Breast Cancer Diagnostic and Prediction

Ding Tang , Alexander Tsai
Department of Biomedical Engineering
Columbia University

Final Report of APMA E4990: Introduction to Data Science in Industry.

COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science

Introduction

TABLE 1. Estimated New Cancer Cases and Deaths by Sex, United States, 2016*

	ESTIMATED NEW CASES			ESTIMATED DEATHS		
	BOTH SEXES	MALE	FEMALE	BOTH SEXES	MALE	FEMALE
All sites	**1,685,210**	**841,390**	**843,820**	**595,690**	**314,290**	**281,400**
Oral cavity & pharynx	**48,330**	**34,780**	**13,550**	**9,570**	**6,910**	**2,660**
Tongue	16,100	11,700	4,400	2,290	1,570	720
Mouth	12,910	7,600	5,310	2,520	1,630	890
Pharynx	16,420	13,350	3,070	3,080	2,400	680
Other oral cavity	2,900	2,130	770	1,680	1,310	370
Digestive system	**304,930**	**172,530**	**132,400**	**153,030**	**88,700**	**64,330**
Esophagus	16,910	13,460	3,450	15,690	12,720	2,970
Stomach	26,370	16,480	9,890	10,730	6,540	4,190
Small intestine	10,090	5,390	4,700	1,330	710	620
Colon†	95,270	47,710	47,560	49,190	26,020	23,170
Rectum	39,220	23,110	16,110			
Anus, anal canal, & anorectum	8,080	2,920	5,160	1,080	440	640
Liver & intrahepatic bile duct	39,230	28,410	10,820	27,170	18,280	8,890
Gallbladder & other biliary	11,420	5,270	6,150	3,710	1,630	2,080
Pancreas	53,070	27,670	25,400	41,780	21,450	20,330
Other digestive organs	5,270	2,110	3,160	2,350	910	1,440
Respiratory system	**243,820**	**132,620**	**111,200**	**162,510**	**89,320**	**73,190**
Larynx	13,430	10,550	2,880	3,620	2,890	730
Lung & bronchus	224,390	117,920	106,470	158,080	85,920	72,160
Other respiratory organs	6,000	4,150	1,850	810	510	300
Bones & joints	**3,300**	**1,850**	**1,450**	**1,490**	**860**	**630**
Soft tissue (including heart)	**12,310**	**6,980**	**5,330**	**4,990**	**2,680**	**2,310**
Skin (excluding basal & squamous)	**83,510**	**51,650**	**31,860**	**13,650**	**9,330**	**4,320**
Melanoma of the skin	76,380	46,870	29,510	10,130	6,750	3,380
Other nonepithelial skin	7,130	4,780	2,350	3,520	2,580	940
Breast	**249,260**	**2,600**	**246,660**	**40,890**	**440**	**40,450**
Genital system	**297,530**	**191,640**	**105,890**	**57,730**	**26,840**	**30,890**
Uterine cervix	12,990		12,990	4,120		4,120
Uterine corpus	60,050		60,050	10,470		10,470
Ovary	22,280		22,280	14,240		14,240
Vulva	5,950		5,950	1,110		1,110
Vagina & other genital, female	4,620		4,620	950		950
Prostate	180,890	180,890		26,120	26,120	
Testis	8,720	8,720		380	380	
Penis & other genital, male	2,030	2,030		340	340	
Urinary system	**143,190**	**100,920**	**42,270**	**31,540**	**21,600**	**9,940**
Urinary bladder	76,960	58,950	18,010	16,390	11,820	4,570
Kidney & renal pelvis	62,700	39,650	23,050	14,240	9,240	5,000
Ureter & other urinary organs	3,530	2,320	1,210	910	540	370
Eye & orbit	**2,810**	**1,510**	**1,300**	**280**	**150**	**130**
Brain & other nervous system	**23,770**	**13,350**	**10,420**	**16,050**	**9,440**	**6,610**
Endocrine system	**66,730**	**16,200**	**50,530**	**2,940**	**1,400**	**1,540**
Thyroid	64,300	14,950	49,350	1,980	910	1,070
Other endocrine	2,430	1,250	1,180	960	490	470
Lymphoma	**81,080**	**44,960**	**36,120**	**21,270**	**12,160**	**9,110**
Hodgkin lymphoma	8,500	4,790	3,710	1,120	640	480
Non-Hodgkin lymphoma	72,580	40,170	32,410	20,150	11,520	8,630
Myeloma	**30,330**	**17,900**	**12,430**	**12,650**	**6,430**	**6,220**
Leukemia	**60,140**	**34,090**	**26,050**	**24,400**	**14,130**	**10,270**
Acute lymphocytic leukemia	6,590	3,590	3,000	1,430	800	630
Chronic lymphocytic leukemia	18,960	10,830	8,130	4,660	2,880	1,780
Acute myeloid leukemia	19,950	11,130	8,820	10,430	5,950	4,480
Chronic myeloid leukemia	8,220	4,610	3,610	1,070	570	500
Other leukemia‡	6,420	3,930	2,490	6,810	3,930	2,880
Other & unspecified primary sites‡	**34,170**	**17,810**	**16,360**	**42,700**	**23,900**	**18,800**

