
Let's defeat
breast
cancer
together

IEEE 14th International Symposium on Embedded Multicore/Many-core Systems-on-Chip (MCSoC-2021)
Singapore University of Technology and Design, Singapore
December 20-23, 2021

# THE ROLE OF LINEAR DISCRIMINANT ANALYSIS FOR ACCURATE PREDICTION OF BREAST CANCER

BY

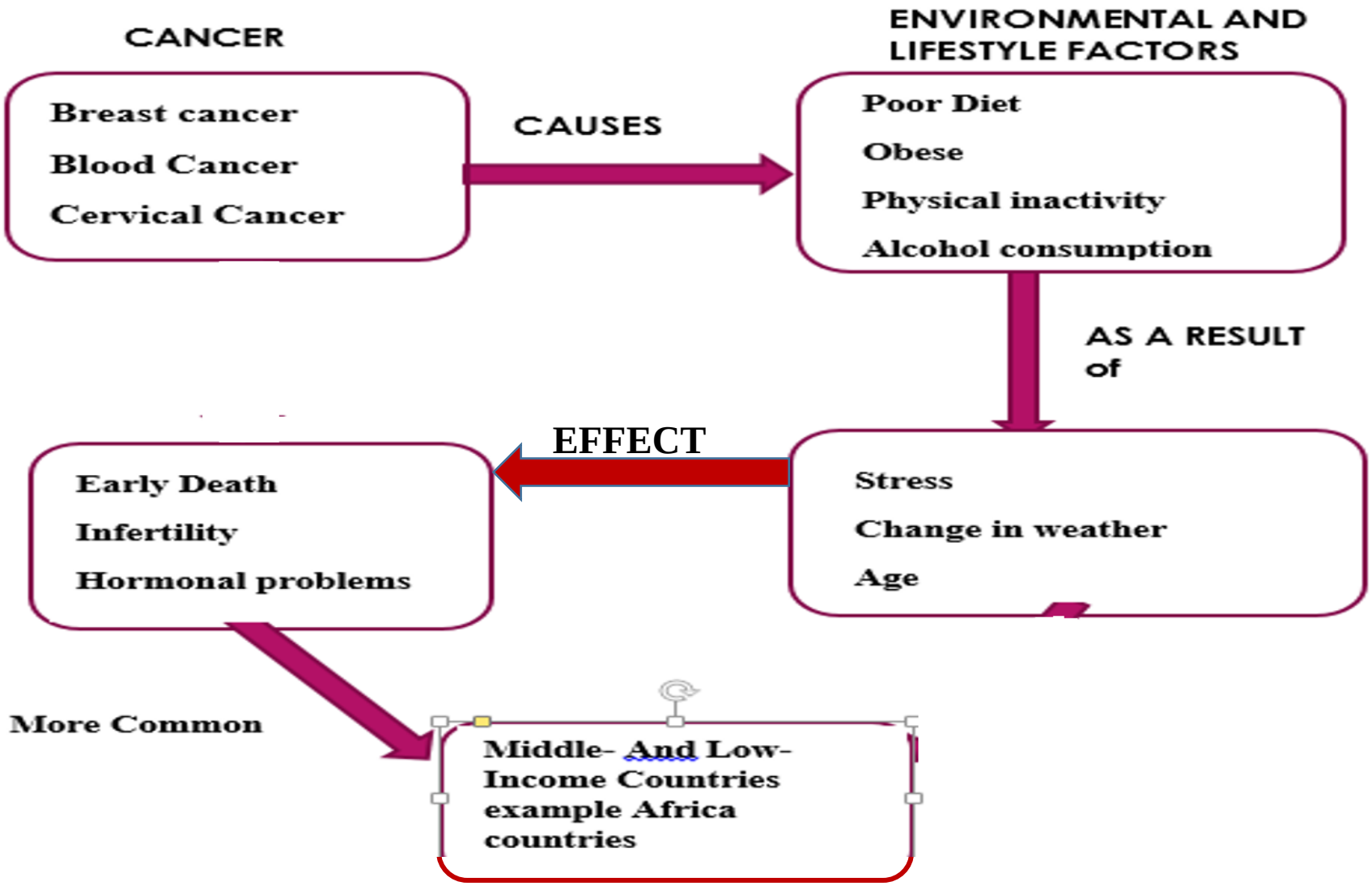
EGWOM ONYINYECHI JESSICA

Bayero University, Kano

# OVERVIEW

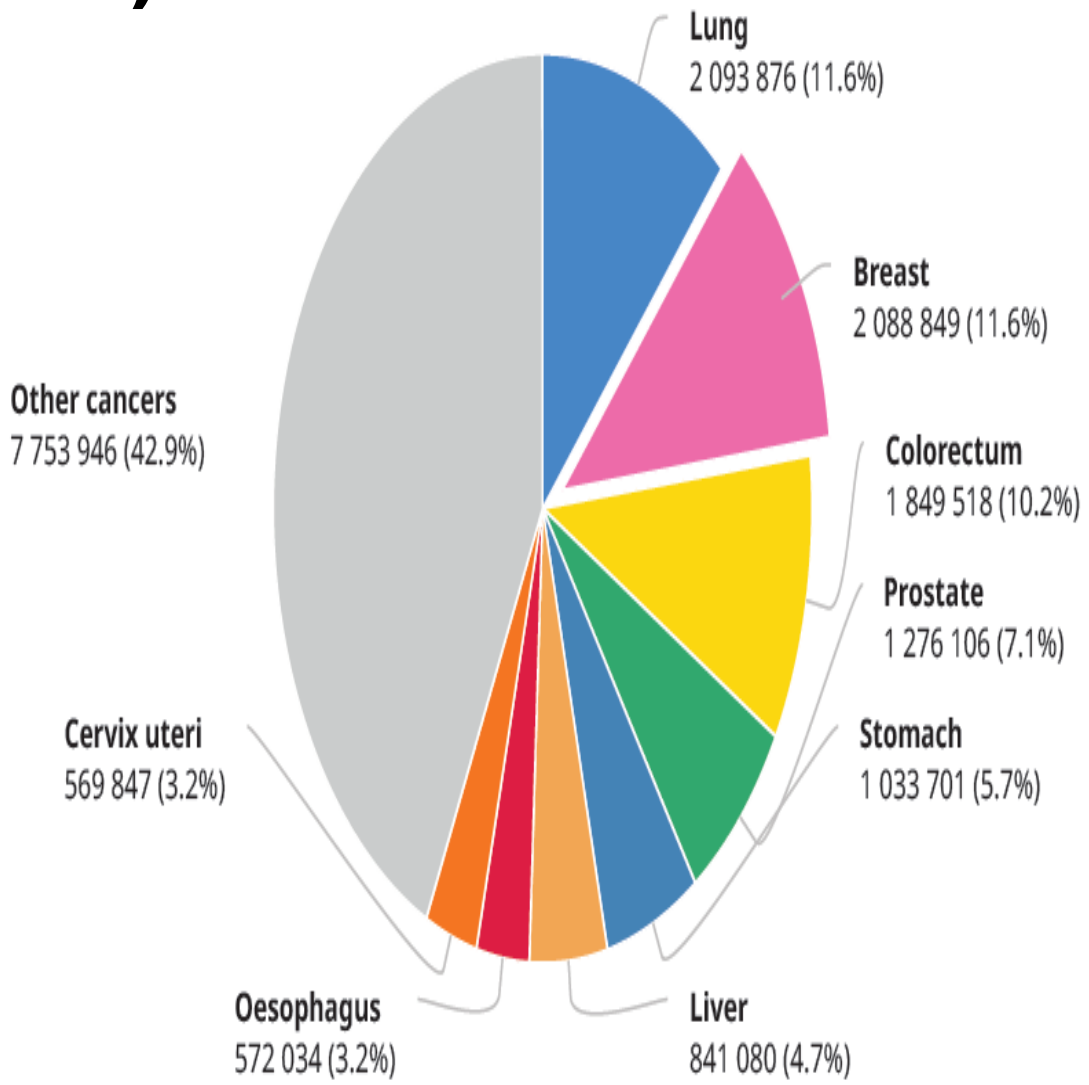
1. Introduction
2. Research problem
3. Aim and objectives
4. Methodology
5. Results
6. Conclusion
7. Recommendations

