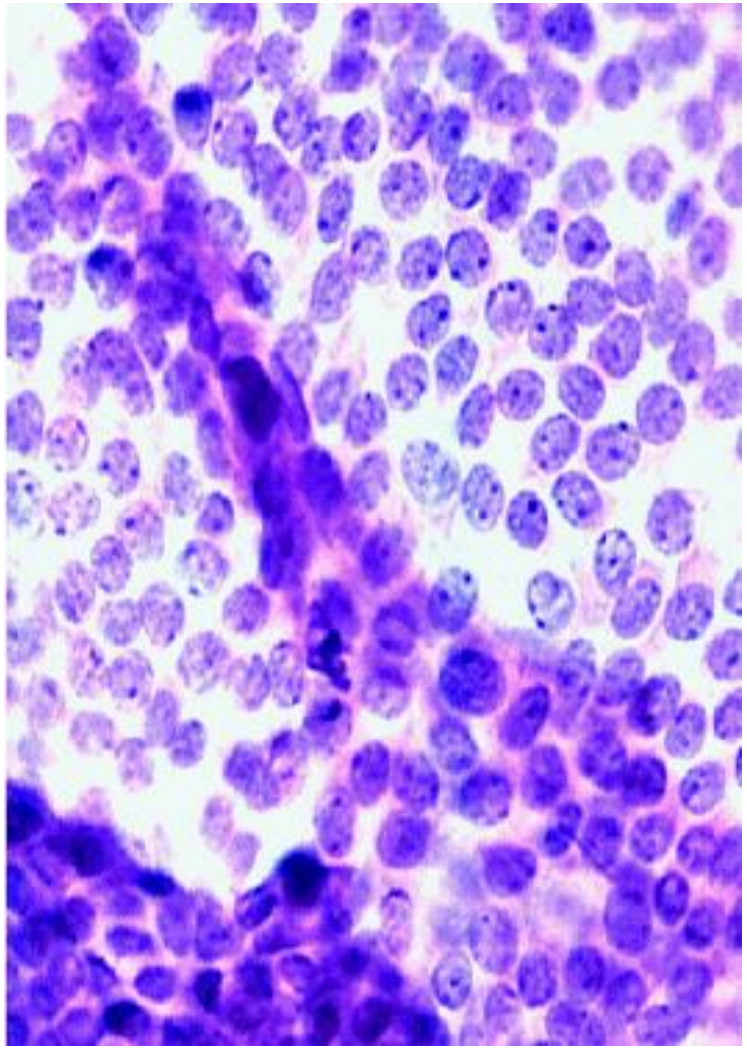
XGBClassifier Most Favorable for Breast Cancer Diagnosis

