Journal of Applied Sciences, Information and Computing

Volume 5, Issue 1, June 2024

School of Mathematics and Computing, Kampala International University

ISSN: 1813-3509 https://doi.org/10.59568/JASIC-2024-5-1-10

Exploring Dimensionality Reduction Techniques for Improved Breast Cancer Diagnosis

[1]Akampurira Paul, [2]Semalulu Paul, [3]Elly Gamukama, [4]Kareyo Margaret

[1]Kampala International University, Kampala, Uganda, akampurira.paul@kiu.ac.ug
[2]Kampala International University, Kampala, Uganda, paul.semalulu@kiu.ac.ug
[3]Kampala International University, Kampala, Uganda, elly.gamukama@kiu.ac.ug
[4]Kampala International University, Kampala, Uganda, margaret.kareyo@kiu.ac.ug

Abstract:

Breast cancer diagnosis is a critical area in medical research, where the challenge lies not only in accurate identification but also in managing the inherent complexity of high-dimensional datasets. This paper navigates this challenge by exploring dimensionality reduction techniques to enhance diagnostic accuracy. The primary objective of this research was to employ dimensionality reduction methods to refine breast cancer diagnosis, with a focus on improving accuracy and interpretability. The study investigates the impact of preprocessing techniques on a high-dimensional dataset, aiming to uncover meaningful patterns for effective diagnostic models. Starting with a dataset including 569 observations and 30 attributes, careful examination reveals imbalances in the dataset (63% benign, 37% malignant). To deal with multicollinearity, we use the coefficients of Pearson correlation to find and eliminate highly correlated features. Subsequent data transformation, utilizing min-max normalization, ensures uniform weighting. Principal Component Analysis (PCA) is then leveraged for comprehensive dimensionality reduction. Visualizations through scree plots and bi-plots underscore the efficacy of early principal components in distinguishing benign from malignant cases. Our results demonstrate a notable 24% reduction in data dimensionality, affirming the process's efficiency. This abstract beckons the exploration of detailed findings, emphasizing dimensionality reduction's pivotal role in refining breast cancer diagnosis for more accurate, efficient, and interpretable models.

Keywords: High-dimensional datasets, early diagnosis, breast cancer, dimensionality reduction, artificial intelligence, and machine learning.

I. Introduction

In 2018, cancer was the second-leading cause of death worldwide accounting for about 9.6 million deaths (WHO, 2018). Breast cancer, a prevalent and fatal form, causes around 2.09 million deaths annually, with 70% occurring in low- and middle-income countries. The urgency to address breast cancer arises from its status as the most common cancer among women, constituting 25% of all

cancer cases. Alarmingly, late-stage presentations, inaccessible diagnosis, and limited treatment options persist, particularly in low-income countries. Breast cancer is a major global health concern, causing millions of deaths annually, particularly affecting women. The incidence is high, with about 2.1 million cases reported yearly, contributing to approximately 15% of all cancer-related deaths in women. Late-stage presentation and inadequate diagnostic and treatment accessibility, especially in low-income countries, are common issues (Louise Wilkinson, 2021). In 2018, Uganda reported 22,000 cancer-related deaths, emphasizing the dire consequences of late-stage diagnoses (Uganda National Cancer Institute, 2019). The statistic that If found early, 30% of cancer cases are curable emphasizes the vital need for preventative measures. The mortality rate is exacerbated by late diagnoses. Early diagnosis strategies are crucial, aiming to increase early-stage identification through improved access to breast cancer treatment and effective diagnostic services (Tobore Onojighofia, 2019).

The urgency to address breast cancer as a global health concern, particularly in low- and middle-income countries, underscores the critical need for innovative and efficient diagnostic approaches. Breast cancer, claiming millions of lives annually, manifests as a pervasive and fatal disease, necessitating a strategic shift in diagnostic methodologies. The high incidence of breast cancer, constituting a significant portion of all cancer cases among women, coupled with persistent challenges like late-stage presentations and limited accessibility to accurate diagnosis and treatment, amplifies the urgency for transformative solutions (Louise Wilkinson, 2021).