# INTRODUCTION

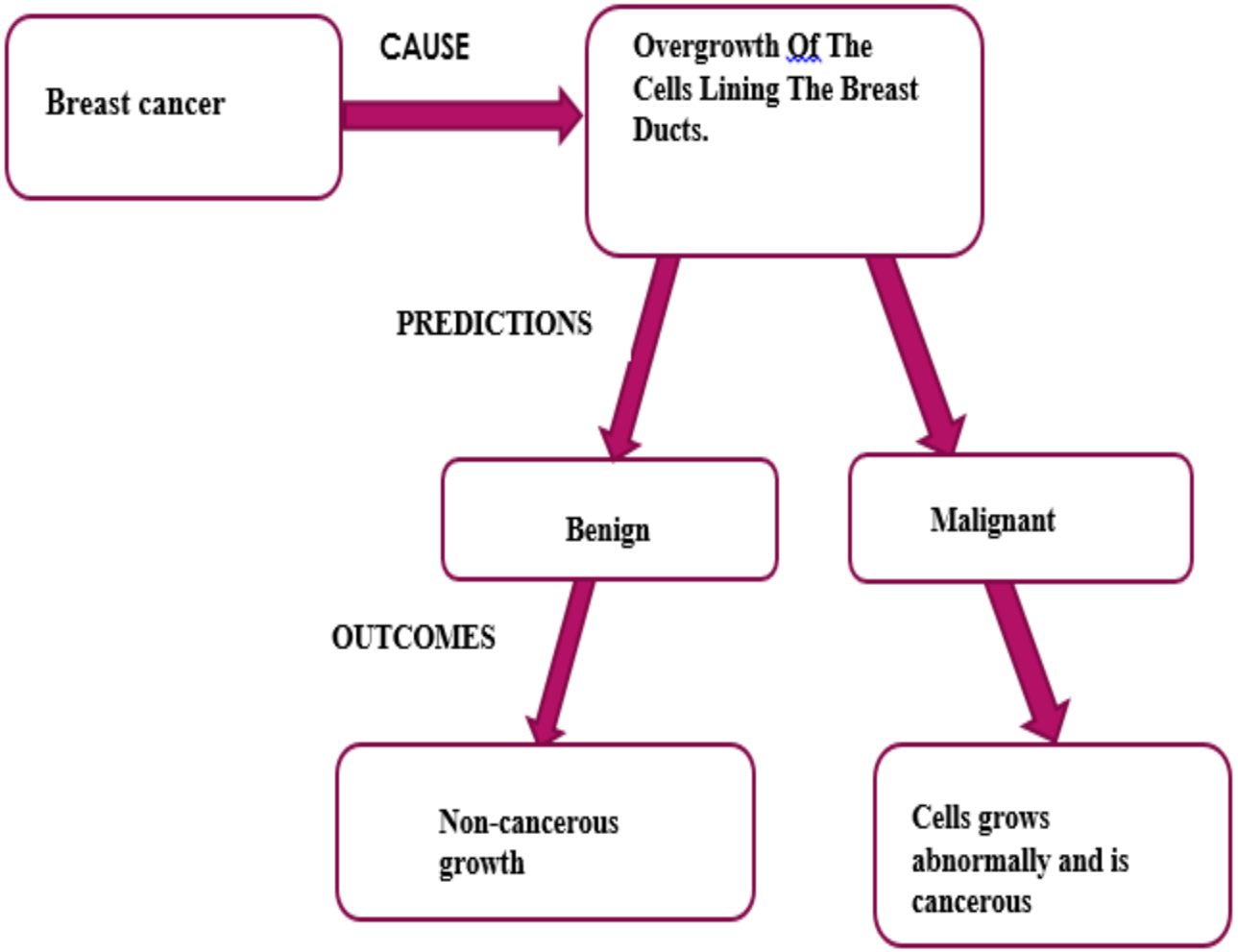


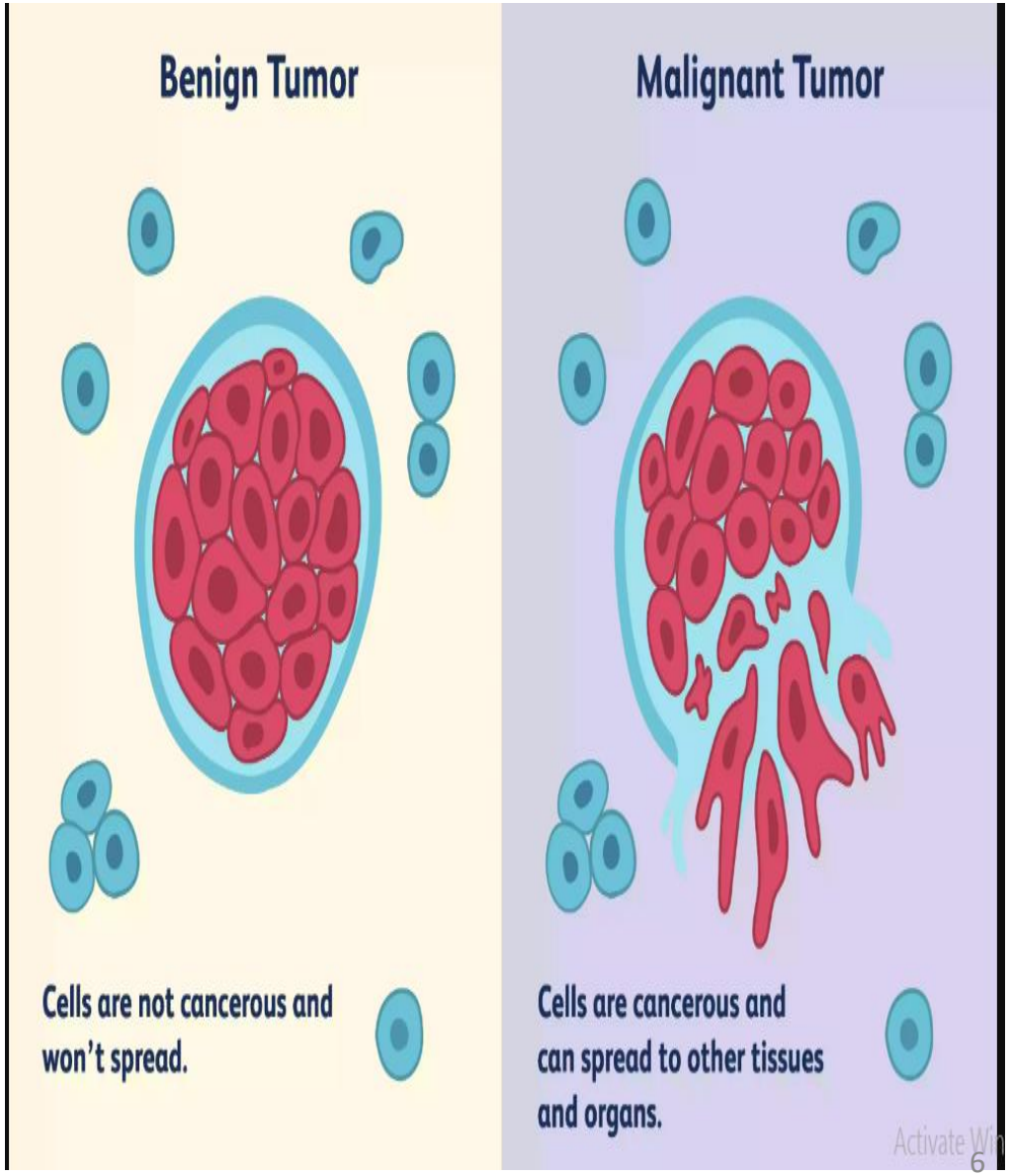
# INTRODUCTION ( cont...)

- **Breast cancer is a disease in which cell in the breast tissues change and divide uncontrollable which results to lump.**
- **A major type of cancers among women in developing countries.**



