BREAST CANCER DIAGNOSTIC CLASSIFICATION

Presented By:
Divisha Makkar (358)
Ananya Aggarwal (438)
Anjuli Jain (548)
Maitreyee Katre (967)
Astha Ray (979)

CONTENTS

01 INTRODUCTION

02 PURPOSE OF STUDY

03 KEY CONCEPTS USED

04 DATA DESCRIPTION – KEY TERMS INCLUDED

05 ANALYSIS

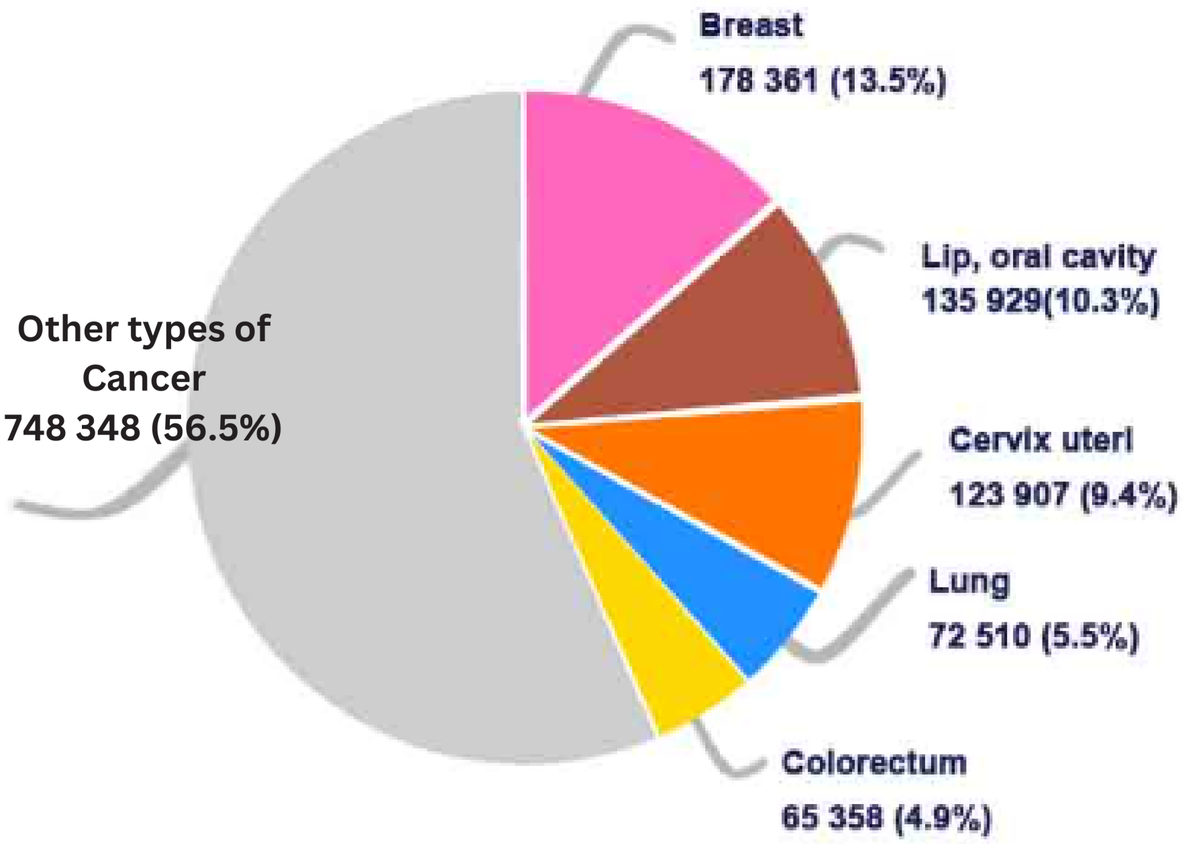
06 CONCLUSION

07 FURTHER SCOPE

ABOUT

The national average of cancer cases for 2022 is 100.4 per 100,000, with a large number of women (105.4 per 100,000) being diagnosed with breast cancer, a preventible disease. By comparison, 95.6 men per 100,000 have been diagnosed with lung cancer.

What is more alarming is that it is being increasingly diagnosed at a younger age (a decade earlier) in India compared to the West. With 90,000 deaths per annum, tragically, a woman loses her life to breast cancer every eight minutes in the country. For every two women diagnosed with breast cancer, one dies of it.



