DABN14

Fall semester 2024

LUNDS UNIVERSITET

Project 2: Advanced Machine learning

PCA & Autoencoders for enhancing logistic regression and SVM

Felix Hult

Evagelos Ifantidis

1. Introduction

Real world data usually suffers from high dimensionality and usually requires some kind of preprocessing before applying any one machine learning algorithm for prediction. However, not all features contribute with the same importance and too many features might make the dataset too complex hence the introduction of different dimensionality reduction techniques. This paper will examine a real world data set regarding classification of different kinds of breast cancer diagnoses with features extracted from digitized images. We deem that this data is of such complexity that dimension reduction techniques might be justified in the preprocessing of the data to improve the predictive power of the models. We aim to evaluate both a linear dimension reduction technique in PCA as well as a non linear technique in Autoencoders and their effectiveness in enhancing the predictive power of logistic regression and SVM in a binary classification setting.

2. Literature Review

Dimensionality reduction is the transformation of features into a lower dimensional space while retaining meaningful representation, mitigating the *curse of dimensionality* (van der Maaten, Postma & Herik, 2007). The traditional way of reducing dimension was through PCA, which does this in a linear fashion. However, these methods have a hard time handling more complex non-linear data which has led to methods that can handle dimension reduction in a non-linear fashion such as the use of Autoencoders (van der Maaten, Postma & Herik, 2007).

A study done by Zhang & Zhao (2017) utilized PCA for predicting fetal well-being using CTG data. Zhang & Zhao (2017) applied PCA for preprocessing of the data and then utilized Adaboost algorithm integrated with SVM. This method helped medical staff make more quick and efficient decision from CTG data, where the proposed model had a accuracy of 93.0% and 98.6%

A paper done by Miotto, Li, Kidd & Dudley (2016) shows how autoencoders could enhance healthcare modeling using data from electronic health records with data from 700 000 patients. They used a three-layer stack of denoising autoencoders which led to an improved prediction accuracy for various diseases like diabetes and cancers compared to the raw electronic health data. Autoencoders in this case enabled better clinical predictions.

3. Method

3.1 Data

The dataset used in this project is the Diagnostic Breast Cancer Wisconsin Database accessed from the UC Irvine Machine Learning Repository (Wolberg et al., 1993a). This dataset 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. An example image of FNA can be seen in Figure 1.

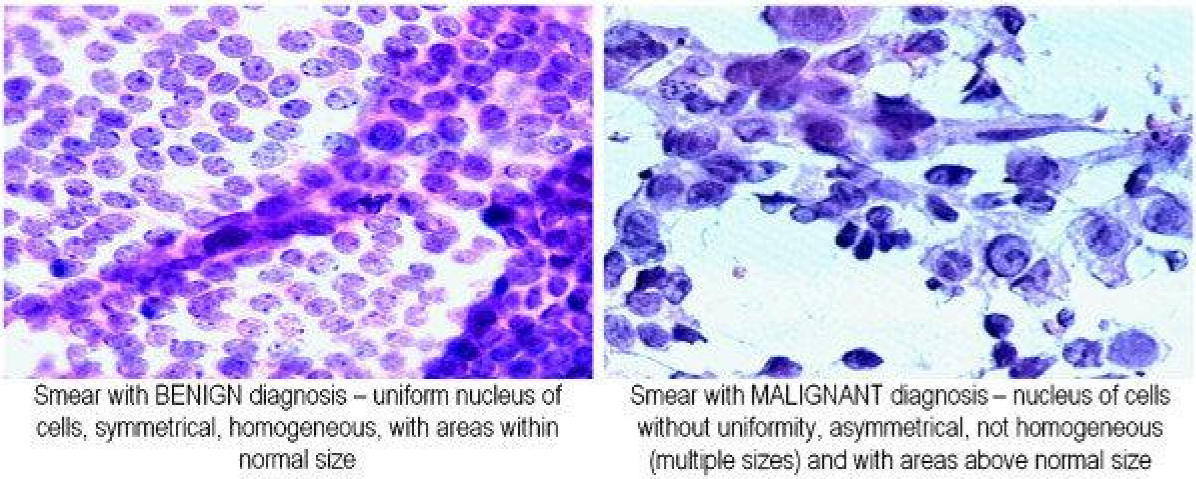
Smear with BENIGN diagnosis – uniform nucleus of cells, symmetrical, homogeneous, with areas within normal size
Smear with MALIGNANT diagnosis – nucleus of cells without uniformity, asymmetrical, not homogeneous (multiple sizes) and with areas above normal size