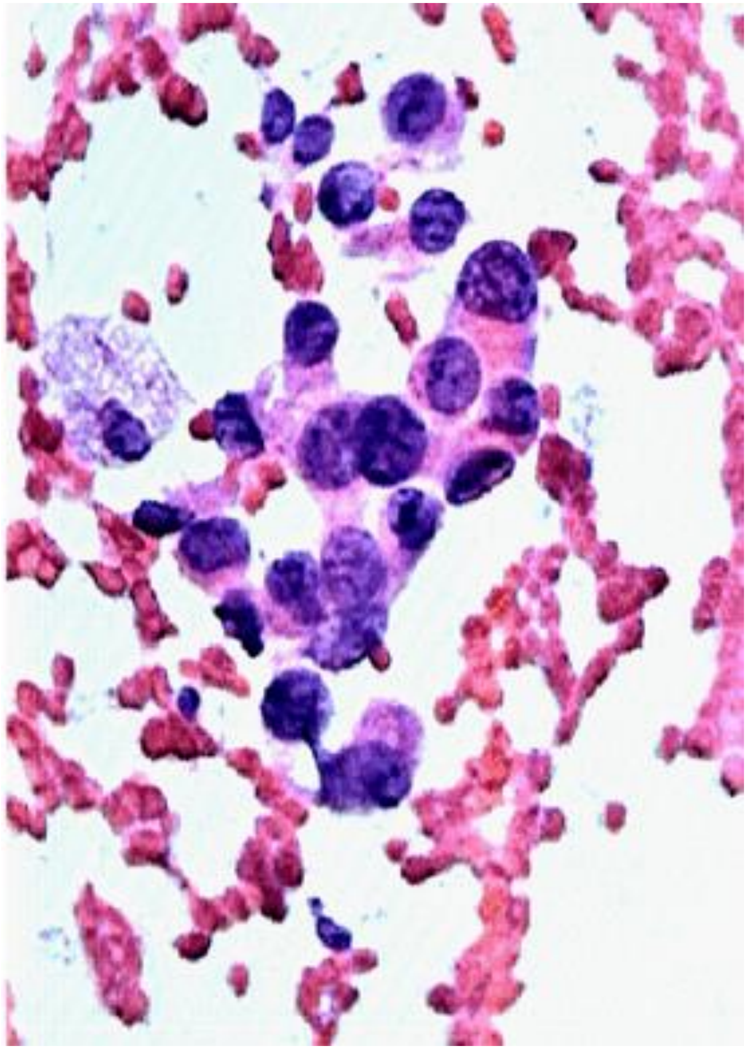
Rachel Khoo

Thinkful Capstone 2

September 2020

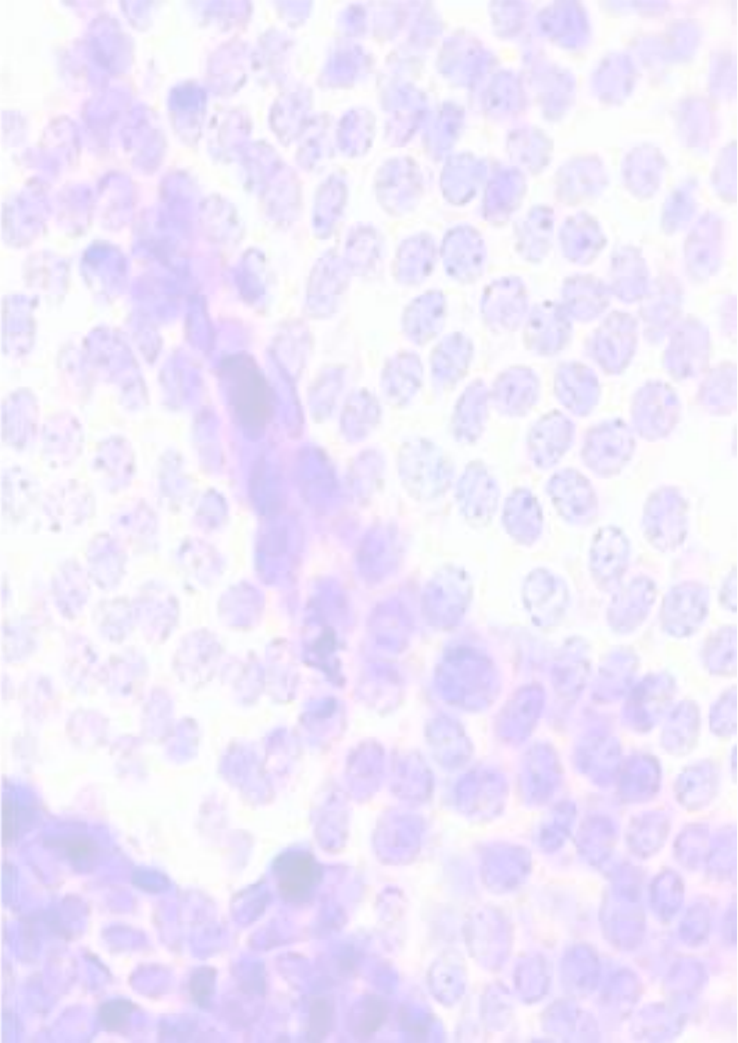
Who has cancer?

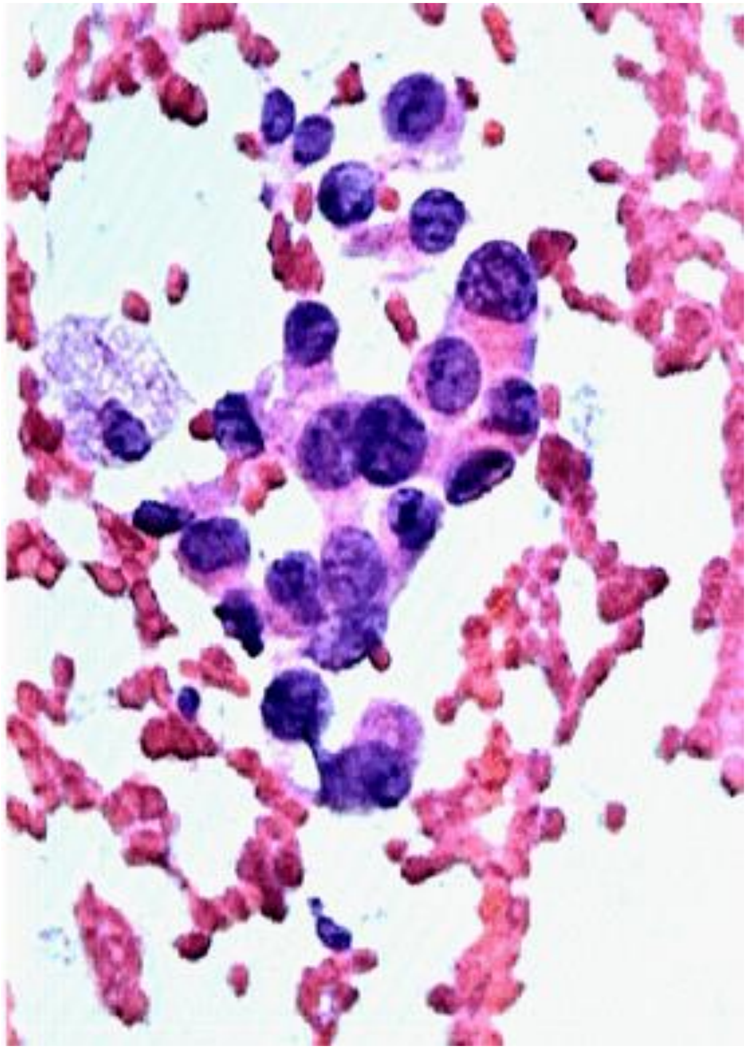




Eickhoff, Carsten. (2014). Crowd-powered experts: helping surgeons interpret breast cancer images. ACM International Conference Proceeding Series. 53-56. 10.1145/2594776.2594788.