*Rounded to the nearest 10; cases exclude basal cell and squamous cell skin cancers and in situ carcinoma except urinary bladder.

About 61,000 cases of carcinoma in situ of the female breast and 68,480 cases of melanoma in situ will be diagnosed in 2016.

†Deaths for colon and rectum cancers are combined because a large number of deaths from rectal cancer are misclassified as colon.

‡More deaths than cases may reflect lack of specificity in recording underlying cause of death on death certificates and/or an undercount in the case estimate.

New cases: 249,260 in the USA in 2016.
Deaths: 40,890 in the USA in 2016.

About 61,000 cases of carcinoma in situ of the female breast and 68,480 cases of melanoma in situ will be diagnosed in 2016.

Siegel, Rebecca L., Kimberly D. Miller, and Ahmedin Jemal. *CA: A Cancer Journal for Clinicians* 66.1 (2016): 7-30.

Introduction

Estimated New Cases

	Males			Females		
Prostate	180,890	21%		Breast	246,660	29%
Lung & bronchus	117,920	14%		Lung & bronchus	106,470	13%
Colon & rectum	70,820	8%		Colon & rectum	63,670	8%
Urinary bladder	58,950	7%		Uterine corpus	60,050	7%
Melanoma of the skin	46,870	6%		Thyroid	49,350	6%
Non-Hodgkin lymphoma	40,170	5%		Non-Hodgkin lymphoma	32,410	4%
Kidney & renal pelvis	39,650	5%		Melanoma of the skin	29,510	3%
Oral cavity & pharynx	34,780	4%		Leukemia	26,050	3%
Leukemia	34,090	4%		Pancreas	25,400	3%
Liver & intrahepatic bile duct	28,410	3%		Kidney & renal pelvis	23,050	3%
All Sites	**841,390**	**100%**		**All Sites**	**843,820**	**100%**

Estimated Deaths

	Males			Females		
Lung & bronchus	85,920	27%		Lung & bronchus	72,160	26%
Prostate	26,120	8%		Breast	40,450	14%
Colon & rectum	26,020	8%		Colon & rectum	23,170	8%
Pancreas	21,450	7%		Pancreas	20,330	7%
Liver & intrahepatic bile duct	18,280	6%		Ovary	14,240	5%
Leukemia	14,130	4%		Uterine corpus	10,470	4%
Esophagus	12,720	4%		Leukemia	10,270	4%
Urinary bladder	11,820	4%		Liver & intrahepatic bile duct	8,890	3%
Non-Hodgkin lymphoma	11,520	4%		Non-Hodgkin lymphoma	8,630	3%
Brain & other nervous system	9,440	3%		Brain & other nervous system	6,610	2%
All Sites	**314,290**	**100%**		**All Sites**	**281,400**	**100%**

Ten Leading Cancer Types for the Estimated New Cancer Cases and Deaths by Sex, United States, 2016.

Estimates are rounded to the nearest 10 and cases exclude basal cell and squamous cell skin cancers and in situ carcinoma except urinary bladder.

