May 2022

TEXAS
The University of Texas at Austin

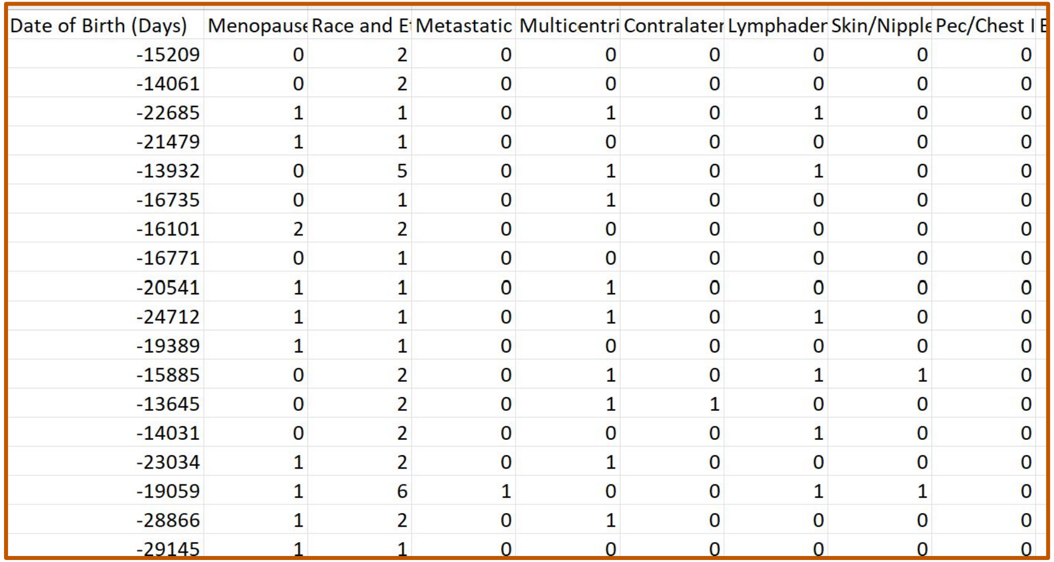
Multimodal Learning for Breast Tumor Molecular Subtype Diagnosis

Greg Holste, Adrian Velasquez, Ever Olivares, Michelle Sanchez Guererro, Olivia Parker
The University of Texas at Austin

Our Dataset

Duke Breast MRI Data

- 922 cases:
    - Images
    - 7 important Clinical Characteristics
    - Radiomics Data



Date of Birth (Days)	Menopause	Race and E	Metastatic	Multicentri	Contralater	Lymphader	Skin/Nipple	Pec/Chest I	E
-15209	0	2	0	0	0	0	0	0	
-14061	0	2	0	0	0	0	0	0	
-22685	1	1	0	1	0	1	0	0	
-21479	1	1	0	0	0	0	0	0	
-13932	0	5	0	1	0	1	0	0	
-16735	0	1	0	1	0	0	0	0	
-16101	2	2	0	0	0	0	0	0	
-16771	0	1	0	0	0	0	0	0	
-20541	1	1	0	1	0	0	0	0	
-24712	1	1	0	1	0	1	0	0	
-19389	1	1	0	0	0	0	0	0	
-15885	0	2	0	1	0	1	1	0	
-13645	0	2	0	1	1	0	0	0	
-14031	0	2	0	0	0	1	0	0	
-23034	1	2	0	1	0	0	0	0	
-19059	1	6	1	0	0	1	1	0	
-28866	1	2	0	1	0	0	0	0	
-29145	1	1	0	0	0	0	0	0	