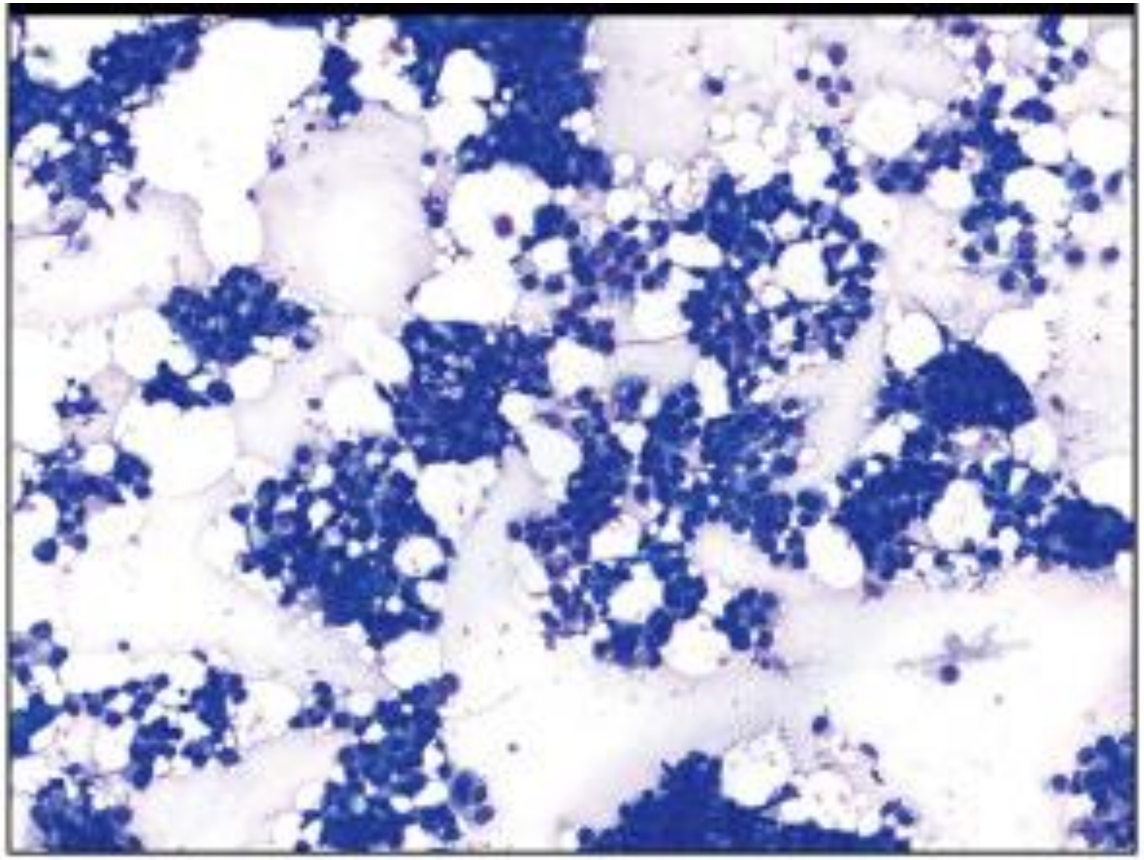
Who has cancer?





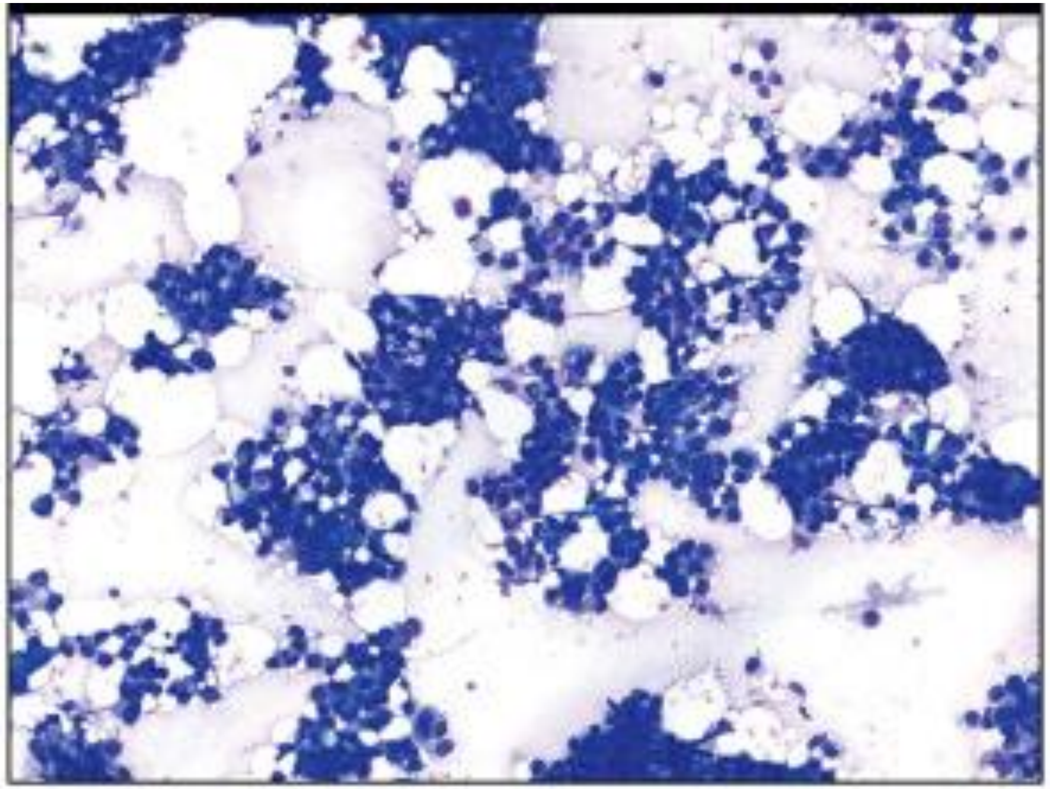
Eickhoff, Carsten. (2014). Crowd-powered experts: helping surgeons interpret breast cancer images. ACM International Conference Proceeding Series. 53-56. 10.1145/2594776.2594788.