Number of new cases in 2020
(Source: Globacon 2020)

PURPOSE OF STUDY

Breast Self-Examination (BSE) and Clinical Breast Exam (CBE) are the two methods that women can go for. However, these methods are not capable of detecting the cancer at its earliest stage. Mammography is the popular technique designed to image the breast. It comprises of an X-ray system that permits scanty application of X-ray, high contrast, and high resolution detectors.

Mammography has been proven to be most effective in screening and diagnosis. **In this presentation, we have classified malignant (harmful) and benign (not harmful) tumors based on the Wisconsin Breast Cancer Database (WBCD) using Discriminant Analysis.**

DISCRIMINANT ANALYSIS

Discriminant analysis builds a predictive model for group membership. The model is composed of a discriminant function (or, for more than two groups, a set of discriminant functions) based on linear combinations of the predictor variables that provide the best discrimination between the groups. The functions are generated from a sample of cases for which group membership is known; the functions can then be applied to new cases that have measurements for the predictor variables but have unknown group membership.

Dapendant Variable - Categorical
Independent Variable - Metric

Y = A + B1X1 + B2X2 + --- + BnXn
It is called Discriminant function.
Y: Dependent Variable/Grouping Variable
X1, X2, X3,---Xn : Independent Variable/predictor Variable

DATA DESCRIPTION

The data used in this study involving Breast Cancer data extracted from UCI Machine Learning Repository. Dr. WIlliam H. Wolberg (physician) University of Wisconsin Hospitals Madison, Wisconsin, USA has been collected the data since 1989 to 1991. The dataset is available for everyone for research.

There are 699 instances of 11 attributes with 19 missing values. The first attribute is for id number and it is unnecessary for research that's why we removed it from dataset. The number 10 attribute represent class value which have two value 2 and 4 where 2 represents benign and 4 represents malignant. Rest of attributes are ranged from 1 to 10.

Pathologist assigned these numbers based on their characteristics. Large value represents greater chance of malignancy. The detailed information about WDBC is given below in Table 01

Table 01. Attributes information of WDBC

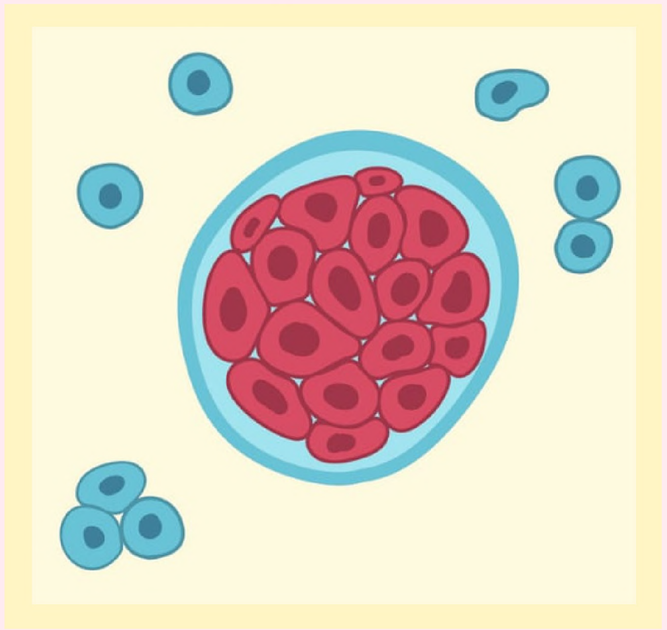
Attributes Name	Value
Clump-thickness	1 – 10
Uniformity of Cell Size	1 – 10
Uniformity of Cell Shape	1 – 10
Marginal Adhesion	1 – 10
Single Epithelial Cell Size	1 – 10
Bare nuclei	1 – 10
Bland chromatin	1 – 10
Normal Nucleoli	1 – 10
Mitoses	1 – 10
Class	2 for benign, 4 for malignant

