PCA and K-means clustering on the Breast Cancer Wisconsin data set

Ross Erskine,[1,*] Ho Him Lee,[1,†] and Prakhar Prakarsh[1,‡]

[1] *School of Physics and Astronomy, University of Nottingham, Nottingham, NG7 2RD, UK*

(Dated: May 4, 2023)

Breast cancer is a common disease for middle-aged women. For this project, we focus on discovering the possibilities to perform features engineering on a breast cancer dataset collected by [1] at the University of Wisconsin-Madison Hospital. This dataset contains 699 instances with 9 features and two classes, benign and malignant. Two features engineering method are used in this project, the first one is Principal component analysis which aims to reduce the dimension of the features. The second method we used is K-means Clustering to simulate the two classes in the dataset.

I. INTRODUCTION

Breast cancer can occur due to abnormal growth of cells in the breast. The Breast is composed of different blood vessels with connective tissues, lymph vessels, lobules, and milk ducts. Tumors can form when tissues grow abnormally and create cell divisions. Tumors can be categorised in to two types: benign tumors are classed as non-cancerous which are produced from minor structural changes in the breast [2]. While Malignant tumors are classed as cancerous and can be invasive; which spread to surrounding organs or non-invasive; that remained confined. Classifying these tumours using *Machine learning* (ML) models can help control treatment pain for patients, decrease mortality risks and increase survival rates.

In this study the data set used, is the Wisconsin breast cancer data from [3] which was collected by [1] at the University of Wisconsin-Madison Hospital. The data has 683 entries (after pre-processing) of assessed nuclear features of fine needle point aspirates taken from the breast of patients. Each entry has 9 attributes: *Clump thickness, Uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, Mitoses*, and a response variable that is either *malignant* 239 cases, and *benign* 444 cases. In studies on [3] breast cancer data-set in recent years, the data-set has been a very popular data-set from which to experiment ML models on, for instance some *Supervised learning classification* methods such as : *Artificial neural networks* (ANN) [4–9], *Support-vector machines* (SVM) [5, 10–13], *Decision tree's* (DT) [6, 11, 14–17], *Naive bayes* (NB) [6, 11], and *Logistic regression* (LR) [18].

In recent Studies including PCA on breast cancer such as [14] created a Random forest PCA, which transformed the data so, that five principle components could describe 91.17% of the data or seven principle components that could describe 95.99% of the Wisconsin breast cancer data. [19] reduced the features from nine to six, [9] reduced the features form nine to 4, [7] extracted two principle components which accounted for 37% of the variance in the data analysed, whilst [12, 20] also used PCA on Wisconsin breast cancer data.

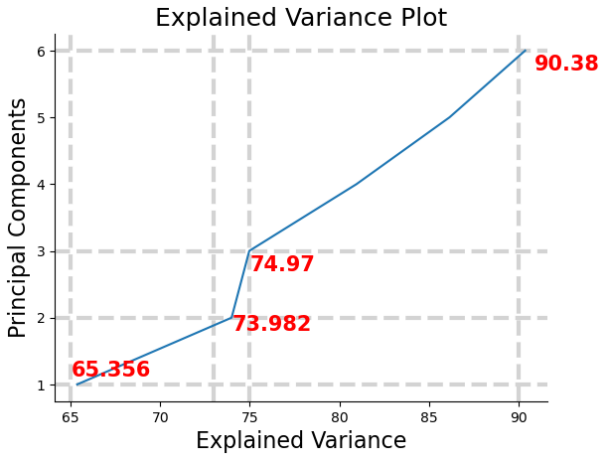

FIG. 1. Variance explained describes the variance from the principle components, which states that with six components we can describe 90% of the data

Studies including K-means clustering on Wisconsin breast cancer data-set are [21] compared foggy or random centroids, along with measuring Euclidean distance versus Manhattan and Pearson correlation, and found that Euclidean and Manhattan out performed Pearson's correlation. [20] used K-means as a dimension reduction technique and compared against PCA and found that K-means as a dimension reduction technique performed well against PCA. [22] created a hybrid K-means + *Gaussian mixture model* (GMM) that out performed other standalone models as K-means, GMM, SVM and thresholding. [23] argued that K-means is not a unsupervised method as we have to initialise clusters before hand, so created an algorithm that automatically picks the number of clusters, and [24] used K-means for outlier detection.