Who has cancer?



https://www.captodayonline.com/cytopathology-standardized-reporting-breast-fnab-cytology/

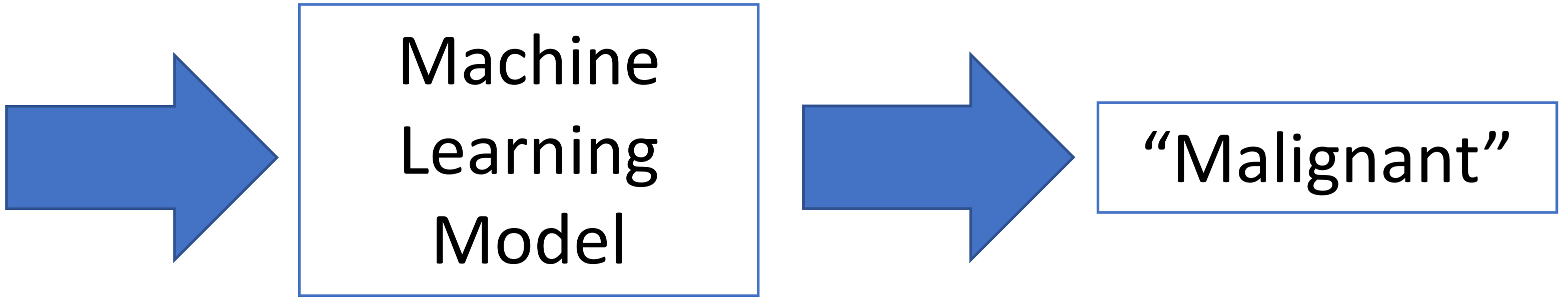
Machine Learning can make diagnosis easier



- Area: 1001.0
- Texture: 10.38
- Compactness: 0.27760
- Concavity: 0.3001

Machine Learning can make diagnosis easier

- Area: 1001.0
- Texture: 10.38
- Compactness: 0.27760
- Concavity: 0.3001



Machine Learning Model

“Malignant”