# INTRODUCTION (cont...)

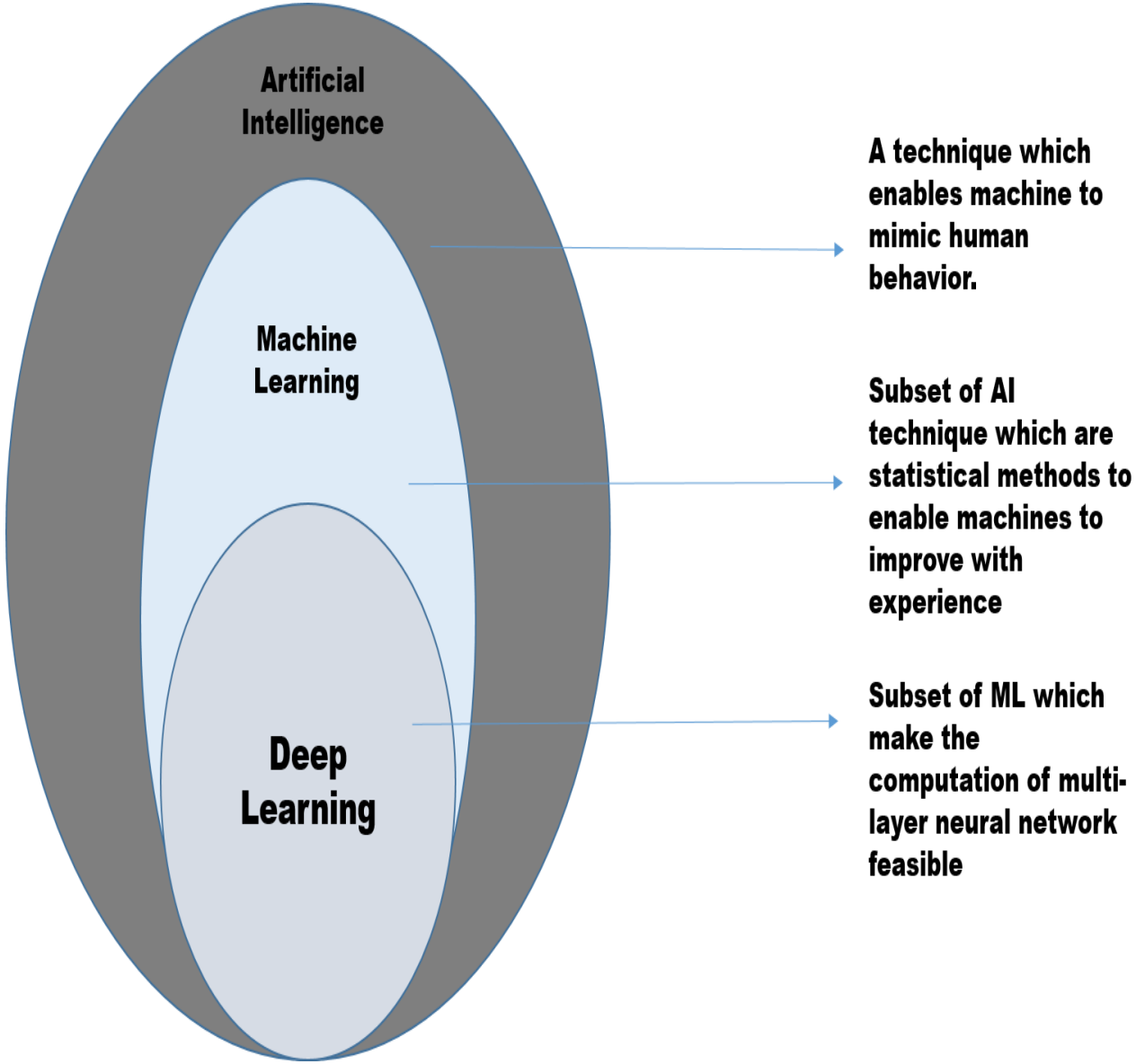



When detected at early stage the survival rate is high

# INTRODUCTION( cont...)

- **Machine learning is a branch of Artificial Intelligence (AI).**
- **It learns and predict new concept from the input data.**
- **learning comes from past experience from input data.**



# INTRODUCTION( cont...)

## FEATURE EXTRACTION

- **Aims to reduce numbers of features in the original dataset.**
- **Newly created features summarize the necessary information in the original set of features.**

# Research Problem

**A good number of research work has shown some limitations such as:**

**Choosing appropriate method to fit the model without considering feature extraction, inability to choose a proper feature extraction techniques to effectively reduce dimensionality for better prediction of the disease and the issues of handling missing values.**

# AIM AND OBJECTIVES

**Aim**

**To develop a  model for breast cancer prediction using Linear Discriminate Analysis (LDA) feature extraction technique and support Vector Machine algorithm .**