Figure 1. *Digitized images of fine needle aspirates (Sizilio et al., 2012)*

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

1. *ID Number*: A unique identifier for each sample.
2. *Diagnosis*: The target variable, indicating whether the sample is malignant (M) or benign (B).
3. Feature Variables: 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.
    a. *Radius*: Mean of distances from the center to points on the perimeter.
    b. *Texture*: Standard deviation of gray-scale values.
    c. *Perimeter*: The boundary length of the nucleus.
    d. *Area*: The size of the nucleus.
    e. *Smoothness*: Local variation in radius lengths.
    f. *Compactness*: Calculated as (perimeter^2 / area) - 1.0.
    g. *Concavity*: Severity of concave portions of the contour.
    h. *Concave Points*: Number of concave portions of the contour.
    i. *Symmetry*: Proportion of the nucleus' reflective symmetry.
    j. *Fractal Dimension*: Approximates the complexity of the contour as a coastline.

To check for linear relationships between the feature variables, a correlation matrix was computed, seen in Figure 2. Most notably, there is high correlation between the mean of the variables and their corresponding largest value (labeled as *worst* in the dataset). Also, there is high correlation between the *radius*, *shape*, and *perimeter* variables.

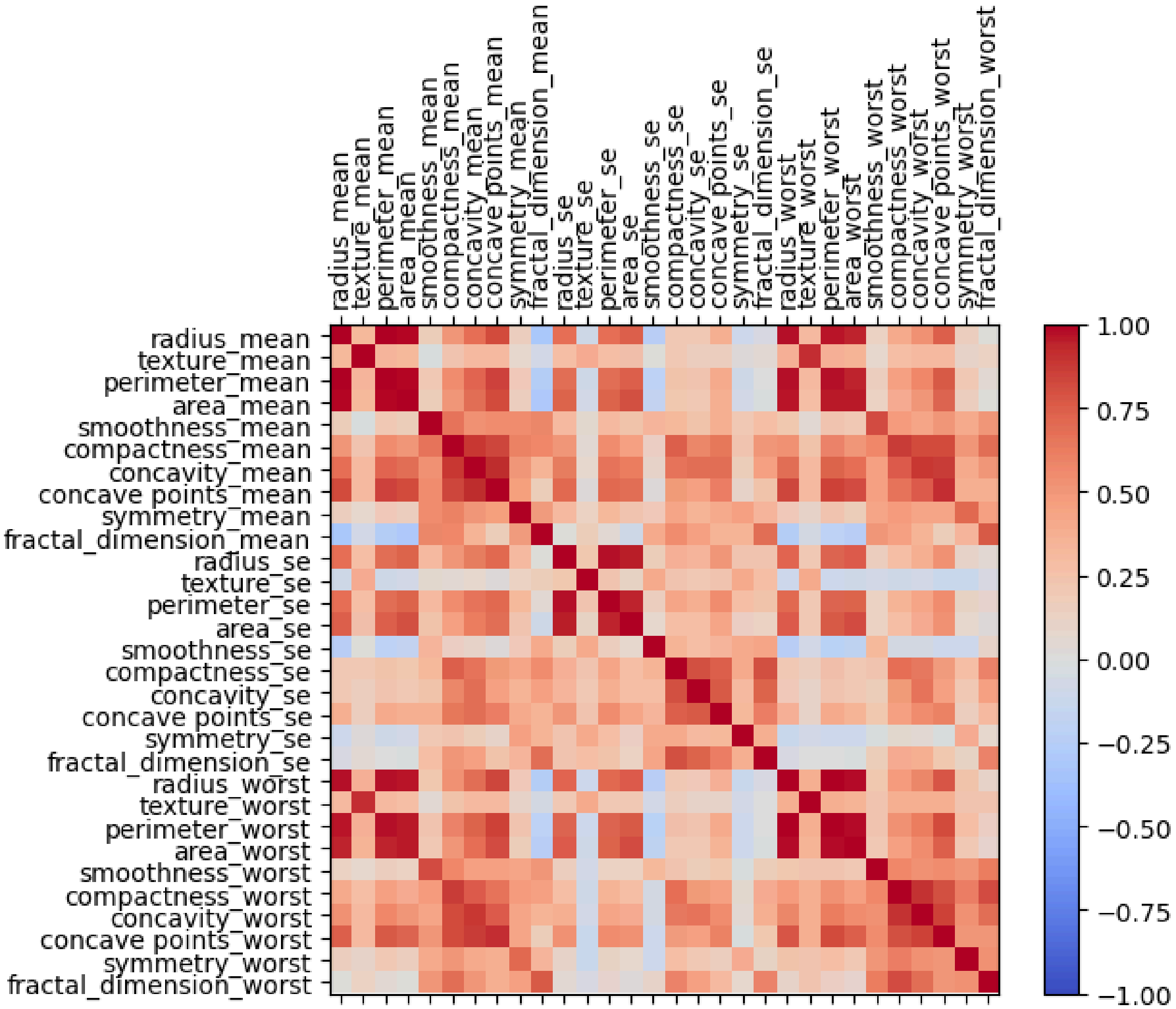

Figure 2. *Heatmap representing the correlation matrix of the feature variables.*

3.2 Method

3.2.1 Binary Classification Models

The machine learning models chosen for our binary classification problem are logistic regression and support vector machines (SVM). Logistic regression is widely used for binary classification, as it models the probability of an outcome by applying a logistic (sigmoid) function to a linear combination