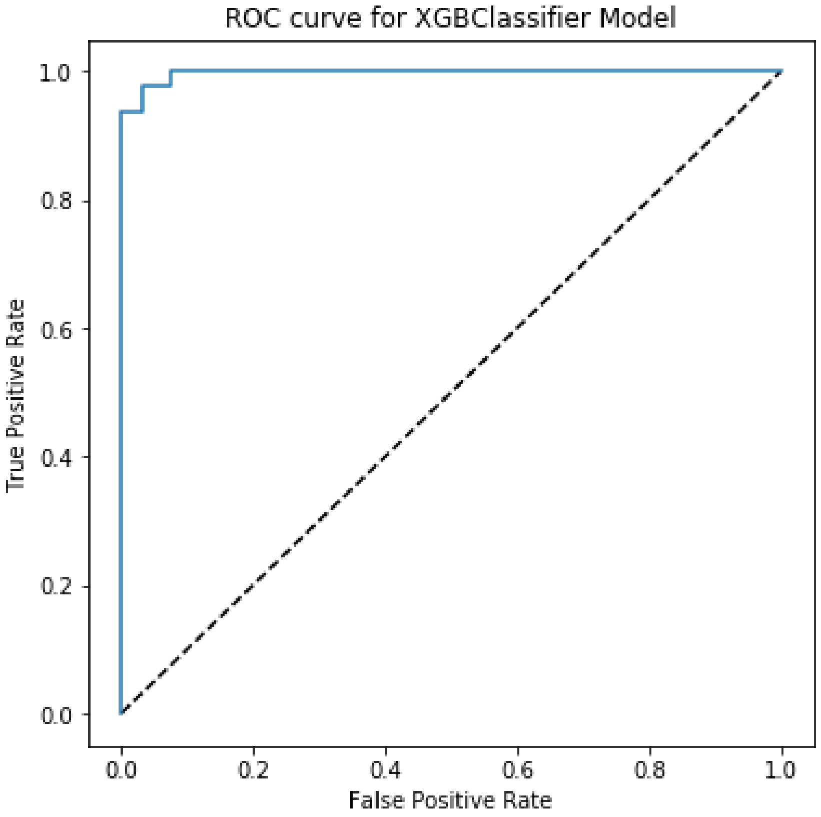
XGBClassifier is the best model

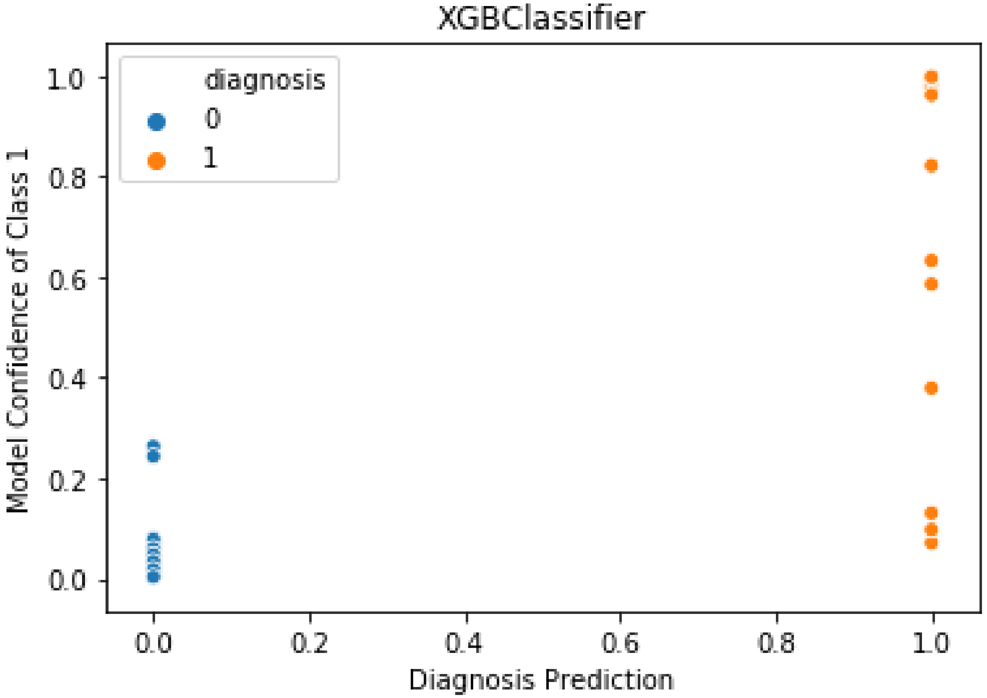
- AUPRC: 0.9699

- Recall: 90%

	Predicted Benign	Predicted Malignant
True Benign	68	4
True Malignant	4	38

XGBClassifier is confident and accurate





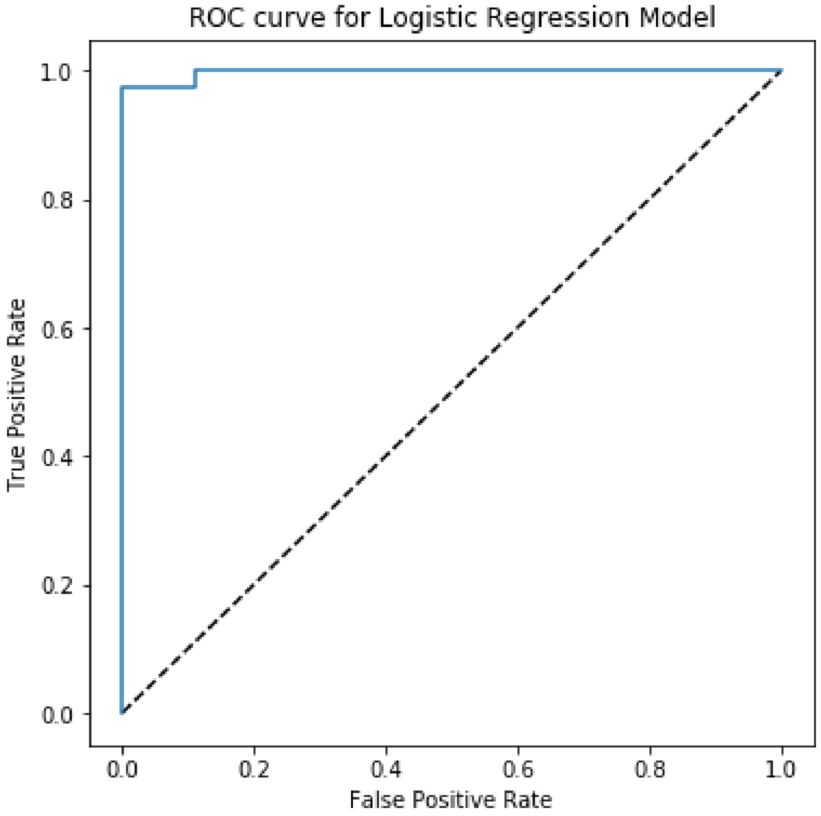
Logistic Regression: Accuracy isn’t everything