# AIM AND OBJECTIVES

**OBJECTIVES**

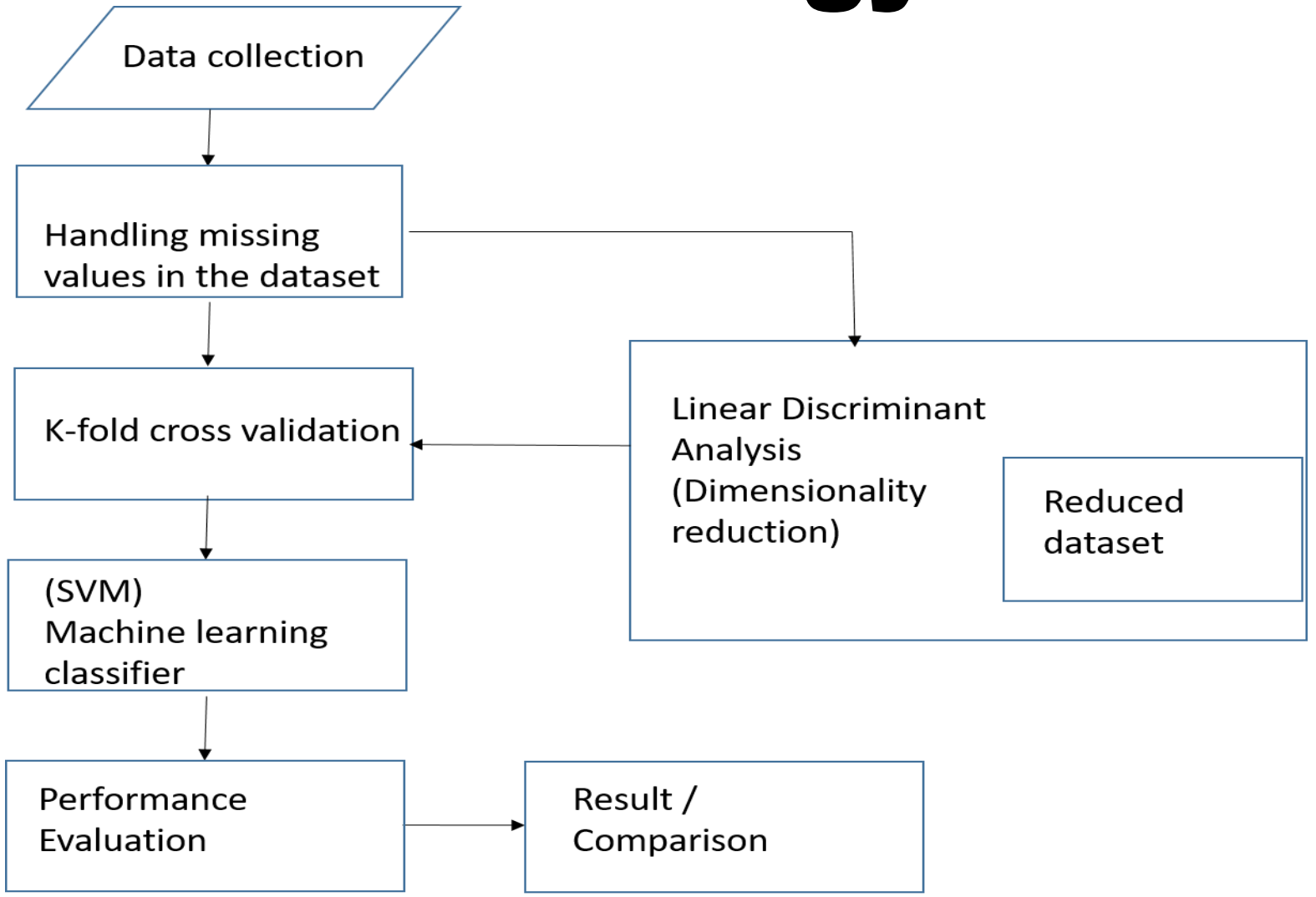
**1. To show the effect of missing values in dataset on model performance.**

**2. Use LDA feature extraction techniques on the existing dataset for dimensionality reduction.**

**3. Design a model for breast cancer prediction that uses SVM classifier on the reduced dataset.**

**4. Use evaluation metrics to evaluate the performance of the proposed model.**

# Methodology




**Prediction Model**

# DATA ACQUISITION

## Dataset was made available from the Wisconsin UCI breast cancer repository having 699 instances and 10 features

| Number | Attribute | Domain |
|---|---|---|
| 1 | Clump Thickness | 1–10 |
| 2 | Uniformity of Cell Size | 1–10 |
| 3 | Uniformity of Cell Shape | 1–10 |
| 4 | Marginal Adhesion | 1–10 |
| 5 | Single Epithelial Cell Size | 1–10 |
| 6 | Bare Nuclei | 1–10 |
| 7 | Bland Chromatin | 1–10 |
| 8 | Normal Nucleoli | 1–10 |
| 9 | Mitoses | 1–10 |
| 10 | Class | 2 for benign, 4 for malignant |

# Computing missing Values

**Missing values was computed using the median of the row. The median is calculated by arranging the number in ascending order to get the in the middle X[n/2]**