of input features, ensuring predicted probabilities range between 0 and 1. SVM, on the other hand, is a versatile supervised learning algorithm that identifies the optimal hyperplane to separate data points of different classes in a high-dimensional space. By using support vectors to define the margin, SVM achieves robust classification. 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.

3.2.2 Dimension Reduction

The dimension reduction techniques used in this report are principal component analysis (PCA) and auto-encoders. Since the dataset consists of 30 feature variables, some of which are highly correlated, these algorithms are expected to significantly reduce the data's complexity. This reduction may lead to improved model performance in our binary classification problem.

An autoencoder is a neural network that compresses data into a lower-dimensional representation, retaining key information. It consists of an encoder that maps input data to a latent space and a decoder that reconstructs the data from this representation. The encoder and decoder are trained to minimize reconstruction error, capturing important features of the data (Lindholm et al., 2022). We begin by standardizing the data to ensure that it has zero mean and unit variance. 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.

Principal Component Analysis (PCA) reduces dimensionality by projecting data onto principal axes, capturing the directions of greatest variance. PCA can be viewed as a linear autoencoder, where both the encoder and decoder are linear, and by retaining the first few principal components, it provides a compact, low-dimensional representation that is effective for data compression and noise reduction (Lindholm et al., 2022). In our approach, 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.

