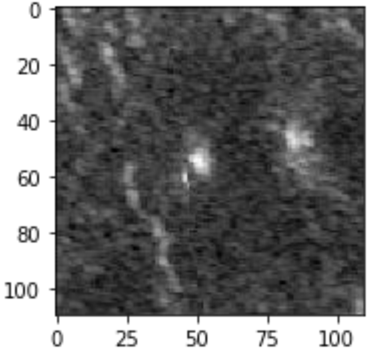
```
1    0.999979
Name: 692, dtype: float64
```



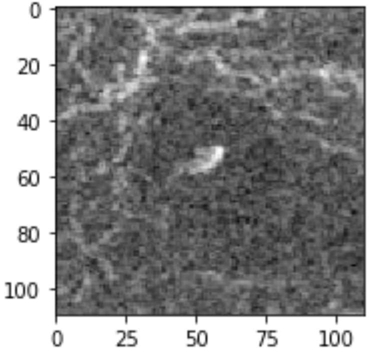
```
1    0.999964
Name: 1041, dtype: float64
```



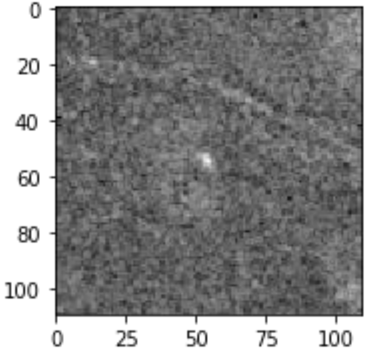
```
1    0.999971
Name: 1223, dtype: float64
```



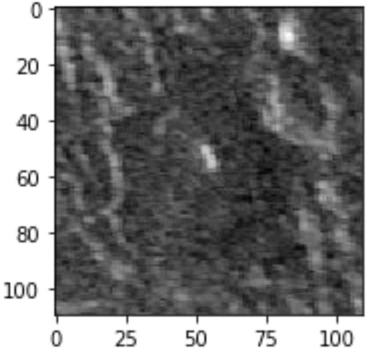
```
1    0.999995
Name: 1420, dtype: float64
```



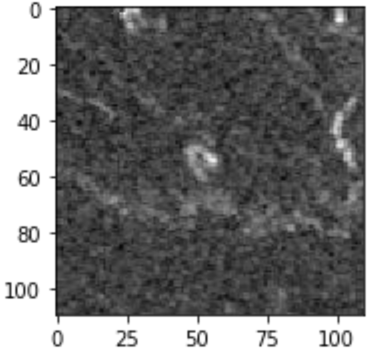
```
1    0.999999
Name: 1607, dtype: float64
```



```
1    0.999979
Name: 1834, dtype: float64
```

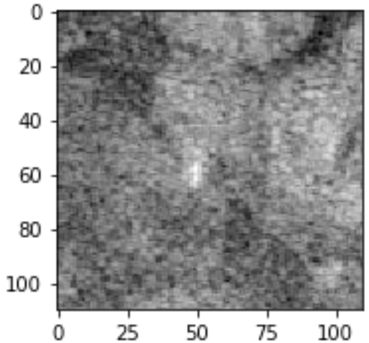


```
1    0.999976
Name: 2351, dtype: float64
```

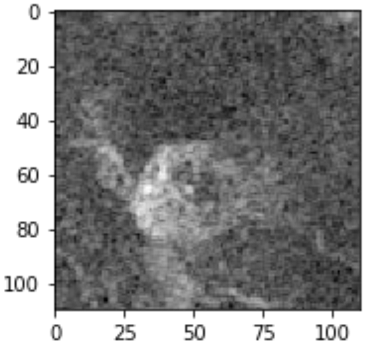


```
1    0.999989
Name: 2543, dtype: float64
```

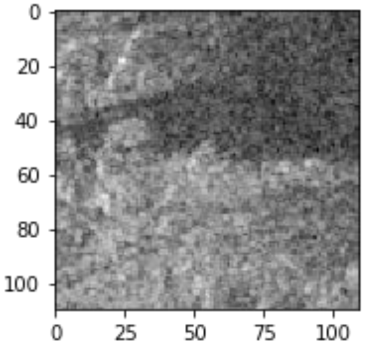
Type I Errors



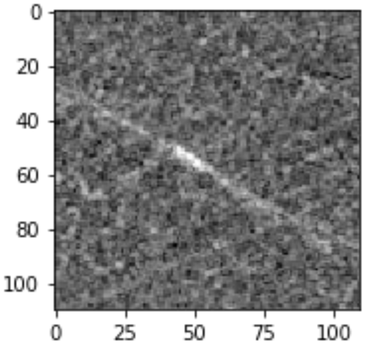
Volcano=0.0
Prediction=1.0
Probability=0.9452041