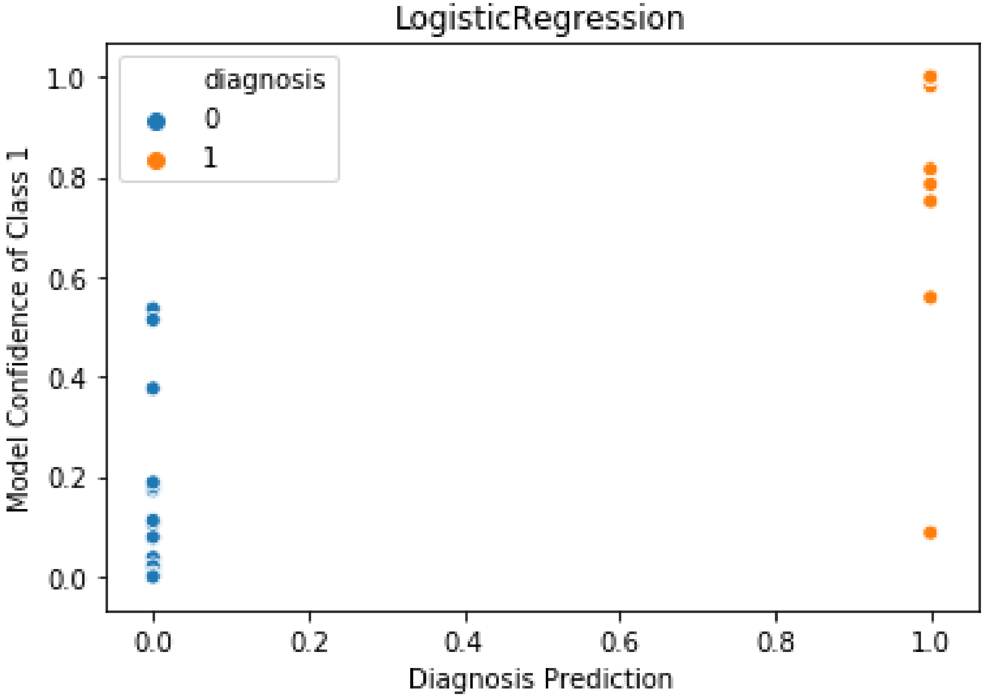
- AUPRC: 0.9466

- Recall: 93%

	Predicted Benign	**Predicted Malignant**
True Benign	71	1
True Malignant	3	39

Logistic Regression is more confidently wrong





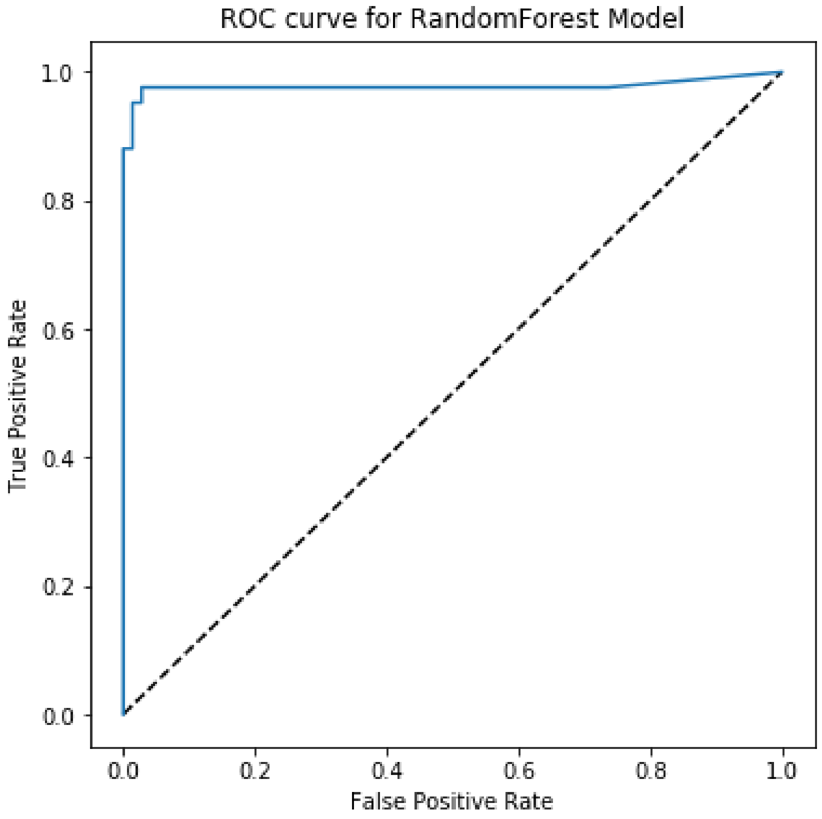
Random Forest: good accuracy, bad recall