The aim of this project is to try and understand the difference in features for malignant and benign tumours, using Unsupervised ML methods such as PCA and K-mean algorithm. The rest of this report will be methodologies used, findings and results, we will then finish with

* ppxre1@nottingham.ac.uk
† ppxhl1@nottingham.ac.uk
‡ ppxpp1@nottingham.ac.uk

a further work section followed by a conclusion of our report..

II. METHODOLOGY

Two different approaches were used to pre-process the data. The first is to remove any rows which had a column containing a null value (represented by '?' for this dataset). By removing these rows, the dataset reduced its size from 699 rows to 683. Considering the size of this dataset is relatively small, we felt it important to keep as much data as possible, therefore, we implemented the second approach.

For our second approach, we replaced all of the missing values by interpolating them based on other existing values in the corresponding column. For that, we utilise pandas.DataFrame.interpolate and set the method parameter to 'pad', this method involves replacing the empty values with the most closely related existing value in another row, which prevents issues such as type errors from occurring. This method meant we were able to keep and utilize the entire dataset for our models.

After filling in the missing data values, we produced a normalized version of the data using Z-Score Normalization $X^{'} = \frac{X-\mu}{\sigma}$. Z-Score Normalization or Standardization will transform the features by subtracting from the mean and dividing by standard deviation. All the processed data will be then saved to a .csv file and were ready to be used in the later stage.

Principle component analysis (PCA) is a linear transformation, which transforms the data into a new set of coordinates, The new set of coordinates aligns in the direction with the features which has the largest variation in the data; is known as the first *principle component*, the second set of coordinates which should be orthogonal to the first principle component, and has the second largest variation in the data; is known as the second *principle component*. The n^{th} component is orthogonal to the $n-1$ principle component, whilst accounting to be the n^{th} largest variation in the data. PCA is a technique that is used in application such as lossy data compression, feature extraction or dimension reduction [25]. The aim of the algorithm is to find the maximum variations and achieve dimensional reduction, this can be achieved in the following steps: *Step 1:* Input original data matrix X :

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ x_{m1} & x_{m2} & \dots & x_{mn} \end{pmatrix} \tag{1}$$

Step 2: Subtract the average value from the original data to the new centralized data:

$$\text{mean} = \overline{\mathbf{X}} = \frac{1}{N}\sum_{j=1}^{N}\mathbf{X}^{(j)} \tag{2}$$

Step 3: Calculate the covariance matrix:

$$\mathbf{\Sigma} = \frac{1}{w-1}\sum_{j=1}^{N}(\mathbf{X}^{(j)} - \overline{\mathbf{X}}) \cdot (\mathbf{X}^{(j)} - \overline{\mathbf{X}})^{T} \tag{3}$$

Step 4: Calculate eigenvalues λ and eigenvectors $\mathbf{V}_\lambda$ of $\mathbf{\Sigma}$:

$$\mathbf{\Sigma} \cdot \mathbf{V}_\lambda = \lambda \mathbf{V}_\lambda (\text{with} |\mathbf{V}_\lambda| = 1) \tag{4}$$

$$\det(\mathbf{\Sigma} - \lambda \Pi) = 0, \forall \lambda \text{ finds } \mathbf{V}_\lambda \tag{5}$$

Step 5: Sort the eigenvalues from large to small, and select the largest k of them. Then the corresponding k eigenvectors are used as row vectors to form eigenvector matrix:

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_D \tag{6}$$

$$\mathbf{V}_{\lambda_1} = \text{PC}_1 \tag{7}$$

$$\mathbf{V}_{\lambda_2} = \text{PC}_2 \tag{8}$$

K-means clustering is an unsupervised learning method. given a set of variables $X = x_1, \dots, x_n$ our goal is to partition the data-set, in to some K-clusters and find groups of points that are close to each other, however far away from others, using a metric such as *Euclidean* distance $|X^{(i)} - X^{(j)}|$. Each cluster is defined entirely by it's centroid/center, or mean value u_k.