Siegel, Rebecca L., Kimberly D. Miller, and Ahmedin Jemal. *CA: A Cancer Journal for Clinicians* 66.1 (2016): 7-30.

Introduction

Table 3. Diagnostic Accuracy of Digital and Film Mammography with the Use of the Seven-Point Malignancy Scale after 455 Days of Follow-up.*

Variable	Malignancy Score on Digital Mammography	Malignancy Score on Film Mammography	Difference†	
Women <50 yr old				
Sensitivity	0.49±0.06	0.35±0.06	0.14 (−0.01 to 0.28)	0.06
Specificity	0.97±0.001	0.98±0.001	−0.003 (−0.007 to 0.0003)	0.07
Positive predictive value	0.08±0.01	0.07±0.01		
Premenopausal and perimenopausal women				
Sensitivity	0.47±0.05	0.38±0.05	0.09 (−0.04 to 0.22)	0.20
Specificity	0.97±0.001	0.98±0.001	−0.002 (−0.006 to 0.001)	0.20
Positive predictive value	0.10±0.01	0.09±0.01		
Women with heterogeneously dense or extremely dense breasts				
Sensitivity	0.38±0.04	0.36±0.04	0.03 (−0.07 to 0.12)	0.69
Specificity	0.97±0.001	0.97±0.001	−0.002 (−0.005 to 0.002)	0.33
Positive predictive value	0.10±0.01	0.10±0.01		

Digital Mammography vs. Conventional Film Mammography

Digital mammography, which is available for computational analysis, has shown higher accuracy comparing with conventional film mammography in breast cancer diagnosis. Computation-based diagnosis is the tendency of breast cancer.

Pisano, Etta D., et al. *New England Journal of Medicine* 353.17 (2005): 1773-1783.

Contents

Project 1: Breast Cancer Diagnostic

In this section, our goal is to build a model to diagnose the breast cancer based on the morphology of cell nuclei. The diagnostics are classified into malignant (M) and benign (B), along with the cell nuclei data (features) provided.

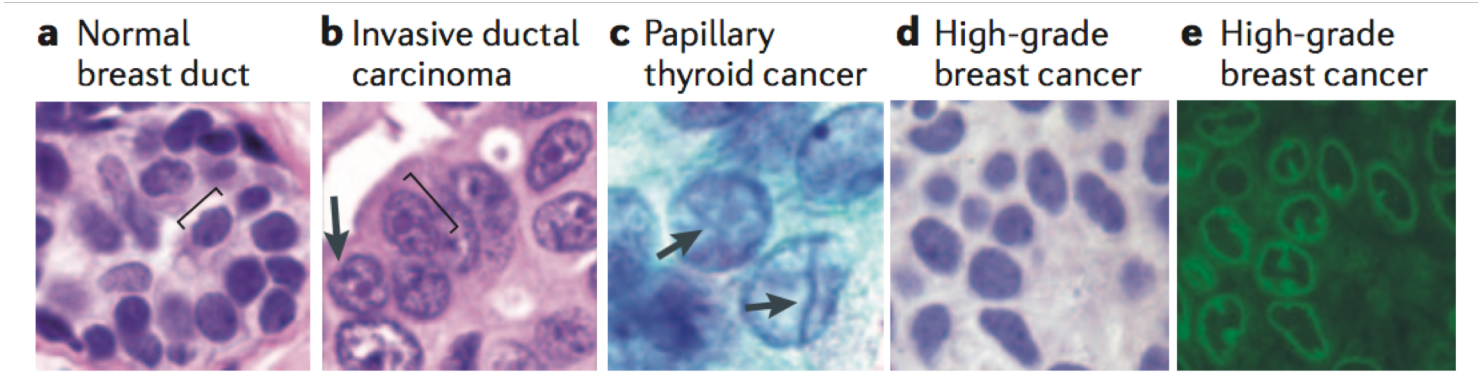
Project 2: Breast Cancer Prediction

In this section, our goal is to build a model to classify the breast cancer type based on the mammography images. The predictions are classified into NORM (Normal), SPIC (Spiculated masses), MISC (Other, ill-defined masses), CIRC (Well-defined/circumscribed masses), CALC (Calcification), ARCH (Architectural distortion) and ASYM (Asymmetry).