Computer-aided diagnostic (CAD) systems play a vital role in classifying malignant and benign cancers, enhancing physician performance by reducing misdiagnoses and diagnosis time (Chhatwal, 2010). Machine learning (ML), a subset of artificial intelligence, has been extensively employed in cancer detection and diagnosis, utilizing various classification algorithms. Despite technological advancements, challenges persist, especially in low-income countries. AI, coupled with Electronic Medical Records (EMRs), presents a transformative potential for healthcare services (Blümel et al., 2020). However, the applicability and success of ML in low-resource settings, including low-income countries, are underexplored. The need for accurate diagnostic tools in resource-poor environments is evident, and AI applications, such as Natural Language Processing (NLP), are already making strides in guiding cancer treatments (Chaurasia et al., 2018). The contextual background recognizes the potential of ML in reshaping healthcare delivery, emphasizing the need for efficient and accurate diagnostic tools in diverse settings (Blümel et al., 2020). Also, ensemble learning, a promising approach, combines multiple classifiers to improve predictive performance (Rokach, 2010).

Breast cancer diagnosis faces challenges posed by high-dimensional datasets, necessitating advanced techniques for effective model development.
The challenge in breast cancer diagnosis is exacerbated by the multitude of features contributing to the determination of malignancy or benignity. Human interpretation, often subjective and dependent on personal experience, poses limitations in accurately representing the facts, especially as the number of samples increases.
In the backdrop of the escalating impact of breast cancer, the adoption of CAD systems and ML holds promise. However, the inherent complexity of high-dimensional datasets poses a substantial hurdle. The multitude of features contributing to malignancy or benignity demands a nuanced approach to enhance accuracy and interpretability. Traditional diagnostic methods, often reliant on subjective human interpretation, falter in the face of increasing sample sizes and diverse datasets. CAD systems, are designed to reduce misdiagnosis and expedite the diagnostic process. However, the effectiveness of these systems is contingent on overcoming the dimensionality of the data, making the case for advanced computational techniques (Chhatwal, 2010).

High-dimensional datasets not only strain computational resources but also risk introducing noise and irrelevant features, potentially hampering the accuracy of diagnostic models (Ricvan, 2018). By reducing dimensionality, the focus shifts to the most informative features, enhancing the efficiency and interpretability of the diagnostic process. The significance of dimensionality reduction is magnified by its potential to address late-stage presentations and limited accessibility to accurate diagnosis, particularly in resource-poor environments (Vogelstein, 2021). By streamlining datasets and uncovering meaningful patterns, dimensionality reduction techniques offer a pathway to more efficient, accurate, and accessible diagnostic models. The exploration of these techniques aligns with the transformative potential of artificial intelligence in reshaping healthcare delivery.

Our study therefore embarked on practical experimentation to uncover the intrinsic importance within the capabilities of techniques used in dimensionality reduction towards breast cancer diagnosis (Akampurira, 2022).

II Research Methodology:

A dataset comprising 30 features and 569 observations related to breast cancer cases. The dataset exhibited inherent complexities, including an

 https://doi.org/10.59568/JASIC-2024-5-1-10

imbalance with 63% benign and 37% malignant cases. The diverse nature of the dataset posed a challenge in developing accurate diagnostic models.

Data Preprocessing:
To address the issue of multicollinearity, Pearson correlation coefficients were utilized to discern and eliminate highly correlated features. This procedural step was pivotal in fortifying the resilience of subsequent analyses. Following the correlation analysis, min-max normalization was implemented to ensure the uniform weighting of features, thereby alleviating the impact of disparate scales on the model development process.

Detection, neutralization, and or removal of Outliers:
Outliers can be delineated as anomalous data or values that deviate from the norm in comparison to the majority of observations. Typically stemming from measurement errors, coding discrepancies, or, at times, representing naturally occurring abnormal values, these non-representative samples wield considerable influence on later model outcomes. Consequently, a meticulous examination of the data was conducted to identify and either neutralize or expunge outliers, contingent upon their perceived impact.