Algorithm 1 K-Means

1: Fix k Clusters
2: choose k random centroids
3: Assign each data element to a centroid; define clusters
4: Move centroids
5: repeat steps 3 & 4 until convergence

Centroids can be calculated by:

$$\mu^{(k)}, k = 1, 2, \dots K \tag{9}$$

$$S_k = \left\{ \mathbf{X}^{(j)} : |\mathbf{X}^{(j)} - \mu^{(k)}|^2 \leq |\mathbf{X}^{(i)} - \mu^{(k)}|^2 \right\} \tag{10}$$

$$\mu^{(k)} = \frac{1}{|S_k|}\sum_{\mathbf{X} \in S_k}\mathbf{X} = \text{ Near cluster k} \tag{11}$$

$$\{S_1, \dots, S_k\}^{\arg\min} \sum_{k=1}^{K}\sum_{\mathbf{X} \in S_k} |\mathbf{X} - \mu^{(k)}|^2 = \sum_{K} |S_k| \text{Var}_k(x) \tag{12}$$

In order to evaluate our k-means model we have opted to initialise K=2 to represent the *benign* and *malignant* classes, although K-means is a *Unsupervised* learning method, meaning we do not need class-labeled training examples for our model, however, by comparing the results against the original we can observe how the clustering performed which is a method also used by [20]. It also important to bare in mind that we could set the clusters to any number and there are methods to optimise this number such as [23] this would then be used as a dimension reduction technique, whereas the number of clusters would become the number of dimensions, then used with a classification algorithm such as SVM; whereas here we are using the K-means as a classifier. To evaluate our K-means model to compare the clusters against the original class, and create a confusion matrix, that consists of success instances: classes predicted correctly; true positives(TP) and true negative(TN),and error instances: classes predicted incorrectly; false positive(FP) and false negative (FN). Classification error rate can be to optimistic especially on unbalanced set so we will also look at *recall* = TP / (TP+FN), and *precision* = TP / (TP+FP), which gives us more insight into classes behaviour in regards to the positive class's. *F-measure* will combine the two *recall* and *precision* into a single metric,

$$f\beta = (1+\beta^2)\frac{\text{Precision} \times \text{Recall}}{(\beta^2 \times \text{Precision}\ \ \text{Recall}} \qquad (13)$$

III. RESULTS

PCA aim is to simplify the data, which creates a new coordinate system by altering the data linearly. fig. 2 depicts *principle component* with $n = 2$ with Different colours used for benign and malignant, however, some points are obscured and found not to be completely seperated, although in fig. 3 shows *principle component* with $n = 3$ and the separation is slightly better. PCA is explained by separating the data into n^{th} variations in the data. 1 depicts that variance by, explaining that 90% of the data can be explained with six of the features from the data set, which is in line with other research such as [19], although, [7] claimed that 2 principle components accounted for 37.5% of the data we managed 73.982%, and [14].

In this research, we performed a K-means cluster algorithm, on the breast cancer Wisconsin data using K= 2 to simulate the two classes, benign and malignant. fig. 4 Shows the two clusters separated by two different colours and Two centroids depicted in the center of each cluster. We then compared our clusters with the original classification from the data. The results are shown in table I.

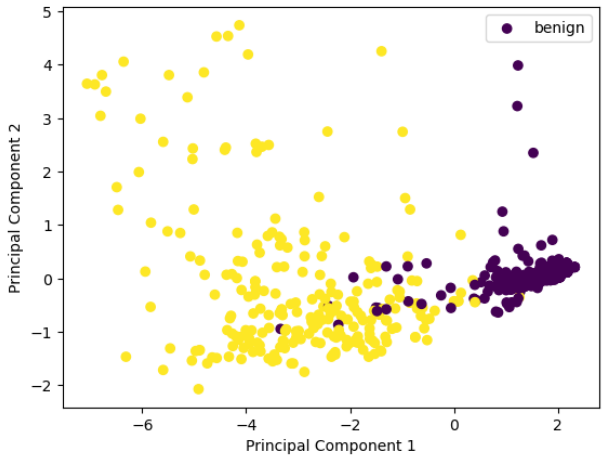

FIG. 2. Shows top 2 principle components from the breast cancer Wisconsin data set.Which shows benign and malignant tumors separated by colour. Although some points seem to be obscured and not completely separated.