Preprocessing & Task

Goal: Predict “luminal-like” molecular subtype of patient tumor

Data Split: 70% train | 10% val | 20% test

Image Data: 64x64 tumor images

- Maximum intensity projection of (3D) MRI scan

Clinical Data: 7 patient features

- Ex: age, race/ethnicity, menopause status

Radiomics Data: 530 “radiomics” features

- Handcrafted local descriptors of tumor image

May 2022

TEXAS
The University of Texas at Austin

Analysis 1: Clinical Characteristics

Adrian Velasquez, Olivia Parker
The University of Texas at Austin

Clinical Data Analysis- Logistic Regression Model

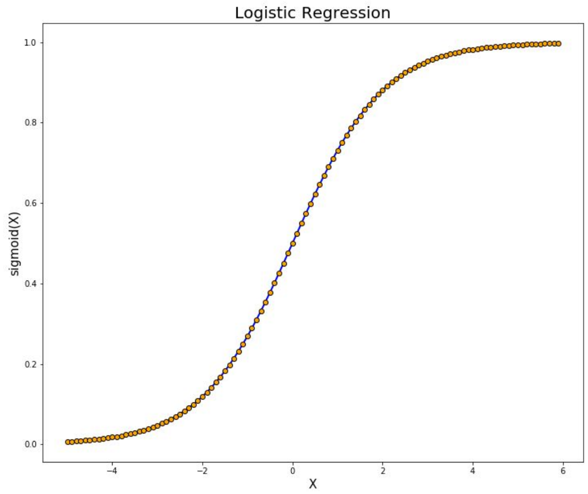
Purpose: Determining a baseline accuracy using given clinical data.

Data input:

- Demographics: race, ethnicity etc.
- Menopause status
- Tumor characteristics: metastatic, multicentric etc.