Project 1: Breast Cancer Diagnostic

• Building a model to diagnose the breast cancer based on the morphology of cell nuclei. The diagnostics are classified into malignant (M) and benign (B), along with the cell nuclei data (features) provided.

• The dataset is provided by University of Wisconsin & University of California, Irvine:
https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29
https://www.kaggle.com/uciml/breast-cancer-wisconsin-data/data



Project 1: Data Set

diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean
M	13.71	20.83	90.2	577.9	0.1189	0.1645	0.09366
M	13	21.82	87.5	519.8	0.1273	0.1932	0.1859
M	12.46	24.04	83.97	475.9	0.1186	0.2396	0.2273

concave points_mean	symmetry_mean	fractal_dimension_mean	radius_se	texture_se	perimeter_se	area_se	smoothness_se
0.1471	0.2419	0.07871	1.095	0.9053	8.589	153.4	0.006399
0.07017	0.1812	0.05667	0.5435	0.7339	3.398	74.08	0.005225
0.1279	0.2069	0.05999	0.7456	0.7869	4.585	94.03	0.00615

concavity_se	concave points_se	symmetry_se	fractal_dimension_se	radius_worst	texture_worst	perimeter_worst
0.05373	0.01587	0.03003	0.006193	25.38	17.33	184.6
0.0186	0.0134	0.01389	0.003532	24.99	23.41	158.8
0.03832	0.02058	0.0225	0.004571	23.57	25.53	152.5

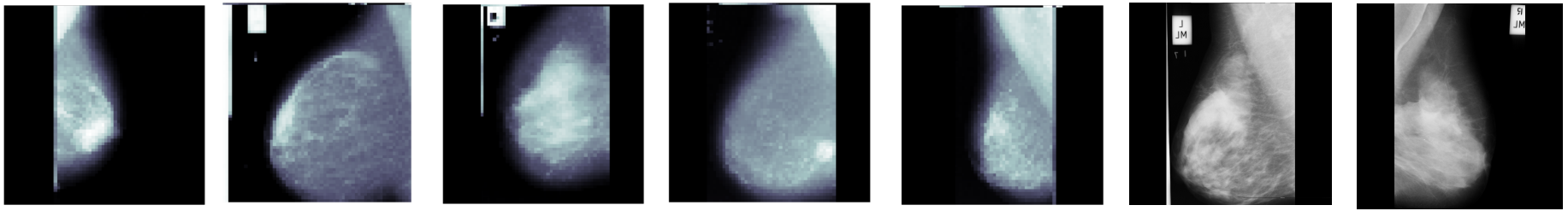
area_worst	smoothness_worst	compactness_worst	concavity_worst	concave points_worst	symmetry_worst	fractal_dimension_worst
2019	0.1622	0.6656	0.7119	0.2654	0.4601	0.1189
1956	0.1238	0.1866	0.2416	0.186	0.275	0.08902
1709	0.1444	0.4245	0.4504	0.243	0.3613	0.08758