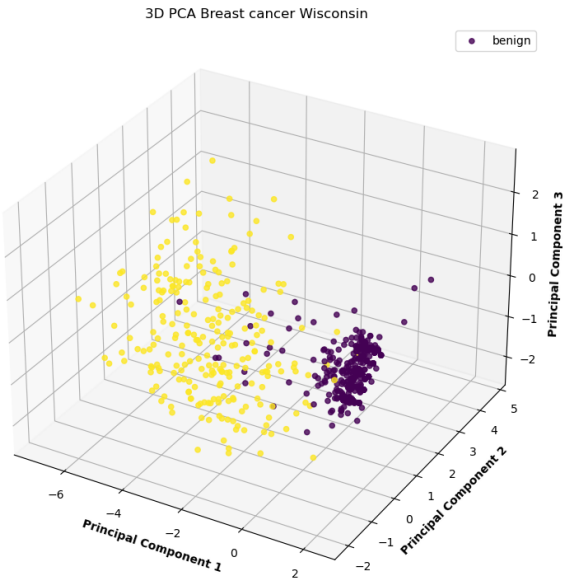

FIG. 3. Shows top 3 principle components from the breast cancer Wisconsin data set.Which shows benign and malignant tumors separated by colour, With 3 principle components shows better separation than with 2 principle components.

IV. FUTURE WORK

While this project has utilised two different approaches, Principal component analysis and K-means clustering, to have a deeper understanding of the aim of this project, which is to learn the difference in features for malignant and benign tumours, we felt that exploration of other Machine Learning methods is needed. Due to the limited time frame of this project, we decided to only implement PCA and K-means clustering, but in future

TABLE I. K-Means Results

Precision	Recall	F1-score	accuracy
0.95	0.97	0.96	0.95

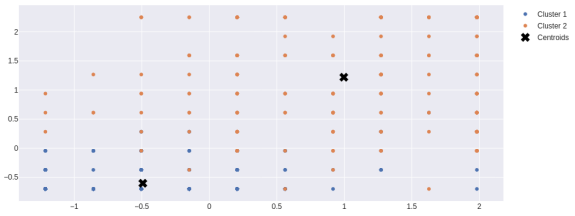

FIG. 4. Shows K-means clustering where $K = 2$ the centroids are shown in the center of each cluster.

work for this project, we agreed that Auto Encoder can be a great dimension reduction algorithm to implement. While PCA is restricted to a linear procedure, autoencoders can have nonlinear decoders and encoders, it will be great to compare the result with PCA.

V. CONCLUSIONS

In this project, we have performed Principal Component Analysis and K-means Clustering on a dataset which is the Winconsin breast cancer data collected by the [1] University of Wisconsin-Madison Hospital. The dataset includes 239 cases of malignant and 444 cases of benign tumours and the aim of this project is to understand the difference in the nine features for this dataset using Unsupervised ML methods like PCA and K-mean Clustering. We first pre-process the data by replacing the missing values with pandas.DataFrame.interpolate and normalize the values using Z-Score Normalization. For PCA, we implemented it with both $n = 2$ and $n = 3$, the features were reduced from nine to six and 90% of the data can be explained. For K-means Clustering, we implemented it using $K = 2$ and got the result shown in table table I.

[1] W. Wolberg and O. Mangasarian, Multisurface method of pattern separation for medical diagnosis applied to breast cytology., In Proceedings of the National Academy of Sciences, **87**, 9193 (1990).

[2] S. M. Shah, R. A. Khan, S. Arif, and U. Sajid, enArtificial intelligence for breast cancer analysis: Trends & directions, Computers in Biology and Medicine **142**, 105221 (2022).

[3] UCI Machine Learning Repository: Data Sets.

[4] A. Marcano-Cedeño, J. Quintanilla-Domínguez, and D. Andina, enWBCD breast cancer database classification applying artificial metaplasticity neural network, Expert Systems with Applications **38**, 9573 (2011).

[5] E. D. Übeyli, enImplementing automated diagnostic systems for breast cancer detection, Expert Systems with Applications **33**, 1054 (2007).