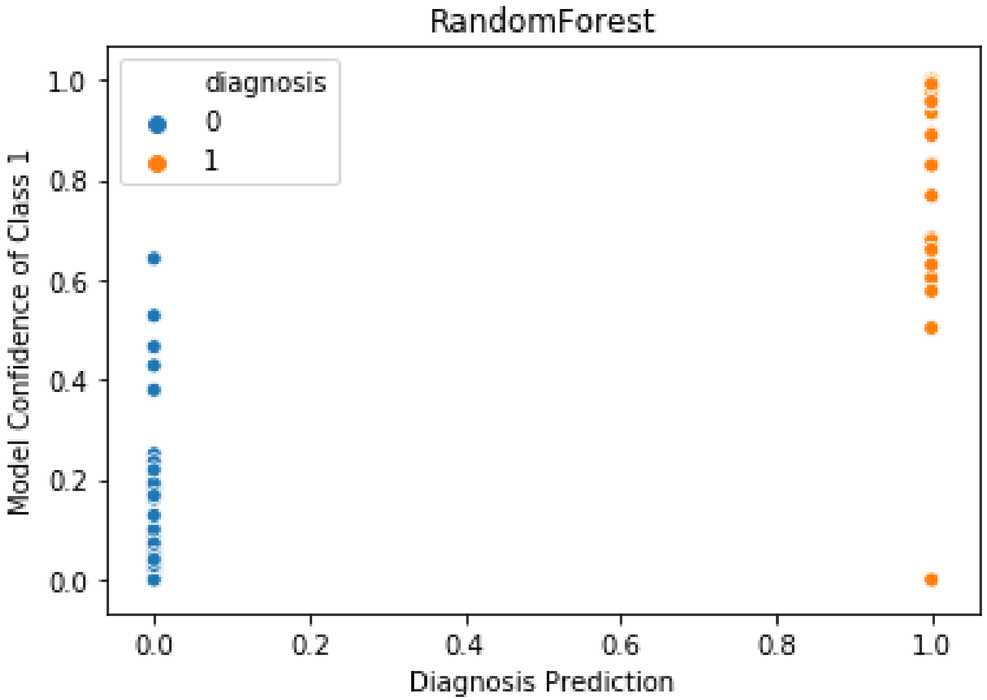
- AUPRC: 0.9395

- Recall: 88%

	Predicted Benign	Predicted Malignant
True Benign	71	1
True Malignant	5	37

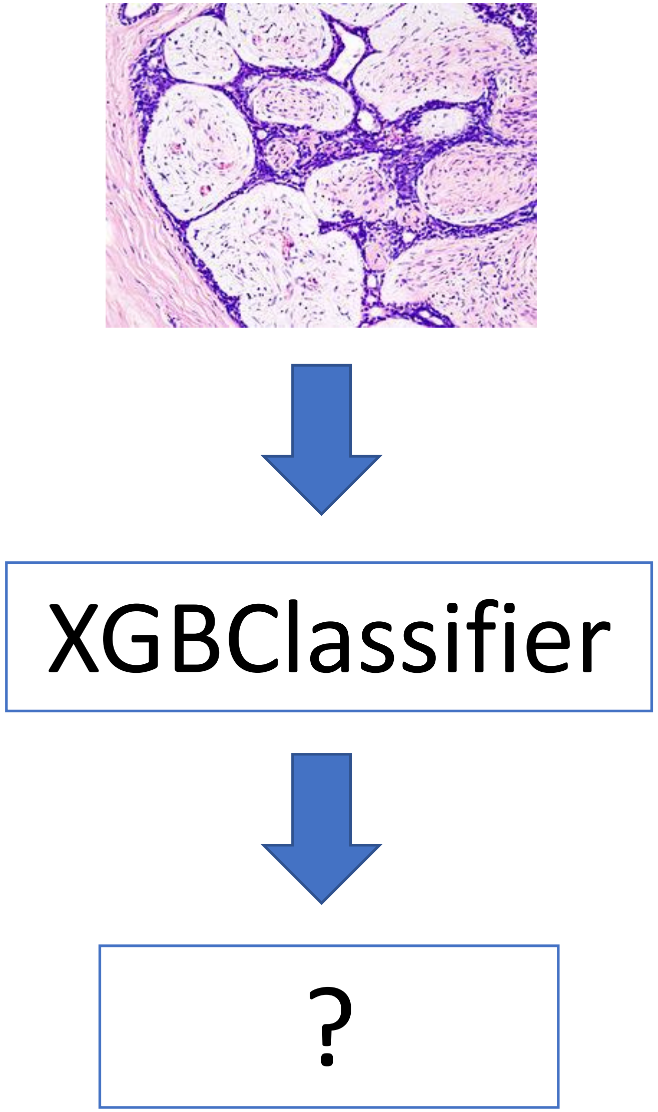
RandomForest can still be useful





Every model has limitations

- XGB
  - Can’t predict outside of sample
- Logistic Regression
  - Can be slow
- KNN
  - Slower
  - Not easily interpretable
  - Can’t predict outside of sample