Handling missing data
The most straightforward recourse for addressing missing data involves downsizing the dataset by discarding all samples with incomplete values. This approach is particularly applicable to expansive datasets where missing values constitute a negligible proportion relative to the entirety of the dataset. Alternatively, if the researcher opts against discarding samples with missing values, efforts must be made to impute suitable values in their stead.
Normalization
There are several approaches to data normalization, such as Z-score normalization, Min-Max normalization, and decimal scaling. Because the former works with most of the methods used in the normalization process, it was used in this investigation.

Data reduction employing feature selection and extraction
In order to determine the aspects of the dataset's relevance for the result or target variable and their interrelationships, the researcher dug further in dataset during this phase. Features that were judged irrelevant were removed, collinearity tests were carried out, and features with a high degree of correlation were carefully handled. Furthermore,

Dimensional space was reduced by the application of Principal Component Analysis (PCA), a flexible method for lowering the dimensionality of data and fine-tuning feature selection criteria.

The researcher used Principal Component Analysis (PCA) for this research even though there are other methods for reducing dimensionality, such as the relief approach, entropy-based feature ranking, Chi Merge, value elimination, and case reduction. This is because PCA employs procedures that are straightforward but comprehensive. The dataset, which is represented by vector samples, was changed into a new collection of vector samples with generated dimensions using PCA.

III. Results and Discussion:

Dataset:
The Wisconsin Breast Cancer Database (WBCD) dataset, which is often used in research studies, was employed in this study. The feature values derived from a digital picture of a Fine Needle Aspirate (FNA) of a breast mass make up the WBCD dataset for breast cancer diagnosis. The attributes of the cell nuclei shown in the picture are described by these features. The UW CS file transfer protocol (FTP) server, located at ftp.cs.wisc.edu/cd math-prog/cpo-dataset/machine-learn/WDBC, is another way for you to access this database. It is advised that data science initiatives make use of this publicly available standard dataset.
The data included different attributes including ID number, Diagnosis (M = malignant, B = benign), and For each cell nucleus, the following ten real-valued features are calculated: "compactness (perimeter^2 / area - 1.0), concavity (severity of concave sections of the contour)," "concave points (number of concave portions of the contour)," "texture (standard deviation of gray-scale values), a perimeter, area, smoothness (local variation in radius lengths)," and symmetry, fractal dimension ("coastline approximation" - 1)."

The mean, average, standard deviation, and "worst" or worst (mean of the three most significant values) were also computed for each image, resulting in a total of 30 features. The Mean Radius, for example, is field 3, the Radius SE is field 13, and the Worst Radius is field 23. Every feature value has four meaningful digits recorded. The downloaded dataset was imported and stored into the RStudio, an integrated development environment (IDE) for R, which provides free and open-source tools for R programming and statistical modeling and is an enterprise-ready professional software for data science.
We employed the function view () to quickly glance at our data in the manner shown below. A glance at the imported data (Table 1) showed that our data is made of 32 columns as features (variables) and 569 rows as examples. No parameter had spaces in their names, and our data was rather clean which would

 https://doi.org/10.59568/JASIC-2024-5-1-10

have been a naming convention that is not compatible with many of the R procedures we would implement.
We also used head() to examine the data's structure, which provides a thorough picture of the information in terms of the feature data structures, as follows;

Table 1: Data import view in R studio

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean
1	87139402	B	12.320	12.39	78.85	464.1	0.10280	0.06981
2	8910251	B	10.600	18.95	69.28	346.4	0.09688	0.11470
3	905520	B	11.040	16.83	70.92	373.2	0.10770	0.07804
4	868871	B	11.280	13.39	73.00	384.8	0.11640	0.11360
5	9012568	B	15.190	13.21	97.65	711.8	0.07963	0.06934
6	906539	B	11.570	19.04	74.20	409.7	0.08546	0.07722
7	925291	B	11.510	23.93	74.52	403.5	0.09261	0.10210
8	87880	M	13.810	23.75	91.56	597.8	0.13230	0.17680