sample code number	Clump thickness	uniformity of cell size	uniformity of cell shape	marginal adhesion	single epithelial cell size	bare nuclei	bland chromatin	normal nucleoli	mitoses	class
1000025	5	1	1	1	2	1	3	1	1	2
1002945	5	4	4	5	7	10	3	2	1	2
1015425	3	1	1	1	2	2	3	1	1	2
1016277	6	8	8	1	3	4	3	7	1	2
1017023	4	1	1	3	2	1	3	1	1	2
1017122	8	10	10	8	7	10	9	7	1	4
1018099	1	1	1	1	2	10	3	1	1	2
1018561	2	1	2	1	2	1	3	1	1	2
1033078	2	1	1	1	2	1	1	1	5	2
1033078	4	2	1	1	2	1	2	1	1	2
1035283	1	1	1	1	1	1	3	1	1	2
1036172	2	1	1	1	2	1	2	1	1	2
1041801	5	3	3	3	2	3	4	4	1	4
1043999	1	1	1	1	2	3	3	1	1	2
1044572	8	7	5	10	7	9	5	5	4	4
1047630	7	4	6	4	6	1	4	3	1	4
1048672	4	1	1	1	2	1	2	1	1	2
1049815	4	1	1	1	2	1	3	1	1	2
1050670	10	7	7	6	4	10	4	1	2	4
1050718	6	1	1	1	2	1	3	1	1	2
1054590	7	3	2	10	5	10	5	4	4	4

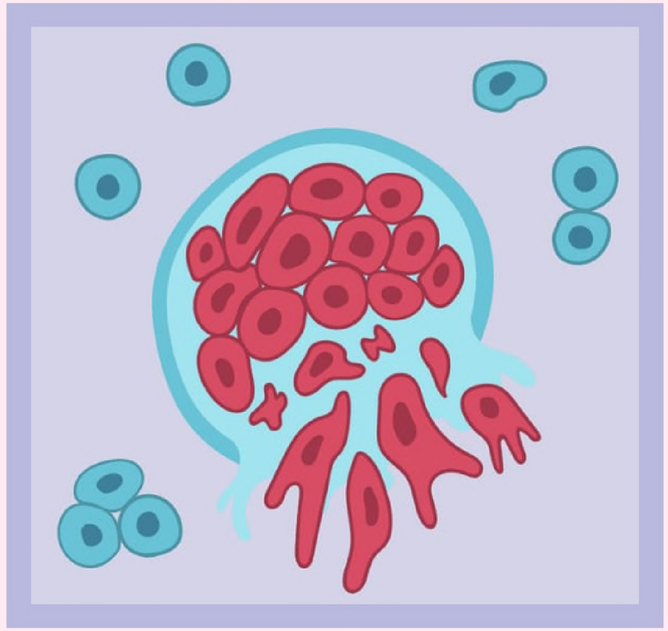
ndicates id number and it is not useful for the experiment

GROUPING VARIABLES

Benign Tumor



Malignant Tumor



Benign Tumor

- Have a slow growth rate.
- Do not invade the issues around them.
- Do not spread to the other parts of the body.
- Easy to cure.
- Most of the cells in Benign tumor are normal.

Malignant Tumor

- Have a fast growth rate.
- Invade the issues around them.
- Spread to the other parts of the body.
- Difficult to cure.
- Cells have abnormal DNA and chromosomes, which make the nucleus larger and darker.

MEDICAL TERMINOLOGIES USED

- Clump Thickness
- Uniformity of Cell size
- Uniformity of Cell Shape
- Marginal Adhesion
- Single Apithelial Cell Size
- Bare Nuclei
- Normal Nucleoli
- Mitoses
- Bland Chromatin