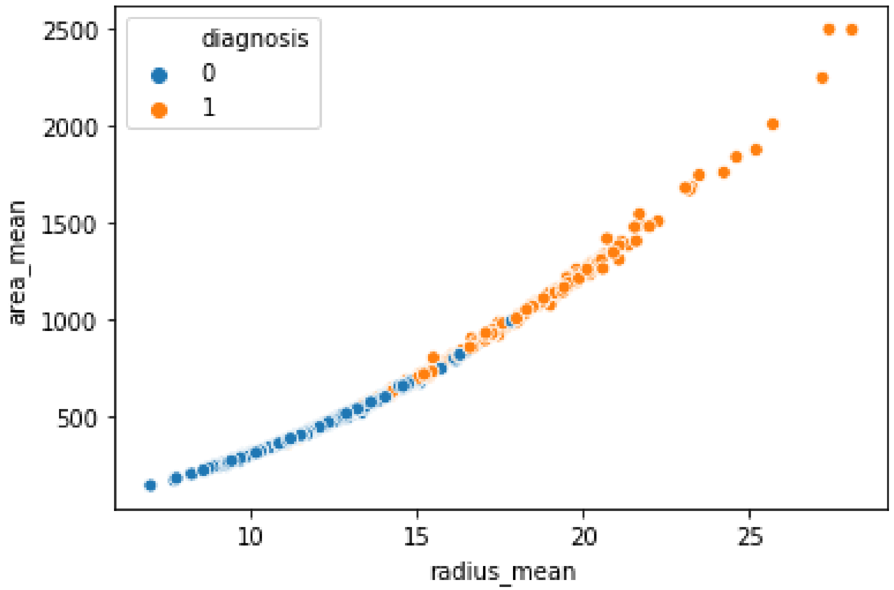


XGBClassifier

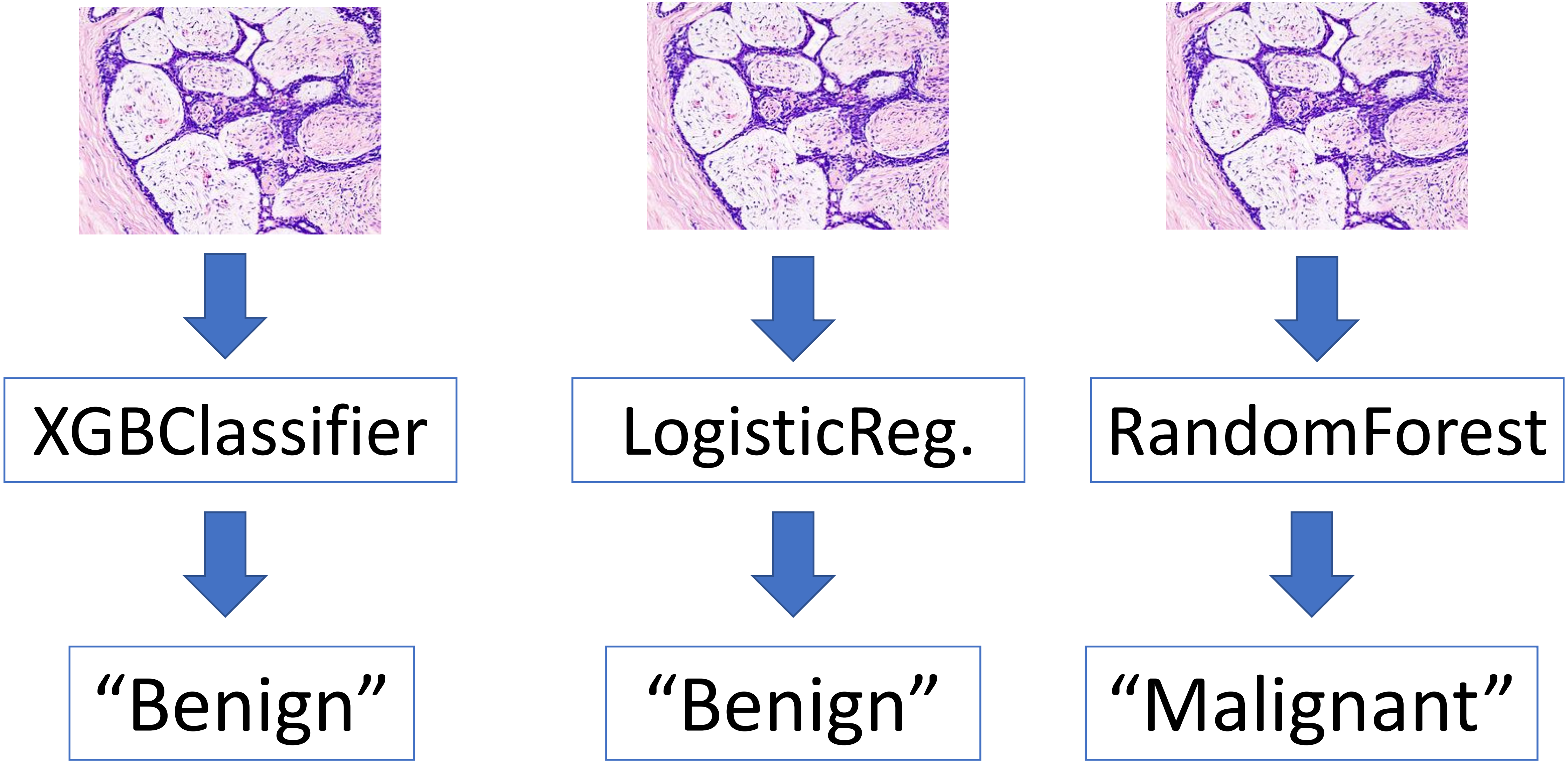
?

How can we improve accuracy?

- More data
- Spend more time tuning hyperparameters
- PCA to reduce complexity redundancy



An ensemble method could be even better



XGBClassifier

“Benign”

LogisticReg.

“Benign”

RandomForest

“Malignant”