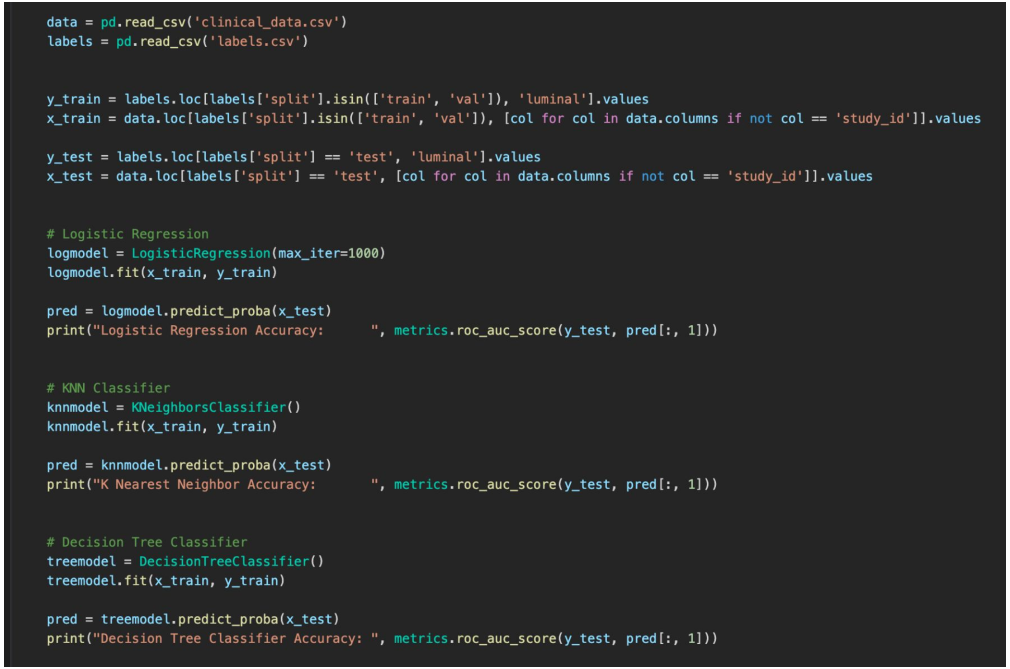
Variable Tracked

- Luminal-Like Molecular Subtype



Code Structure

1. Filter and parse data
2. Logistic regression analysis
3. K-nearest neighbor analysis
4. Decision Tree classifier



```
data = pd.read_csv('clinical_data.csv')
labels = pd.read_csv('labels.csv')


y_train = labels.loc[labels['split'].isin(['train', 'val']), 'luminal'].values
x_train = data.loc[labels['split'].isin(['train', 'val']), [col for col in data.columns if not col == 'study_id']].values

y_test = labels.loc[labels['split'] == 'test', 'luminal'].values
x_test = data.loc[labels['split'] == 'test', [col for col in data.columns if not col == 'study_id']].values


# Logistic Regression
logmodel = LogisticRegression(max_iter=1000)
logmodel.fit(x_train, y_train)

pred = logmodel.predict_proba(x_test)
print("Logistic Regression Accuracy:      ", metrics.roc_auc_score(y_test, pred[:, 1]))


# KNN Classifier
knnmodel = KNeighborsClassifier()
knnmodel.fit(x_train, y_train)

pred = knnmodel.predict_proba(x_test)
print("K Nearest Neighbor Accuracy:       ", metrics.roc_auc_score(y_test, pred[:, 1]))


# Decision Tree Classifier
treemodel = DecisionTreeClassifier()
treemodel.fit(x_train, y_train)

pred = treemodel.predict_proba(x_test)
print("Decision Tree Classifier Accuracy: ", metrics.roc_auc_score(y_test, pred[:, 1]))
```

Results

Logistic Regression	K-Nearest Neighbor	Decision Tree Classifier
AUC: 59.3%	AUC: 57.7%	AUC: 55.9%

May 2022

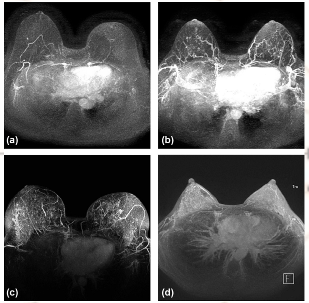
TEXAS
The University of Texas at Austin

Analysis 2: MR Images

Ever Olivares
The University of Texas at Austin

MR Image Analysis- Conv Neural Network

Purpose: Determining a baseline of only image data for training of the overall neural network



Data Input:

- MRI images

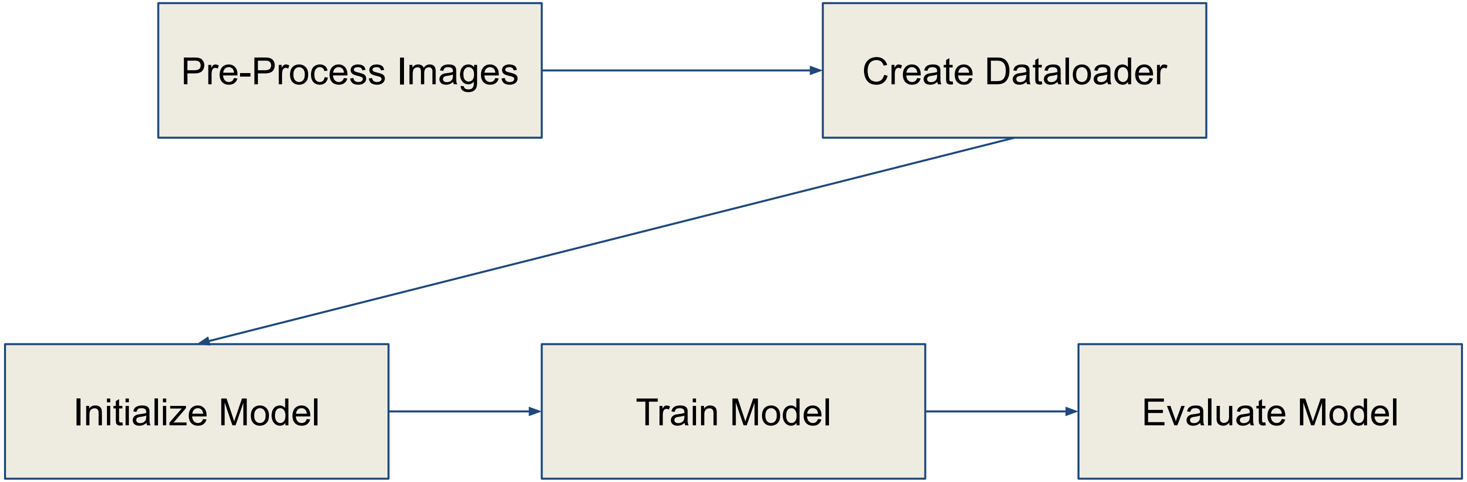
Data Augmentations:

- Random Crop, Random Horizontal Flip, Random Rotation
- **Normalize**

Variable Tracked

- Luminal-Like Molecular Subtype

Code Structure



Pre-Process Images

Create Dataloader

Initialize Model

Train Model

Evaluate Model

Results

Architecture	AUC	Accuracy
Resnet18 Pre-Trained Imagenet Weights	AUC: 55.3%	Accuracy- 60.5%

May 2022

TEXAS
The University of Texas at Austin

Analysis 3: Radiomics Data

Adrian Velasquez, Olivia Parker
The University of Texas at Austin

Radiomics Data Analysis- Logistic Regression Model

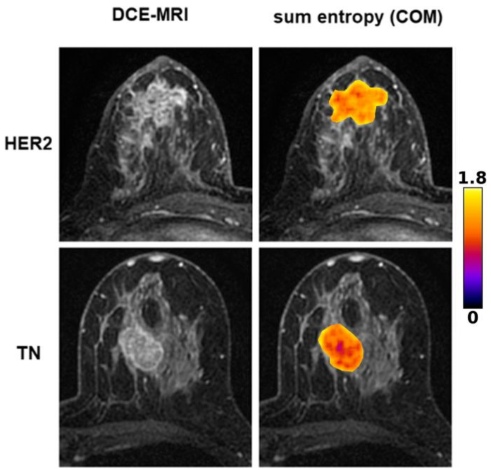
Purpose: Determining baseline accuracy using radiomics data

Data input:

- Radiomics data containing 530 features

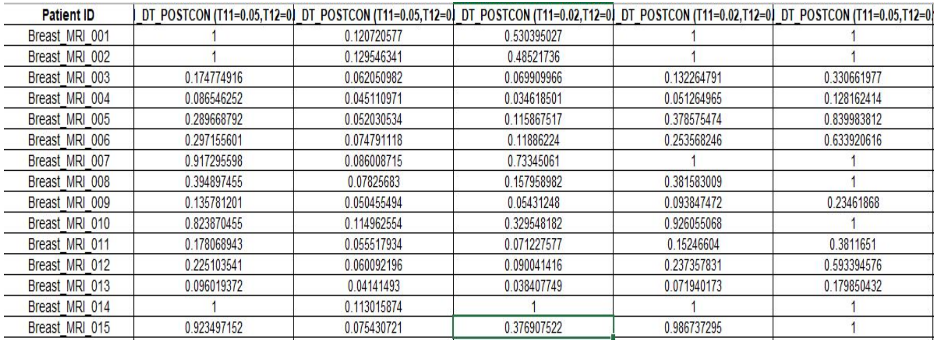
Variable Tracked

- Luminal-Like Molecular Subtype



Data Preprocessing

1. Recursive Feature Elimination
2. Removal of correlated features
3. Removal of features with a p-value higher than 0.05