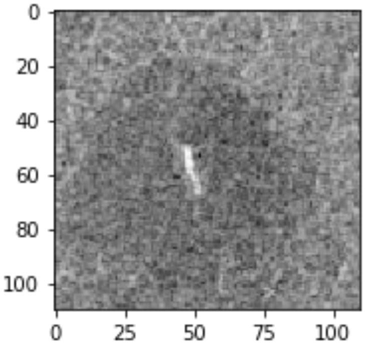
Volcano=0.0
Prediction=1.0
Probability=0.90660644



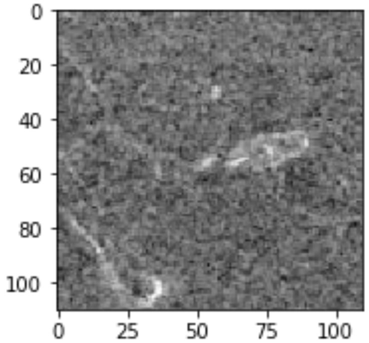
Volcano=0.0
Prediction=1.0
Probability=0.9536055



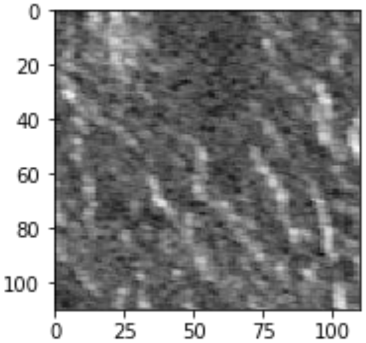
Volcano=0.0
Prediction=1.0
Probability=0.7047649



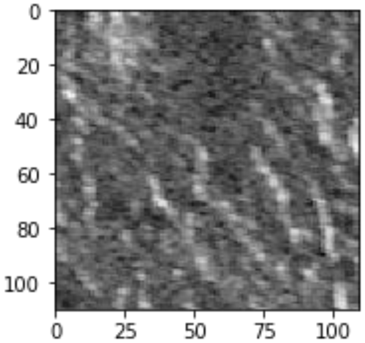
Volcano=0.0
Prediction=1.0
Probability=0.928642



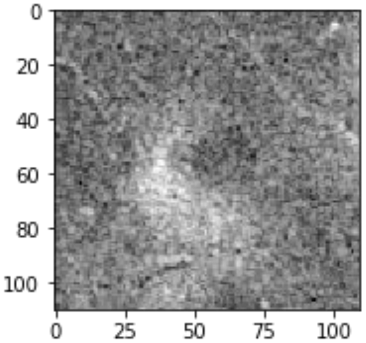
Volcano=0.0
Prediction=1.0
Probability=0.74646395



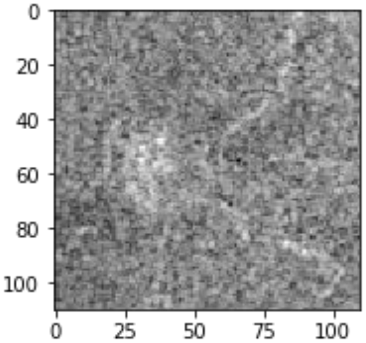
Volcano=0.0
Prediction=1.0
Probability=0.9909244