NORMAL CELLS

CANCEROUS CELLS

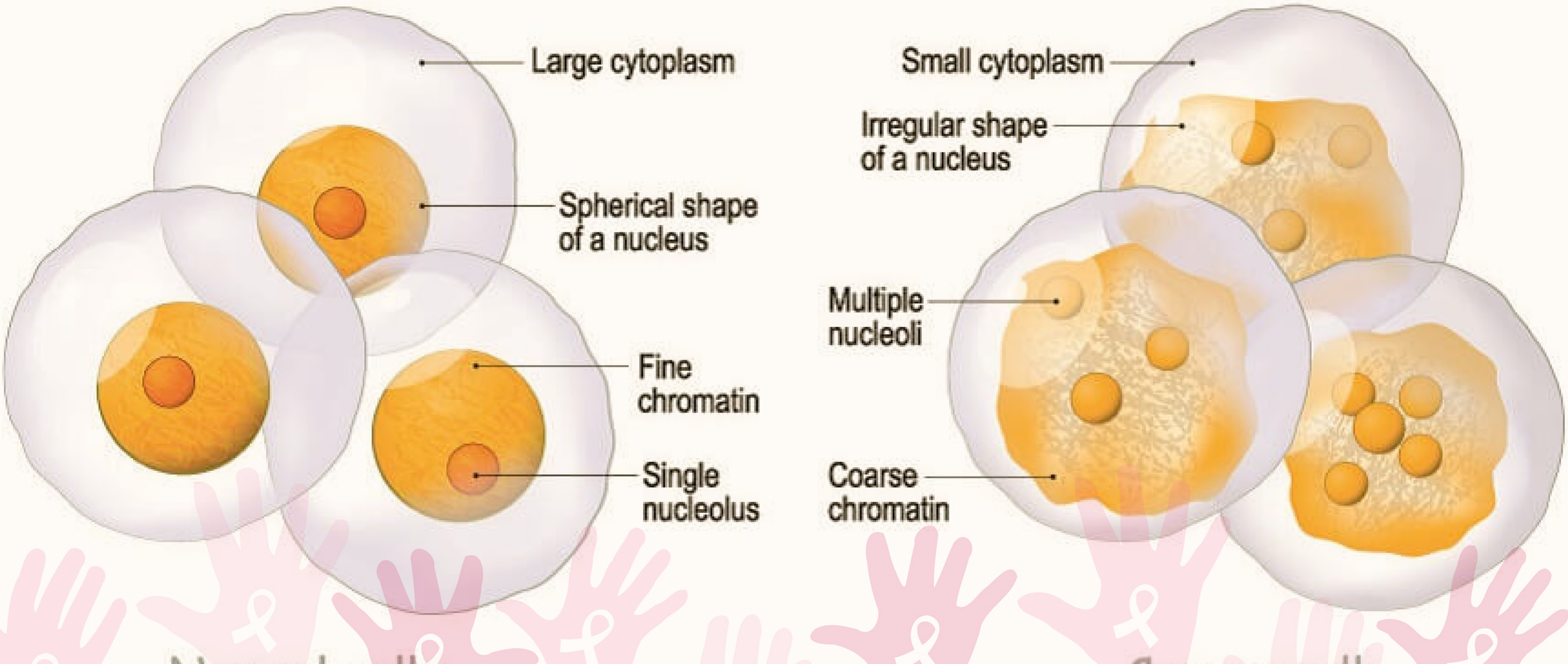
Many cells that ontinue to grow and divide

Variations in size and shapes of cells

Nucleus that is larger and darker than normal

Abnormal number of chromosomes arranged in a disorganized fashion

Cluster of cells without a boundry


Large cytoplasm
Spherical shape of a nucleus
Fine chromatin
Single nucleolus
Small cytoplasm
Irregular shape of a nucleus
Multiple nucleoli
Coarse chromatin
Normal cells
Cancer cells

ASSUMPTIONS

01 Cases should be independent.

02 Predictor variables should have a multivariate normal distribution.

03 Group membership is assumed to be mutually exclusive and collectively exhaustive.

04 The procedure is most effective when group membership is a truly categorical variable; if group membership is based on values of a continuous variable (for example, high IQ versus low IQ), consider using linear regression to take advantage of the richer information that is offered by the continuous variable itself.

ANALYSIS

Whether the grouping variables are distinct or not?

How well the grouping variables discriminate the data?

Relation between grouping variables and predictor variables

Class		Mean	Std. Deviation
Benign	Clump Thickness	2.87	1.671
	Uniformity of Cell Size	1.27	.788
	Uniformity of Cell Shape	1.40	.940
	Marginal Adhesion	1.27	.706
	Single Epithelial Cell Size	2.12	.903
	Bare Nuclei	1.33	1.174
	Bland Chromatin	2.19	1.085
	Normal Nucleoli	1.28	1.011
	Mitoses	1.08	.555
Malignant	Clump Thickness	7.14	2.433
	Uniformity of Cell Size	6.46	2.759
	Uniformity of Cell Shape	6.45	2.546
	Marginal Adhesion	5.53	3.175
	Single Epithelial Cell Size	5.29	2.452
	Bare Nuclei	7.61	3.104
	Bland Chromatin	5.83	2.225
	Normal Nucleoli	5.90	3.326
	Mitoses	2.68	2.586

It gives the descriptives for both the groups on each of the predictor variables.