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 6 | 1 | 1 | 1 | 1 | 2 | 10 | 3 | 1 | 1 | 2 |
| 7 | 2 | 1 | 2 | 1 | 2 | 1 | 3 | 1 | 1 | 2 |
| 8 | 2 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 5 | 2 |
| 9 | 4 | 2 | 1 | 1 | 2 | 1 | 2 | 1 | 1 | 2 |
| 10 | 1 | 1 | 1 | 1 | 1 | 1 | 3 | 1 | 1 | 2 |
| 11 | 2 | 1 | 1 | 1 | 2 | 1 | 2 | 1 | 1 | 2 |
| 12 | 5 | 3 | 3 | 3 | 2 | 3 | 4 | 4 | 1 | 4 |
| 13 | 1 | 1 | 1 | 1 | 2 | 3 | 3 | 1 | 1 | 2 |
| 14 | 8 | 7 | 5 | 10 | 7 | 9 | 5 | 5 | 4 | 4 |
| 15 | 7 | 4 | 6 | 4 | 6 | 1 | 4 | 3 | 1 | 4 |
| 16 | 4 | 1 | 1 | 1 | 2 | 1 | 2 | 1 | 1 | 2 |
| 17 | 4 | 1 | 1 | 1 | 2 | 1 | 3 | 1 | 1 | 2 |
| 18 | 10 | 7 | 7 | 6 | 4 | 10 | 4 | 1 | 2 | 4 |
| 19 | 6 | 1 | 1 | 1 | 2 | 1 | 3 | 1 | 1 | 2 |
| 20 | 7 | 3 | 2 | 10 | 5 | 10 | 5 | 4 | 4 | 4 |
| 21 | 10 | 5 | 5 | 3 | 6 | 7 | 7 | 10 | 1 | 4 |
| 22 | 3 | 1 | 1 | 1 | 2 | 1 | 2 | 1 | 1 | 2 |
| 23 | 8 | 4 | 5 | 1 | 2 | 4 | 7 | 3 | 1 | 4 |
| 24 | 1 | 1 | 1 | 1 | 2 | 1 | 3 | 1 | 1 | 2 |

# DIMENSIONALITY REDUCTION(LDA)

**LDA technique was used to transform the feature into a lower dimensional space $A=FV_K$.**

**This is done to reduce the number of features in the experimental dataset. The newly created features were able to summarize the necessary information contained initially in the original set of features.**

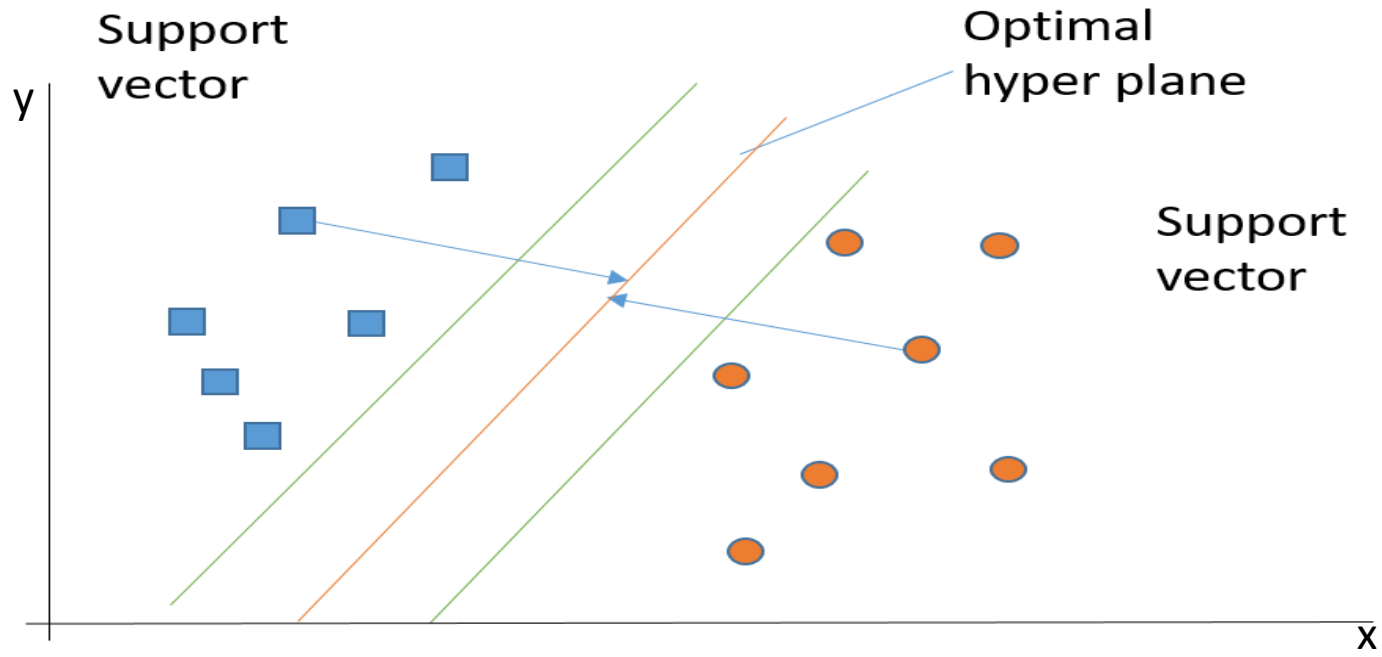
# K FOLD CROSS VALIDATION

**Model was validated by separating one-fifth(k=5) of the data for validation and the rest for training.**

**In each cross-validation process a different one-fifth of the data is selected for validation in such a way that all the data are used for training and also for validation.**