[6] A. Aloraini, enDifferent Machine Learning Algorithms for Breast Cancer Diagnosis, International Journal of Artificial Intelligence & Applications **3**, 21 (2012).

[7] A. Buciński, T. Baczek, J. Krysiński, R. Szoszkiewicz, and J. Załuski, enClinical data analysis using artificial neural networks (ANN) and principal component analysis (PCA) of patients with breast cancer after mastectomy, Reports of Practical Oncology & Radiotherapy **12**, 9 (2007).

[8] M. Karabatak and M. C. Ince, enAn expert system for detection of breast cancer based on association rules and neural network, Expert Systems with Applications **36**, 3465 (2009).

[9] M. S. Uzer, O. Inan, and N. Yılmaz, enA hybrid breast cancer detection system via neural network and feature selection based on SBS, SFS and PCA, Neural Computing and Applications **23**, 719 (2013).

[10] M. F. Akay, enSupport vector machines combined with feature selection for breast cancer diagnosis, Expert Systems with Applications **36**, 3240 (2009).

[11] K. Sivakami, enMining Big Data: Breast Cancer Prediction using DT - SVM Hybrid Model, **1** (2015).

[12] H.-J. Chiu, T.-H. S. Li, and P.-H. Kuo, Breast Cancer–Detection System Using PCA, Multilayer Perceptron, Transfer Learning, and Support Vector Machine, IEEE Access **8**, 204309 (2020), conference Name: IEEE Access.

[13] K. Menaka and S. Karpagavalli, enBreast Cancer Classification using Support Vector Machine and Genetic Programming, International Journal of Innovative Research in Computer and Communication Engineering **1** (2007).

[14] K. Bian, M. Zhou, F. Hu, and W. Lai, enRF-PCA: A New Solution for Rapid Identification of Breast Cancer Categorical Data Based on Attribute Selection and Feature Extraction, Frontiers in Genetics **11**, 566057 (2020).

[15] P. K. P, M. A. B. V, and G. G. Nair, enAn efficient classification framework for breast cancer using hyper parameter tuned Random Decision Forest Classifier and Bayesian Optimization, Biomedical Signal Processing and Control **68**, 102682 (2021).

[16] C.-Y. Fan, P.-C. Chang, J.-J. Lin, and J. Hsieh, enA hybrid model combining case-based reasoning and fuzzy decision tree for medical data classification, Applied Soft Computing **11**, 632 (2011).

[17] M. Seera and C. P. Lim, enA hybrid intelligent system for medical data classification, Expert Systems with Applications **41**, 2239 (2014).

[18] Y. Zhao, N. Wang, and X. Cui, enAided diagnosis methods of breast cancer based on machine learning, Journal of Physics: Conference Series **887**, 012072 (2017).

[19] S. Liang, M. Singh, S. Dharmaraj, and L.-H. Gam, enThe PCA and LDA Analysis on the Differential Expression of Proteins in Breast Cancer, Disease Markers **29**, 231 (2010).

[20] A. Jamal, A. Handayani, A. A. Septiandri, E. Ripmiatin, and Y. Effendi, enDimensionality Reduction using PCA and K-Means Clustering for Breast Cancer Prediction, Lontar Komputer : Jurnal Ilmiah Teknologi Informasi , 192 (2018).

[21] A. K. Dubey, U. Gupta, and S. Jain, enAnalysis of k-means clustering approach on the breast cancer Wisconsin dataset, International Journal of Computer Assisted Radiology and Surgery **11**, 2033 (2016).

[22] P. E. Jebarani, N. Umadevi, H. Dang, and M. Pomplun, A Novel Hybrid K-Means and GMM Machine Learning Model for Breast Cancer Detection, IEEE Access **9**, 146153 (2021), conference Name: IEEE Access.

[23] K. P. Sinaga and M.-S. Yang, Unsupervised K-Means Clustering Algorithm, IEEE Access **8**, 80716 (2020), conference Name: IEEE Access.

[24] G. Gan and M. K.-P. Ng, enk -means clustering with outlier removal, Pattern Recognition Letters **90**, 8 (2017).

[25] C. Bishop, *Pattern Recognition and Machine Learning*, 1st ed. (Springer International Publishing, Singapore Singapore, 2006).