Patient ID	_DT_POSTCON (T11=0.05,T12=0	_DT_POSTCON (T11=0.05,T12=0	_DT_POSTCON (T11=0.02,T12=0	_DT_POSTCON (T11=0.02,T12=0	_DT_POSTCON (T11=0.05,T12=0
Breast_MRI_001	1	0.120720577	0.530395027	1	1
Breast_MRI_002	1	0.129546341	0.48521736	1	1
Breast_MRI_003	0.174774916	0.062050982	0.069909966	0.132264791	0.330661977
Breast_MRI_004	0.086546252	0.045110971	0.034618501	0.051264965	0.128162414
Breast_MRI_005	0.289668792	0.052030534	0.115867517	0.378575474	0.839983812
Breast_MRI_006	0.297155601	0.074791118	0.11886224	0.253568246	0.633920616
Breast_MRI_007	0.917295598	0.086008715	0.73345061	1	1
Breast_MRI_008	0.394897455	0.07825683	0.157958982	0.381583009	1
Breast_MRI_009	0.135781201	0.050455494	0.05431248	0.093847472	0.23461868
Breast_MRI_010	0.823870455	0.114962554	0.329548182	0.926055068	1
Breast_MRI_011	0.178068943	0.055517934	0.071227577	0.15246604	0.3811651
Breast_MRI_012	0.225103541	0.060092196	0.090041416	0.237357831	0.593394576
Breast_MRI_013	0.096019372	0.04141493	0.038407749	0.071940173	0.179850432
Breast_MRI_014	1	0.113015874	1	1	1
Breast_MRI_015	0.923497152	0.075430721	0.376907522	0.986737295	1

Models Architecture

Sequential Model:

- 3 dense layers
- Dropout technique

Logistic regression:

- Used ‘saga’ solver
- Used cross validation with k-fold and sk-fold

Different Model Results

Sequential Model

Accuracy: 63.95%

Logistic Regression

Accuracy: 64.53%

Cross-Validation

Using k-fold → Accuracy: 68.10%
Using sk-fold → Accuracy: 70.48%

May 2022

The University of Texas at Austin

Analysis 4: Multimodal Neural Network

Greg Holste
The University of Texas at Austin

Neural Network Architecture

Inputs:

- MR Image
- Clinical Data
- Radiomics Data

Output:

Neural Network predicting presence of luminal-like molecular subtype in breast tumors