On comparing mean of each predictor variable of both the groups, we can see that their difference is quite high.

Tests of Equality of Group Means

	Wilks' Lambda	F	df1	df2	Sig.
Clump Thickness	.480	631.776	1	583	.000
Uniformity of Cell Size	.336	1150.281	1	583	.000
Uniformity of Cell Shape	.330	1181.317	1	583	.000
Marginal Adhesion	.486	616.098	1	583	.000
Single Epithelial Cell Size	.537	503.573	1	583	.000
Bare Nuclei	.324	1218.937	1	583	.000
Bland Chromatin	.454	701.928	1	583	.000
Normal Nucleoli	.485	618.542	1	583	.000
Mitoses	.814	132.792	1	583	.000

H0: u1=u2 i.e. There is no significant difference between Benign and Malignant tumor

H1: u1 not equal to u2 i.e. There is a significance difference between Benign and Malignant tumor

Box's M Test of Equality of Covariance Matrices

H0: There is no significant difference between group covariance matrices.
H1: There is a significant difference between group covariance matrices..

Test Results

Box's M		3176.995
F	Approx.	69.338
	Sig.	.000

From Box's Test we get an idea of whether we have equal covariance among our groups or not so by looking at significant values we reject H0 i.e. have unequal group variance, i.e. the covariance matrices for the dependent variable are not equal across groups..

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	5.390[a]	100.0	100.0	.918

a. First 1 canonical discriminant functions were used in the analysis.

- Value of canonical correlation indicates that there is a strong correlation between the predictor variables and the outcome.
- $(0.918)^2 = 0.842$ which implies that 84.2% of the variation in grouping variable is explained by the predictor variables.

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	.157	1072.925	9	.000

Since value of Wilk's Lambda is closer to 0, it reflects better discriminating power of the model. It also gives us the statistical significance of the model.

Bare Nuclei has the most explanatory power and Mitoses has the least explanatory power.

Standardized Canonical Discriminant Function Coefficients

	Function 1
Clump Thickness	.378
Uniformity of Cell Size	.200
Uniformity of Cell Shape	.165
Marginal Adhesion	.141
Single Epithelial Cell Size	.122
Bare Nuclei	.538
Bland Chromatin	.164
Normal Nucleoli	.195
Mitoses	-.009

Structure Matrix

	Function 1
Bare Nuclei	.623
Uniformity of Cell Shape	.613
Uniformity of Cell Size	.605
Bland Chromatin	.473
Clump Thickness	.448
Normal Nucleoli	.444
Marginal Adhesion	.443
Single Epithelial Cell Size	.400
Mitoses	.206

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions

Variables ordered by absolute size of correlation within function.

This is also known as **canonical loading** or **discriminant loading** of discriminant functions. It represents the correlations between the standardized predictor variables and the standardized discriminant function scores. These loadings indicate the extent to which each predictor variable contributes to the discriminant function

Classification Results[a,c]

		Class	Predicted Group Membership: Benign	Predicted Group Membership: Malignant	Total
Original	Count	Benign	364	7	371
		Malignant	15	199	214
	%	Benign	98.1	1.9	100.0
		Malignant	7.0	93.0	100.0
Cross-validated[b]	Count	Benign	364	7	371
		Malignant	16	198	214
	%	Benign	98.1	1.9	100.0
		Malignant	7.5	92.5	100.0

a. 96.2% of original grouped cases correctly classified.

b. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

c. 96.1% of cross-validated grouped cases correctly classified.

This table indicates that 98.1% of the Benign cases and 93% of the Malignant cases of Breast cancer are correctly classified.

Overall, 96.2% of the cases are correctly classified.

DISCRIMINANT FUNCTION-

Y = -3.505 + 0.19 (Clump Thickness) + 0.112 (Uniformity of cell size) + 0.097 (Uniformity of cell shape) + 0.07 (Marginal Adhesion) + 0.074 (Single Epithelial cell size) + 0.256 (Bare Nuclei) + 0.103 (Bland Chromatin) + 0.09 (Normal Nucleoli) - 0.06 (Mitoses)

Canonical Discriminant Function Coefficients

	Function 1
Clump Thickness	.190
Uniformity of Cell Size	.112
Uniformity of Cell Shape	.097
Marginal Adhesion	.070
Single Epithelial Cell Size	.074
Bare Nuclei	.256
Bland Chromatin	.103
Normal Nucleoli	.090
Mitoses	-.006
(Constant)	-3.505

Unstandardized coefficients

Larger the absolute value of unstandardized coefficient, better is the predicting power of the variable.