Showing 1 to 8 of 569 entries, 32 total columns

Table.2: Data head Preview

```
> head(bc_data)
# A tibble: 6 x 32
        id diagnosis radius_mean texture_mean perimeter_mean area_mean
     <dbl> <chr>           <dbl>        <dbl>          <dbl>     <dbl>
1 87139402 B                12.3         12.4           78.8      464.
2  8910251 B                10.6         19.0           69.3      346.
3   905520 B                11.0         16.8           70.9      373.
4   868871 B                11.3         13.4           73        385.
5  9012568 B                15.2         13.2           97.6      712.
6   906539 B                11.6         19.0           74.2      410.
# ... with 26 more variables: smoothness_mean <dbl>, compactness_mean <dbl>,
```

The bulk of the data properties were kept in double-precision floating-point (dbl) and character (chr) formats, based on the results shown in Table 2. There were 569 components or instances in the large 32-column array that made up the dataset. Among these are a 'id' column and labels designating the target values as 'B' for Benign and 'M' for Malignant. The data was initially preprocessed
to prepare it for exploration and visualization including rearranging the features (columns), ordering the columns, removing unnecessary features like “id” and replacing the diagnosis labels with full names.

Table 3: diagnosis label redefined

```
> head(bc_data)
# A tibble: 6 x 31
  diagnosis area_mean area_se area_worst compactness_mean compactness_se compactr
  <chr>          <dbl>   <dbl>      <dbl>            <dbl>          <dbl>
1 benign          464.    17.4       549.           0.0698         0.0118
2 benign          346.    27.1       425.           0.115          0.0358
3 benign          373.    13.5       471.           0.0780         0.00936
4 benign          385.    26.3       434            0.114          0.0350
5 benign          712.    17.7       819.           0.0693         0.0148
6 benign          410.    20.3       520.           0.0772         0.0205
```

Subsequently, we removed characteristics such as the ID variable that are not needed at all for data modeling. We observed that the target variable's labels, malignant or non-cancerous, were, respectively, m and b. For easy understanding of the data, we need the full names of the diagnosis field and therefore replaced the labels as malignant and benign as in Table 3. 30 features or predictors and 569 observations are visible in the raw data count following the first round of preprocessing. Furthermore, we observe that every predictor has constant outcomes for observations as well as no values that are absent. We observed that every observation was documented as a series of decimal numbers. Also, a quick count of cancer rates in the data set was done. We quickly counted the samples in our dataset to confirm their quantity and their appropriate categories:

 https://doi.org/10.59568/JASIC-2024-5-1-10

Table 4: cancer rate count

Benign	malignant
357	212

benign	0.63
malignant	0.37

Percentage is the diagnosis, in percentages returns as in the table 5.

The table 5, above shows that the response variable

Table 4 above shows the target parameter for diagnosis, which may be benign or malignant. The tables show that out of 569 observations, 357 were non-cancerous observations or benign, and 212 were cancerous or malignant. We further checked **for balance** in our response (target) variable, which