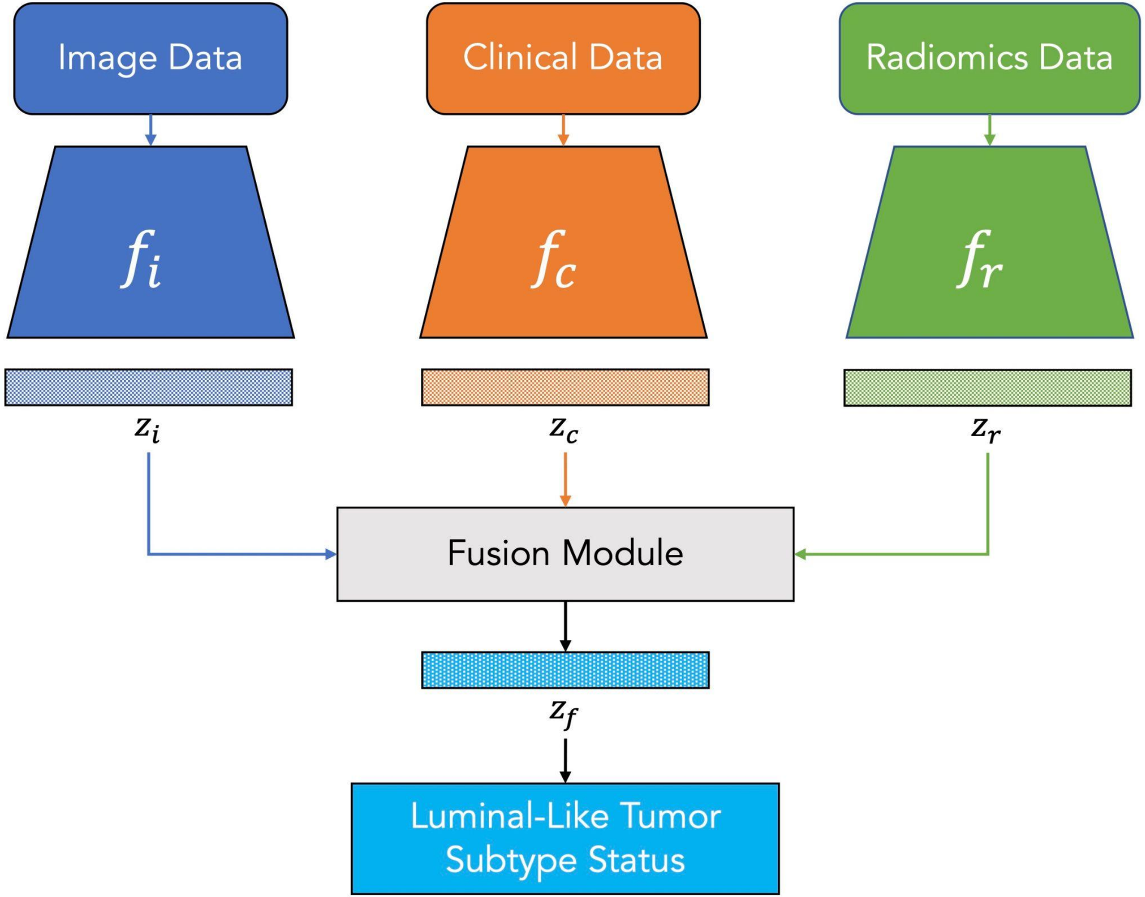


Image Data

Clinical Data

Radiomics Data

f_i

f_c

f_r

z_i

z_c

z_r

Fusion Module

z_f

Luminal-Like Tumor Subtype Status

Results

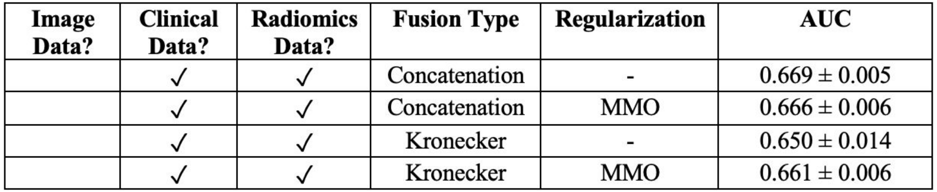


Image Data?	Clinical Data?	Radiomics Data?	Fusion Type	Regularization	AUC
	✓	✓	Concatenation	-	0.669 ± 0.005
	✓	✓	Concatenation	MMO	0.666 ± 0.006
	✓	✓	Kronecker	-	0.650 ± 0.014
	✓	✓	Kronecker	MMO	0.661 ± 0.006

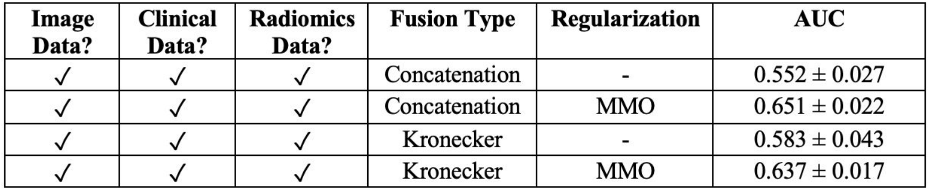


Image Data?	Clinical Data?	Radiomics Data?	Fusion Type	Regularization	AUC
✓	✓	✓	Concatenation	-	0.552 ± 0.027
✓	✓	✓	Concatenation	MMO	0.651 ± 0.022
✓	✓	✓	Kronecker	-	0.583 ± 0.043
✓	✓	✓	Kronecker	MMO	0.637 ± 0.017

Takeaways:

- Multimodal learning improves performance
- Image data hurts performance!
  - Hypothesis: Highly preprocessed images very noisy
  - Solution: incorporate full MRI sequence data
- Concatenation is an adequate fusion method
- MMO regularization only helpful w/ images
  - Hypothesis: image + radiomics info highly correlated
- Difficult task: subtype usually determined by biopsy (not imaging!)