PREDICTION

Functions at Group Centroids

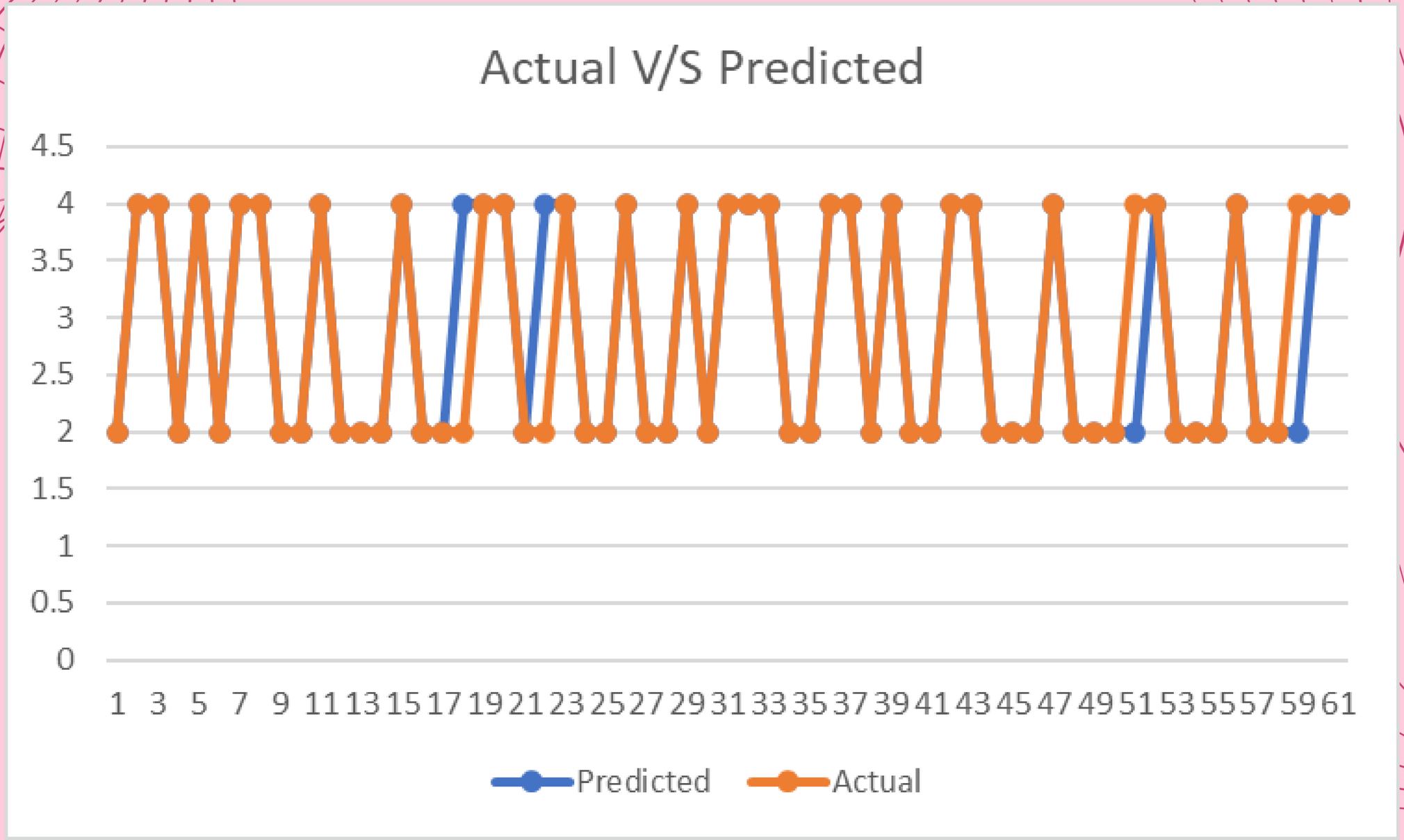
Class	Function 1
Benign	-1.760
Malignant	3.052

Unstandardized canonical discriminant functions evaluated at group means

We use the method of least absolute distance.
We first calculate the discriminant score using the function obtained before and then compute its absolute distance from the centroid of both the classes, i.e., Benign and Malignant.