Table 5: Response variable

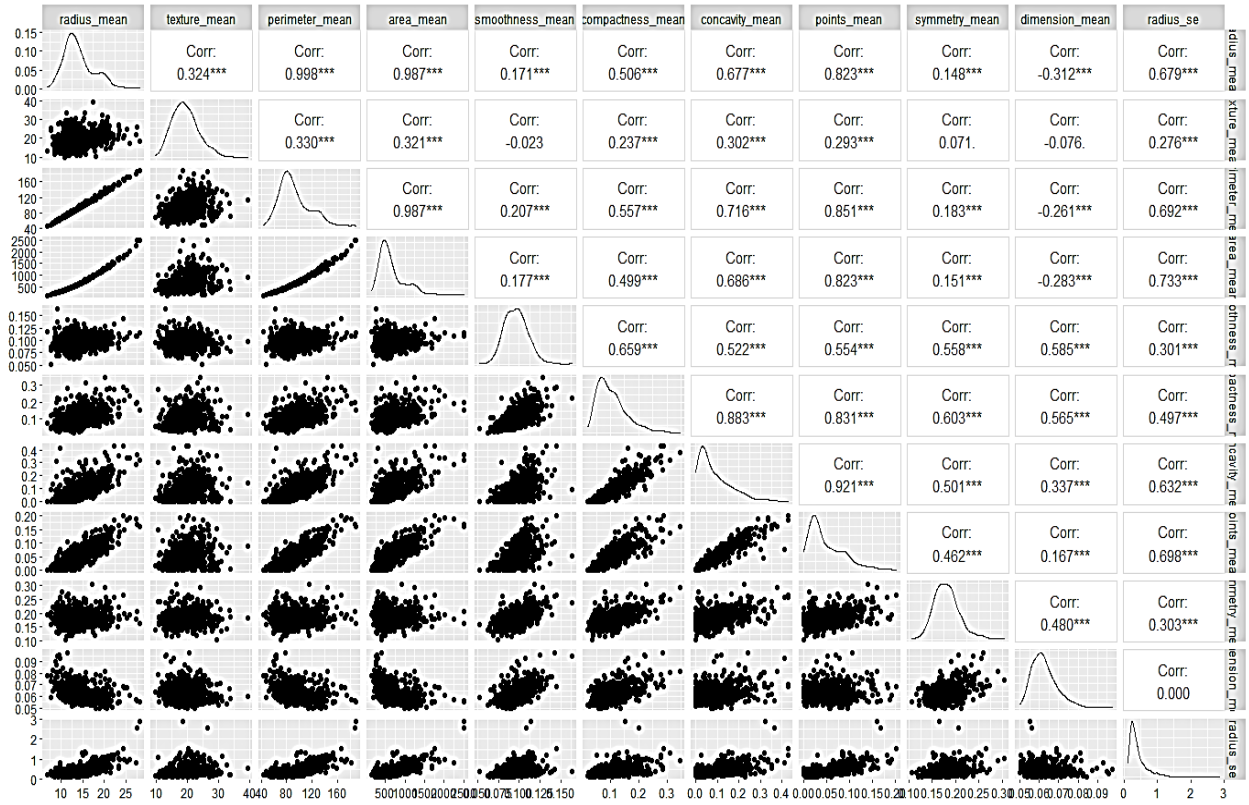


Fig 1: **Visualizing correlations with Corr plot.**

Mathematics, two random variables, x and y, have a tends to benign and malignant are only 37% of the entire dataset. This showed that 37% of the patients were diagnosed with cancerous cells. The balance check therefore shows that the data is a bit unbalanced.

Inspection for multicollinearity

A multicollinearity analysis was performed in an attempt to identify any correlation between the variables. Most machine learning methods require that the variables that predict outcomes be independent of each other in order for the evaluation to be deemed robust. This is why the researcher carried out a study that led to the multicollinearity's identification and removal. We used Pearson correlation to search for relationships between the features in our dataset.

The Pearson correlation coefficient (ρ) can be expressed in the following way.

$$\rho_{x,y} = \frac{Cov(x,y)}{\sigma_x \sigma_y}$$

In this case, y is the standard deviation, σx is the deviation from the mean of x, and Cov (x y) is the covariance of x. The above is accomplished in R using the cor() function as follows:

Fig 1: **Visualizing correlations with Corr plot**

In Figure 1, the size and color intensity of the circles indicate the correlation strength, or the total amount of the correlation coefficient among two variables.

 https://doi.org/10.59568/JASIC-2024-5-1-10

Positive correlates are blue, whereas the negative are red. The graph highlights the presence of linked variables; concavity worst and points worst are two examples of such relationships; these and other examples were covered in later procedural phases. This graphical representation undeniably attests to

Figure 2: Correlation Plots for Dataset Features

The existence of variables with correlations. Any component in our visual representation that registers a value of 0.9 or above indicates an exceptionally strong positive association, according to Pearson correlation coefficients, which range from -1 to 1. On the other hand, features with a correlation of -0.9 or less indicate a strong negative association; hence, their elimination is required for improved model performance. Three prominent instances of characteristics with strong positive correlations are texture_mean, texture_worst, and Arese. The next section outlines the methods that the researcher used to examine strongly linked data using the caret package.

