**Notes**

**Data**

* The dataset used in this project is the **Diagnostic Breast Cancer Wisconsin Database**
* It provides data derived from **digitized images of fine needle aspirates (FNA)** of breast masses. FNA is a medical procedure that uses a thin needle to extract cells from a lump for microscopic examination to diagnose conditions such as breast cancer.
* Each sample is classified as either *malignant* (*M*) or *benign* (*B*) based on the diagnosis.

Variable Description

* The dataset consists of 568 breast mass samples and 32 attributes:

1. *Diagnosis*: The target variable, indicating whether the sample is malignant (M) or benign (B).
2. Feature V ariables: 10 real-valued features are computed for each cell nucleus, describing its geometric and textural characteristics. For each image, the mean, standard error, and largest value of each feature are calculated, resulting in a total of 30 features.
   1. *Radius*: Mean of distances from the center to points on the perimeter.
   2. *Texture*: Standard deviation of gray-scale values.
   3. *Perimeter*: The boundary length of the nucleus.
   4. *Area*: The size of the nucleus.
   5. *Smoothness*: Local variation in radius lengths.
   6. *Compactness*: Calculated as (perimeter² / area) - 1.0.
   7. *Concavity*: Severity of concave portions of the contour.
   8. *Concave Points*: Number of concave portions of the contour.
   9. *Symmetry*: Proportion of the nucleus’ reflective symmetry.
   10. *Fractal Dimension*: Approximates the complexity of the contour as a coastline.

Correlation Matrix

* High correlation between the mean of the variables and their corresponding largest value
* High correlation between the *radius*, *shape*, and *perimeter* variables.

**Method**

Binary Classification

* The machine learning models chosen for our binary classification problem are **logistic regression** and **support vector machines (SVM**).
* To handle non-linear data, we employ an SVM with an **RBF kernel**, which enables non-linear mappings of the input space.
* As a baseline, we begin by performing classification on the unmodified data.

Dimension Reduction

* The dimension reduction techniques used in this report are **principal component analysis (PCA)** and **auto-encoders**.

Autoencoder

* The autoencoder is designed with a **bottleneck dimension of 10**, where the encoder compresses the data into this lower-dimensional representation, and the decoder reconstructs the original data.
* The activation functions used are **ReLU** for the encoder layer and **sigmoid** for the decoder layer.
* The autoencoder is trained using **mean squared error (MSE) loss** on the standardized data.
* After training, we use the encoder to transform the data into its compressed form.
* Finally, we train both logistic regression and SVM on the compressed data and evaluate their performance using classification metrics and accuracy scores.

PCA

* We first apply PCA to the training data and calculate the explained variance ratio for each principal component.
* We then select the number of components that together explain 95% of the variance—specifically, the **first 10 components**.
* Using this reduced representation, we train logistic regression and SVM models, make predictions on the test data, and evaluate their performance using classification metrics and accuracy scores.

**Results**

Baseline

* Logistic regression: 0.97
* SVM: 0.98

With PCA

* Logistic regression: 0.98
* SVM: 0.96

With PCA

* Logistic regression: 0.95
* SVM: 0.95