For Benign,
Distance 1 = | discriminant score - (-1.76) |

For Malignant,
Distance 2 = | discriminant score - (3.052) |


Actual V/S Predicted
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49 51 53 55 57 59 61
Predicted
Actual

RESULTS AND CONCLUSION

- **Group Separation:** Here using test of equality of group of means, we determined that benign and malignant tumor are separable based on the observed characteristics. And since groups are well separated, it indicates that the observed characteristics can be used to predict whether the new patient has benign or malignant tumor.
- Box's M test further strengthens our point by providing that there is no homogeneity between the two grouping variables.
- We get correlation coefficient as 0.918, which suggests a strong correlation between the predictors and the predicted values.
- Value of Wilks' Lambda comes out to be 0.157 and is closer to 0 which reflects better discriminating power of the model. It also gives us the statistical significance of the model.
- Out of all the cases, 98.1% of the Benign cases and 93% of the Malignant cases of Breast cancer are correctly classified.
- **Y = -3.505 + 0.19 (Clump Thickness) + 0.112 (Uniformity of cell size) + 0.097 (Uniformity of cell shape) + 0.07 (Marginal Adhesion) + 0.074 (Single Epithelial cell size) + 0.256 (Bare Nuclei) + 0.103 (Bland Chromatin) + 0.09 (Normal Nucleoli) - 0.06 (Mitoses)**
- From the above equation we can see the relative importance of each predictor variable in predicting whether the tumors benign or malignant

LIMITATIONS

- There is lack of description from the source of data. Thus, we have no or limited idea about the classification of factors on the scale of 1-10 by the pathologist.
- There could be more factors that could possibly impact the classification of breast cancer tumors but our data is limited to these 9 factors.
- Assumption of Normality: Discriminant analysis assumes that the distribution of the predictor variables is normal. If the distribution is not normal or is skewed, it can lead to inconsistent and unreliable results.
- Multicollinearity: Discriminant analysis assumes that the predictor variables are uncorrelated. If there is high multicollinearity between the predictor variables, it may lead to instability of the discriminant function and inaccurate classification.
- Class Imbalance: If the proportion of different classes in the dataset is not balanced, then the classifier may be biased towards the majority class and may not accurately predict the minority class.

FURTHER SCOPE

1. Improved Diagnosis: Discriminant analysis and other machine learning tools can help to identify key characteristics and variables that distinguish different subtypes of breast cancer. This can lead to more accurate and precise diagnosis of breast cancer, allowing doctors to identify the most effective treatment strategies for each patient.

2. Personalized Treatment: Discriminant analysis can also be used to identify specific biomarkers or genetic profiles associated with different types of breast cancer. This information can be used to tailor treatment plans for individual patients, based on their unique cancer subtype and overall health.

3. Prognostic Insights: Discriminant analysis can be used to predict the likelihood of recurrence or metastasis in different subtypes of breast cancer. This information can help doctors and patients make more informed decisions about treatment and long-term care.

4. Research Advancements: Classification of breast cancer using discriminant analysis can lead to advancements in breast cancer research. The identification of subtypes and their characteristics may lead to new therapeutic targets and drugs development.

5. **Detailed Analysis:** Performing a detailed age wise and demography wise analysis to further understand the factors affecting the growth of cancerous cells in a woman's body. This will aid us in designing prevention plans for these specific age groups.

Overall, the scope of breast cancer classification using discriminant analysis is extensive and has the potential to improve the lives of countless patients worldwide.

REFERENCES AND RESOURCES

https://www.livemint.com/news/india/icmr-data-shows-unequal-toll-of-cancer-on-women-11670349329355.html

Breast Cancer Wisconsin Data [online]. Available: http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/breast-cancer-wisconsin.data.

http://cancerindia.org.in/cancer-statistics/

https://www.youtube.com/watch?v=tT1kJhQS2Dk&t=321s&pp=ygUVZGlzY3JpbWluYW50IGFuYWx5c2lz

Fight Cancer!

Early detection saves lives. Hope. Fight. Cure.

THANK YOU