4. Results

4.1 Logistic regression & SVM without dimension reduction

Table 1: Model evaluation without dimension reduction

	Accuracy
Log-Regression	0.9736
SVM	0.9826

4.2 Logistic regression & SVM with PCA pre-processing

Table 2: Model evaluation with PCA

	Accuracy
Log-Regression	0.9825
SVM	0.9649

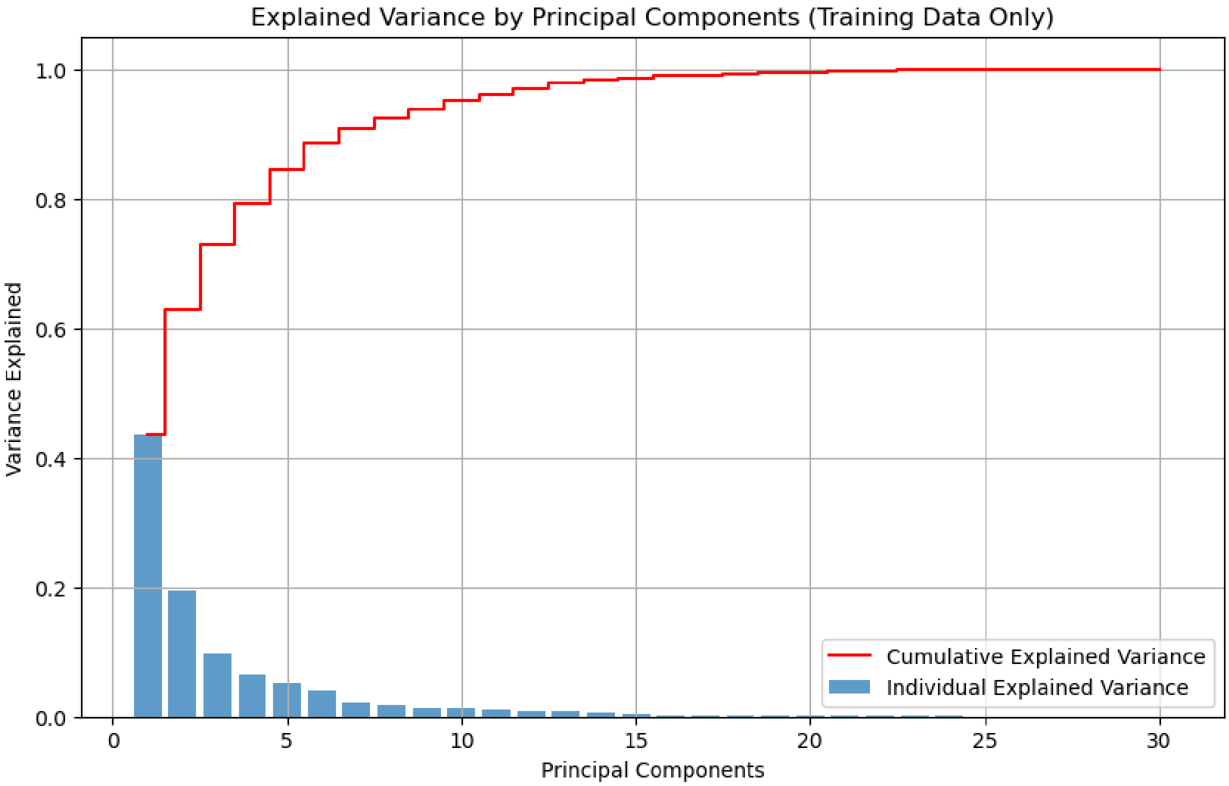

Figure 3. *Explained variance plot*

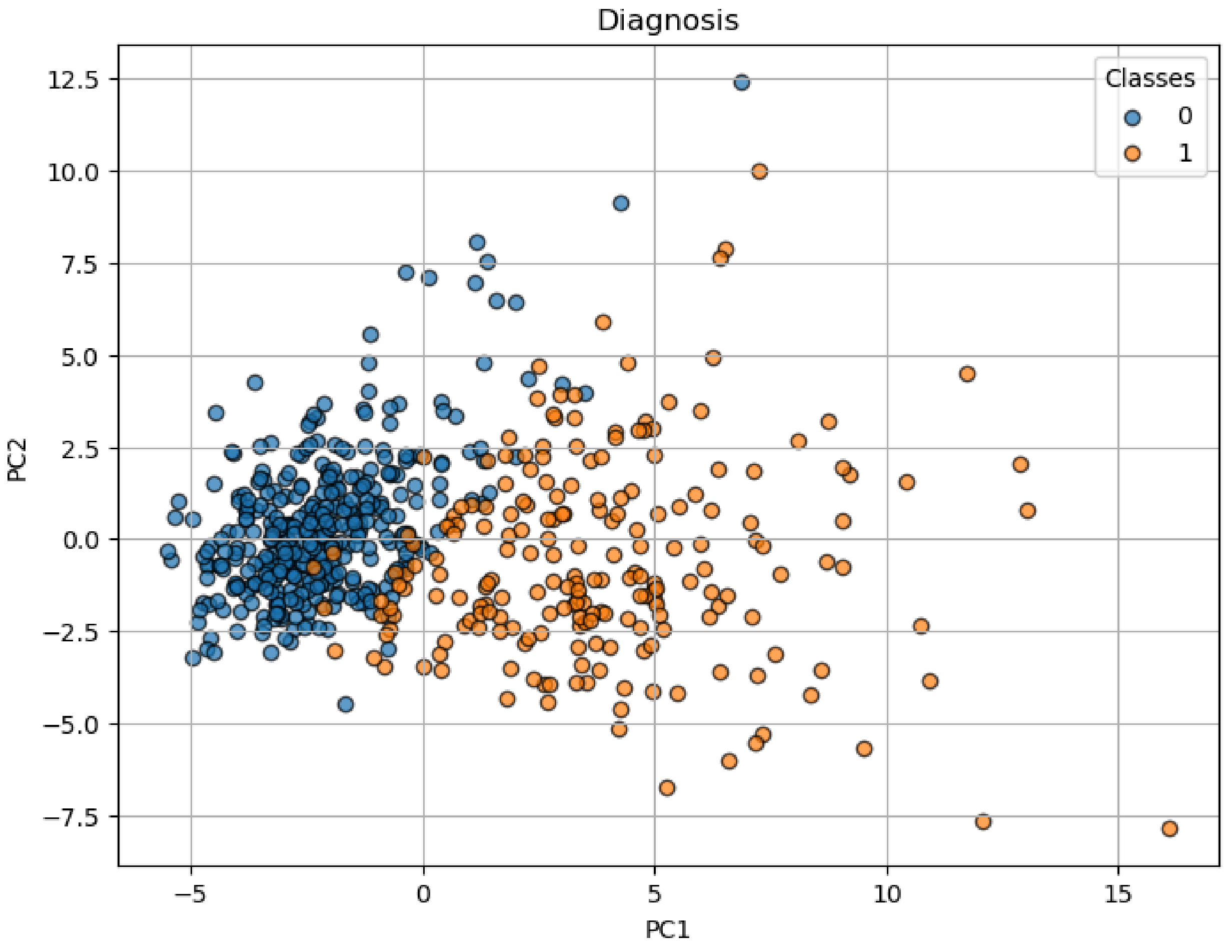

Figure 4. *Scatterplot of classes by PC1 & PC2*

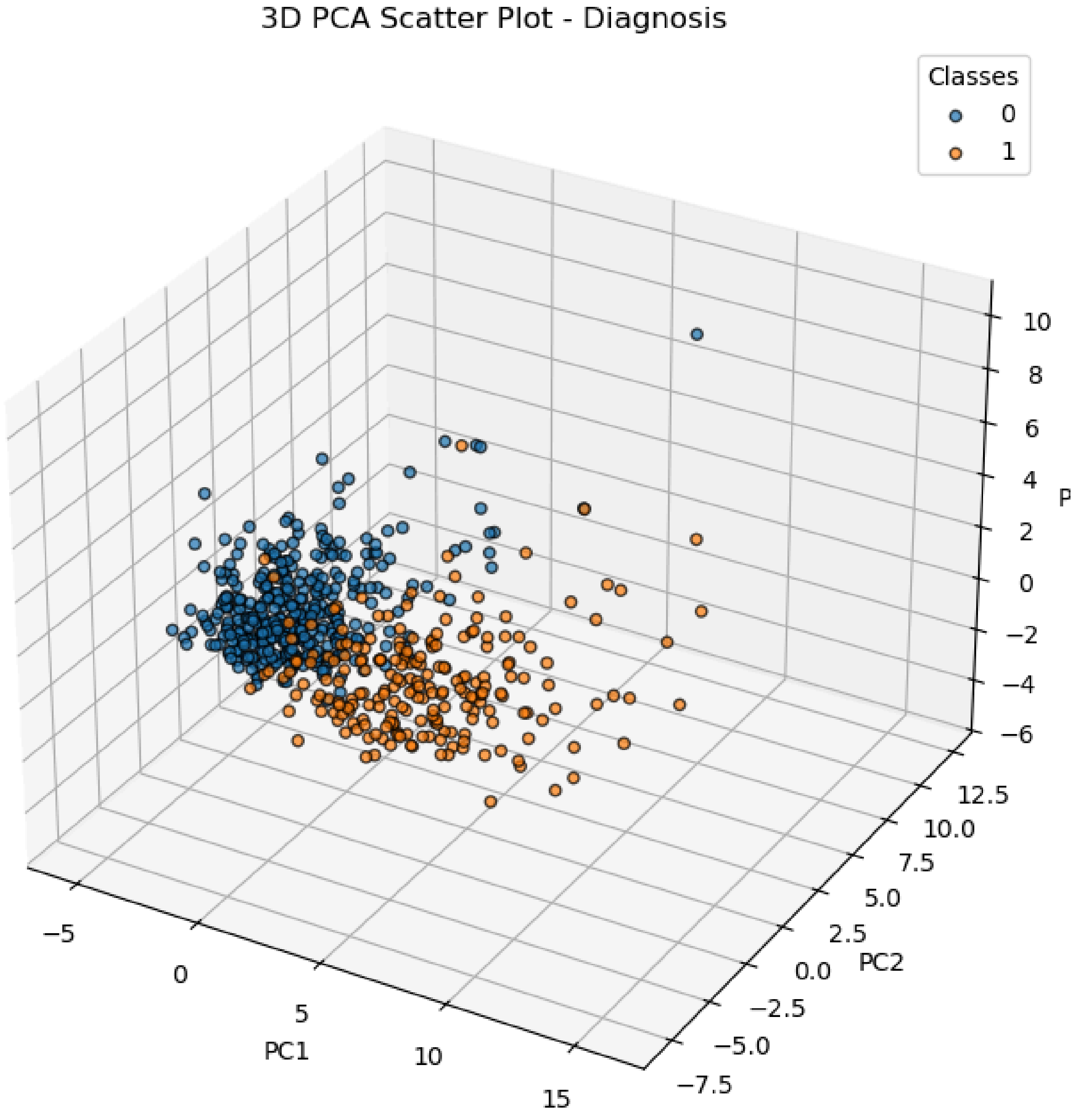

Figure 5*. Scatterplot of classes by PC1,PC2 & PC3*

4.3 Logistic regression & SVM with Autoencoder pre-processing

	Accuracy

Log-Regression	0.9474
SVM	0.9474

5. Conclusion & Discussion

The results show a mixed effectiveness in enhancing the predictive power of SVM and logistic regression. Already without dimension reduction done to the data, the models already have a high accuracy with SVM with a RBF kernel being slightly better. This might be because there are some kinds of non-linear relationships in the data that the logistic regression model can't fully take advantage of. If this is the case then it might explain why the logistic regression model, after having PCA, done before, has a higher accuracy than SVM, which actually sees a decrease in its effectiveness. The reduction of dimension might remove some non-linear relationships in the data making a logistic regression model more generalizable to new data while for SVM, which maps the data to a higher dimension through the RBF-kernel, loses information that the features carried.