Project 2: Breast Cancer Prediction

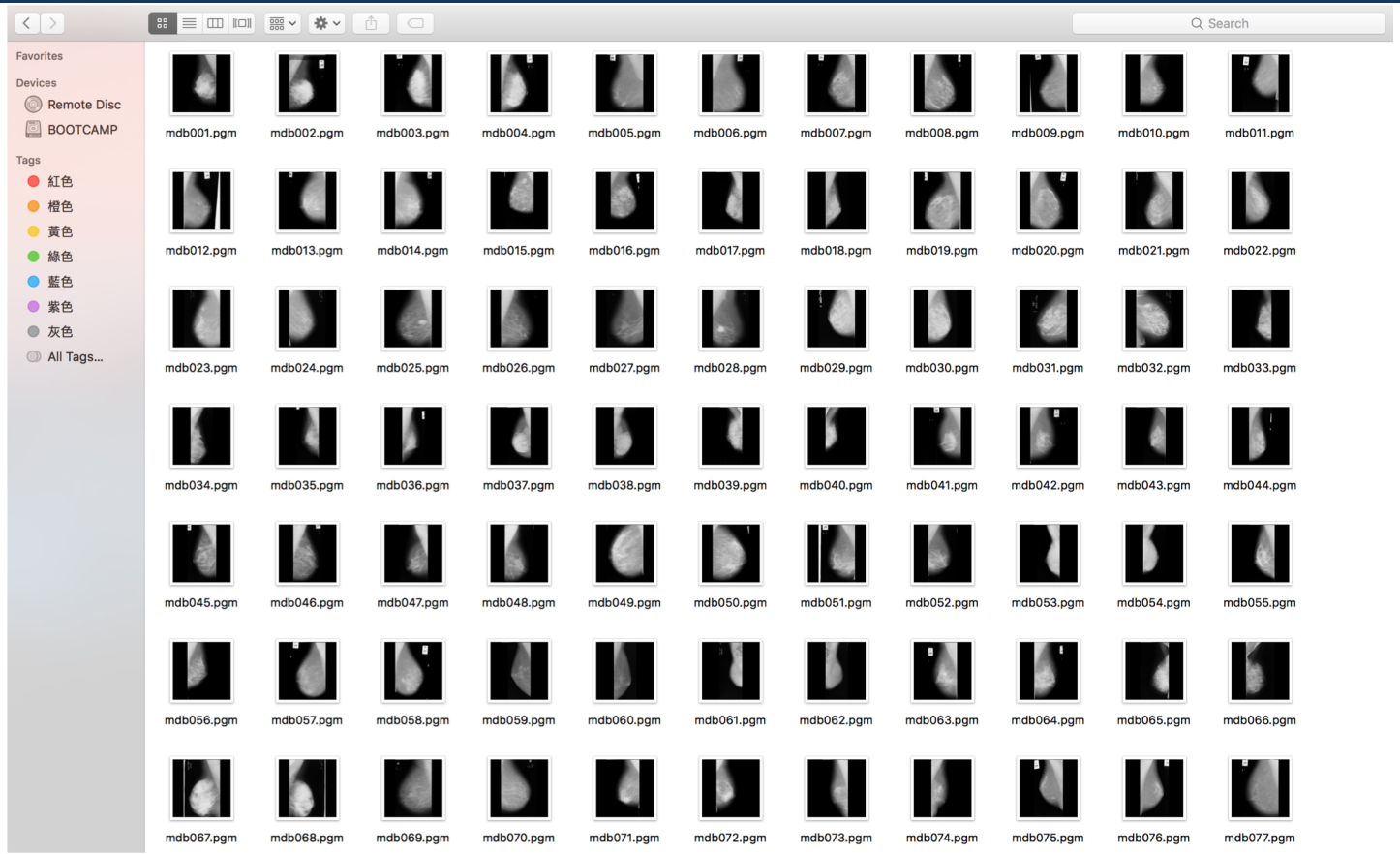
• Building a model to classify the breast cancer type based on the mammography images. The predictions are classified into NORM (Normal), SPIC (Spiculated masses), MISC (Other, ill-defined masses), CIRC (Well-defined/circumscribed masses), CALC (Calcification), ARCH (Architectural distortion) and ASYM (Asymmetry).

• The dataset is provided by Mammographic Image Analysis Society (MIAS):
http://peipa.essex.ac.uk/info/mias.html
https://www.kaggle.com/kmader/mias-mammography/data

NORM SPIC MISC CIRC CALC ARCH ASYM



Project 2: Data Set



Specifically Required URLs

Machine Learning APP:

https://dingandalexander.herokuapp.com

Notebook Presentation:

https://github.com/Columbia-Intro-Data-Science/python-introduction-DingTang/blob/master/Final%20Project/notebook.ipynb

Full Code Base:

https://github.com/Columbia-Intro-Data-Science/python-introduction-DingTang/tree/master/Final%20Project/flask_APP

Completed Project:

https://drive.google.com/file/d/1Z_GvezT9H-pbQzpxXiWi0aH7JWfEprdu/view

Thank you!

Ding Tang , Alexander Tsai
Department of Biomedical Engineering
Columbia University

Final Report of APMA E4990: Introduction to Data Science in Industry.

COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science