Checking for multicollinearity among the features

Through the aforementioned correlation analysis, a nuanced elucidation of the interrelationships among features was attained. The correlation coefficients delineated the extent to which certain features exhibit a pronounced interdependence, thereby potentially compromising the robustness of our modeling outcomes. Consequently, this facilitated the researcher in identifying and subsequently mitigating such correlations, notably exemplified by features such as area mean and radius mean. To address this, the researcher opted for the implementation of principal component analysis, a strategy expounded upon in subsequent sections. Prior to delving into this analytical approach, a more granular examination of correlations was conducted using scatter diagrams, detailed in

The visual representation of correlation plots offered insights into the interconnectedness of distinct features. It is imperative to underscore that correlation, as depicted herein, is not tantamount to causation; rather, it serves as an illustrative indicator of observed associations. Noteworthy patterns emerged, elucidating a robust positive correlation among radius mean, area mean, and perimeter mean. Furthermore, favorable relationships were found between the radius mean the concavity mean and the compactness mean. The intrinsic skewness in the data was clarified by the scatter diagrams, which also disclosed the distributional properties of the features.

In an attempt to mitigate the impact of strongly correlated components, the researcher employed the discover correlation () function from the caret package. Using a heuristic technique, this function consistently identified variables for deletion with a Pearson's correlation coefficient equal to or better than 0.9. The function to remove characteristics with such high correlations was then run by the researcher, and a refined dataset known as bc_data_corr1 was produced. Wisc_bc_data %> bc_data_corr1 <- %-find Relationship (bc_data, cutoff = 0.9)) is selected. > n col (corr1, bc) [1] 22. Following the aforementioned change, the dataset has 10 variables and contains just 22 predictors (bc_data_Corr1).

Normalizing our data

The researcher undertook a consequential measure in the form of data normalization, a pivotal procedure primarily executed to mitigate bias stemming from the incongruity in the significance of absolute quantities compared to their relative counterparts, attributable to variations in scale. The normalization process was instrumental in ensuring parity among variables, thereby conferring uniform weight to each during the modeling phase.

Employing the min-max normalization method, we systematically altered a feature to confine its values within the spectrum of 0 to 1. The normalization of a feature adhered to the subsequent formula:

$$Xnew = \frac{X - min(X)}{\max(X) - \min(X)}$$

In essence, the formula divides by the range of X after subtracting the lowest value of X from each instance of feature X. The resultant normalized feature values can be thought of as the proportion of difference, on an integer ranging from 0 to 100, where the starting value falls between the lowest and maximum values. A normalization function was devised to standardize our data onto a uniform scale. Subsequently, this normalization function, denoted as "normalize ()," was applied to columns 2 through 30 (excluding the diagnosis variable) in the bc_data data frame. The output, converted into a data frame, was then assigned to the variable bc_data_norm. The "_norm" suffix is utilized solely as a mnemonic, underscoring the fact that the values in the dataset have undergone normalization. This method facilitated the creation of a standardized framework, fostering equitable treatment of variables and enhancing the robustness of the modeling process.

Dimensionality lessening through Primary component examination

Dimensionality lessening is a sophisticated procedure entailing the contraction of the feature space, or dimensions, within a dataset before subjecting it to model training. The investigator undertook dimensionality reduction with the primary objective of curtailing the temporal and storage requisites for data processing. This endeavor sought to enhance data visualization and refine models.

 https://doi.org/10.59568/JASIC-2024-5-1-10

Interpretability, and circumvent the deleterious effects of the curse of dimensionality. Fundamentally, the purpose was to excise superfluous and duplicative data, thereby diminishing computational costs and mitigating the risk of overfitting. The principal methodologies encompassed in this pursuit are feature selection and feature extraction.