Volcano=0.0
Prediction=1.0
Probability=0.9909244

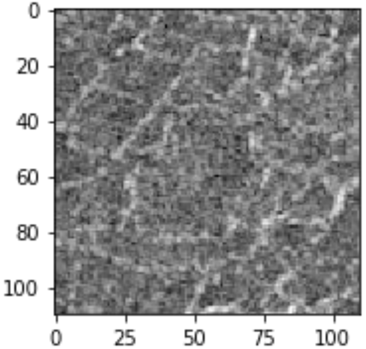


Volcano=0.0
Prediction=1.0
Probability=0.5645607

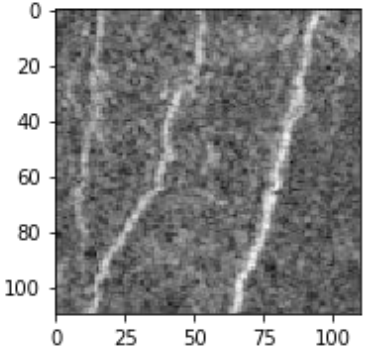


Volcano=0.0
Prediction=1.0
Probability=0.57234716

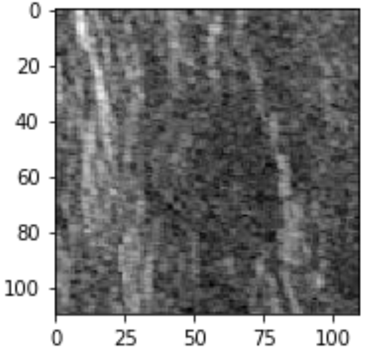
Type II Errors



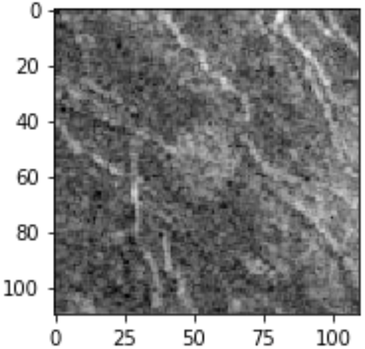
Volcano=1.0
Prediction=0.0
Probability=0.9998109



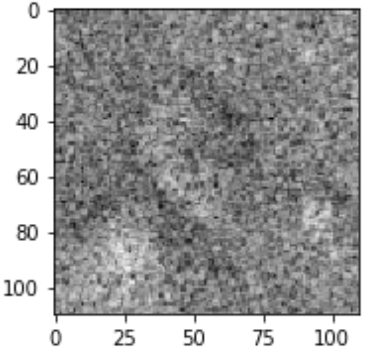
Volcano=1.0
Prediction=0.0
Probability=0.9873268



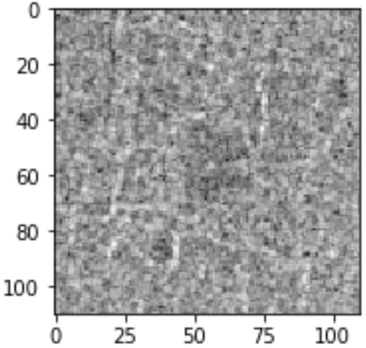
Volcano=1.0
Prediction=0.0
Probability=0.99808013



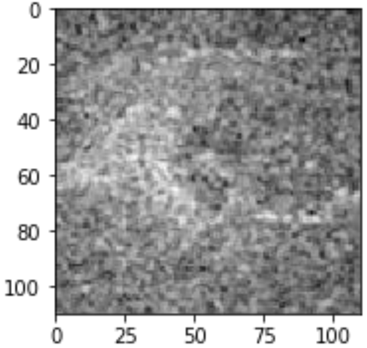
Volcano=1.0
Prediction=0.0
Probability=0.9974285



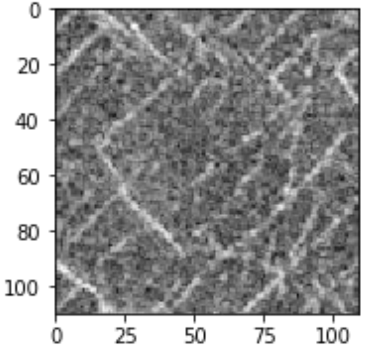
Volcano=1.0
Prediction=0.0
Probability=0.9613874



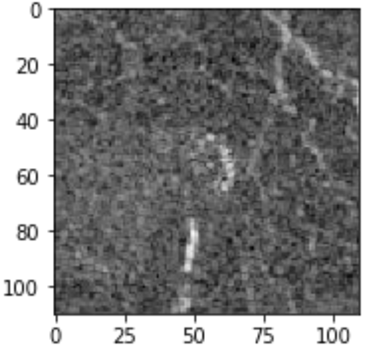
Volcano=1.0
Prediction=0.0
Probability=0.98739547



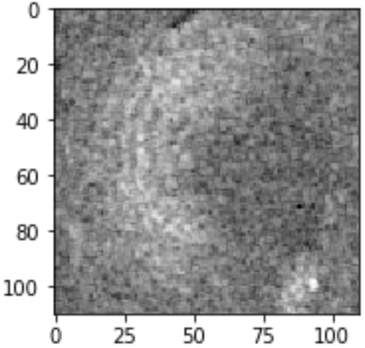
Volcano=1.0
Prediction=0.0
Probability=0.813185



Volcano=1.0
Prediction=0.0
Probability=0.99868625



Volcano=1.0
Prediction=0.0
Probability=0.9909985



Volcano=1.0
Prediction=0.0
Probability=0.7538748

Possible Next Steps

❏ Oversampling the minority class (SMOTE)
❏ Sklearn extract_patches_2d
❏ VGG16 or VGG19
❏ Test on other topographic imagery

Questions?