Finally, using autoencoders led to the worst performance of both models. There isn't a clear cut answer to why this is the case but some common reasons could be that the autoencoders compressed the data so that important information was lost. They are generally used for its ability to make non-linear dimension transformations, if the data did not contain any relevant non-linear relationships then the autoencoders rather captured noise in the data than actual information. Unlike PCA, which explicitly maximizes variance in its components, autoencoders minimize reconstruction error, which does not guarantee the preservation of features relevant to classification. There are also steps that can be taken to make autoencoders more effective through the introduction of regularization and choice of hyperparameters, although this paper didn’t employ these techniques.

In conclusion, the already high baseline accuracy of both models raises the question whether dimension reduction techniques are necessary for this particular data set. Even though we consider this dataset to be of high complexity it could be the case that the features already were very informative and not too sparse and that for these techniques to work the dataset would instead have had many more features making dimension reduction more justifiable.

6. References

James, G., Witten, D., Hastie, T., Tibshirani, R. & Taylor, J. (2023). An Introduction to Statistical Learning: With Applications in Python. Cham: Springer.

Lindholm, A., Wahlström, N., Lindsten, F. & Schön, T.B. (2022). Machine Learning - A First Course for Engineers and Scientists [e-book]. Cambridge: Cambridge University Press, https://smlbook.org/book/sml-book-draft-latest.pdf.

Miotto, R., Li, L., Kidd, B. A. & Dudley, J. T. (2016). Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records, *Scientific Reports,* vol. 6, nr. 1**,** s 26094

Sizilio, G., Leite, C., Guerreiro, A. & Neto, A.D. (2012). Fuzzy method for pre-diagnosis of breast cancer from the Fine Needle Aspirate analysis. *Biomedical Engineering Online*, vol. 11, no. 83, https://doi.org/10.1186/1475-925X-11-83 van der Maaten, L., Postma, E. & Herik, H. (2007). Dimensionality Reduction: A Comparative Review, *Journal of Machine Learning Research - JMLR,* vol. 10, nr.

Wolberg, W., Mangasarian, O. & Street, N.. (1993). Breast Cancer Wisconsin (Diagnostic). Available at: https://doi.org/10.24432/C5DW2B (Accessed 10 January 2025).

Zhang, Y. & Zhao, Z. (2017) Published. Fetal State Assessment Based on Cardiotocography Parameters Using Pca and Adaboost. 2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), 14-16 Oct. 2017 2017. 1-6.