# SVM CLASSIFIER

**Performs binary classification by separating the sets of training vectors for the two different class $(x_1, y_1)$ $(x_2, y_2)$ to find the optimal hyper plane.**



# EVALUATION METRIC

**The effectiveness of the model was evaluated using confusion matrix.**

| | | PREDICTED CLASS | |
|---|---|---|---|
| | | POSITIVE | NEGATIVE |
| ACTUAL CLASS | TRUE | TRUE POSITIVE (TP) | FALSE NEGATIVE (FN) |
| | FALSE | FALSE POSITIVE (FP) | TRUE NEGATIVE (TN) |

# EVALUATION METRIC

- **Precision= shows the ability of the classifier to correctly identify the positive class.**

$$\frac{TP}{TP + FP}$$

- **Recall= The number of times the classifier predicted a negative class out of all the times the class is negative**

$$\frac{TP}{TP+FN}$$

- **Accuracy = Total number of correctly classified patterns divided by total numbers of patterns as shown in equation**

$$\frac{TP+TN}{TP+FP+FN+TN}$$

# Results

**These performance evaluation is done for the models in the following cases.**

- **Case 1: Dropping of rows containing missing values in the breast cancer dataset at the data preprocessing stage.**

- **Case 2: computing of missing values using the median in the breast cancer dataset at the data preprocessing stage.**

# PERFORMANCE EVALUATION USING SVM AFTER DROPPING MISSING VALUES

| Evaluation metrics | Classifier performance |
|---|---|
| Accuracy | 97.1% |
| Precision | 97.0% |
| Recall | 97.0% |

# PERFORMANCE EVALUATION WITH SVM AFTER COMPUTING MISSING VALUES

| Evaluation metrics | Classifier performance |
|---|---|
| Accuracy | 97.8% |
| Precision | 98.0% |
| Recall | 97.0% |

# PERFORMANCE EVALUATION WITH LDASVM AFTER COMPUTING MISSING VALUES WITH MEDIAN

| Evaluation metrics | Classifier performance |
|---|---|
| Accuracy | 99.2% |
| Precision | 98.0% |
| Recall | 99.0% |

# COMPARISON BETWEEN SVM AFTER DROPPING MISSING VALUES AND SVM AFTER COMPUTING MISSING VALUES

| Evaluation metrics | SVM | SVM after computation | Improvement |
|---|---|---|---|
| Accuracy | 97.1% | 97.8% | 97.8 %-97.1%=0.7% |
| Precision | 97.0% | 98.0% | 97.0%-98.0%=1.0% |
| Recall | 97.0% | 97.0% | 97.0%-97.0%=0.0% |

# COMPARISON BETWEEN SVM AND LDASVM AFTER COMPUTING MISSING VALUES

| Evaluation metrics | SVM after computing missing values | LDASVM | IMPROVEMENT |
|---|---|---|---|
| Accuracy | 97.8% | 99.2% | 99.2%-97.8%=1.4% |
| Precision | 98.0% | 98.0% | 98.0%-98.0%=0.0% |
| Recall | 97.0% | 99.0% | 99.0%-97.0%=2.0% |

# Results Comparison

| MODEL | ACCURACY |
| --- | --- |
| LDA-SVM | 99.2% |
| K-SVM by B.Zheng et al [9] | 97.3% |
| DT-SVM by Sivakami & Saraswathi [11] | 91.0% |
| SVM Eldegawy [10] | 98.8% |

# CONCLUSION

**In this research, a novel breast cancer prediction model was proposed.**

**The model was based on SVM and LDA for feature extraction, experiment was conducted using UCI Wisconsin breast cancer dataset.**

**It was found that the model using SVM classifier and LDA feature extraction technique for breast cancer prediction showed a significant improvement in terms of accuracy, Precision and Recall when compared with other approaches in the literature.**

# SIGNIFICANCE OF THE STUDY

**The finding of the current study makes an enticing contribution to the fields of computer science and medicine.**

**In the field of computer science, the study revealed the significance of feature extractions before making predictions. In medicine, it gives some important insight into the process of cancer diagnosis.**

# RECOMMENDATIONS

- what other feature extraction techniques can further improve the prediction accuracy of the model?
- what could be the performance level of the current model when experimented with other breast cancer datasets

# THANKS FOR LISTENING