When it comes to characteristic picking, a selected group of variables is carefully chosen to produce a small number of features, free from redundant or inconsequential attributes, which can be expunged without significantly impacting model performance. In this instance, the variable 'id' was expunged due to its lack of relevance in the modeling process. While alternative methodologies such as the F selection method exist, capable of removing less influential features, a deliberate decision was made to abstain from further feature removal in our dataset.

Table 6: Summary of PCA results

In light of the potential elimination of some

```
> summary(bc_data_norm.pca)
Importance of components:
                          PC1    PC2     PC3     PC4     PC5
Standard deviation     3.6444 2.3857 1.67867 1.40735 1.28403
Proportion of Variance 0.4427 0.1897 0.09393 0.06602 0.05496
Cumulative Proportion  0.4427 0.6324 0.72636 0.79239 0.84734
                           PC6     PC7     PC8    PC9    PC10
Standard deviation     1.09880 0.82172 0.69037 0.6457 0.59219
Proportion of Variance 0.04025 0.02251 0.01589 0.0139 0.01169
Cumulative Proportion  0.88759 0.91010 0.92598 0.9399 0.95157
                          PC11    PC12    PC13    PC14    PC15
Standard deviation     0.5421 0.51104 0.49128 0.39624 0.30681
Proportion of Variance 0.0098 0.00871 0.00805 0.00523 0.00314
Cumulative Proportion  0.9614 0.97007 0.97812 0.98335 0.98649
                          PC16    PC17    PC18    PC19    PC20
Standard deviation     0.28260 0.24372 0.22939 0.22244 0.17652
Proportion of Variance 0.00266 0.00198 0.00175 0.00165 0.00104
```

correlated

Figure 2: Correlation Plots for Dataset Features

however educational elements The investigator's task was to use the feature extraction approach to combine these related features into one entity through the selection of features procedure. It is possible to use feature extraction and projection interchangeably entails the application of mathematical functions to effectuate the transformation of high-dimensional data into lower dimensions, with the newly derived features supplanting their original counterparts. The preeminent methodology for such vibrant feature extraction is Principal Component Analysis (PCA).

As expounded earlier, PCA constitutes a method for extracting linear features from data initially stored in a higher-dimensional space, using a reduced-dimensional space. With the use of this method, the researcher was able to carry out an analysis that would maximize the variation of data in its low-dimensional presentation by mapping the data onto a lower dimension. The proclivity towards employing PCA stemmed from its unparalleled efficacy in scenarios where datasets exhibit a profusion of features coupled with inter-feature redundancy or correlation—a circumstance substantiated by our antecedent investigation into multicollinearity within the dataset.

Consequently, to excise superfluous features characterized by redundancy, PCA was enlisted to transmute high-dimensional data into lower dimensions by condensing features into a concise set of principal components, aptly capturing the majority of the variance inherent in the original features.

The process was done using the following steps;

1. Finding the mean vector $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$ where x_i is the number of points, and shows the data points.
2. Calculating the matrix of covariance $C = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)^T$
3. Calculating the associated eigenvalues and eigenvectors, φ.
4. Selecting and ranking the highest k eigenvectors.
5. Construct a n x k matrix of dimensional eigenvectors, U. In this case, k denotes the number of eigenvectors and n represents the total number of original dimensions.
6. Convert the collected samples to the new subdomain in the formula. $y = U^T.x$

An overview of the PCA findings

The results of the PCA implementations are summarized in the table above. As seen in Table 6, the first five PCs account for 84.73% of the variance, while the first 15 components account for 98.64% of the variance.

Eigen-values and component importance using the covariance matrix

The outcomes of the primary components were obtained by using the predict function: Obtain the Eigenvalues of the correlation matrix to further highlight the significance of the individual parts.

Table 7: **Eigenvalues using covariance matrix**

```
> round(bc_data_norm.pca$sdev ^2,4)
 [1] 13.2816  5.6914  2.8179  1.9806  1.6487  1.2074  0.6752  0.4766
 0.4169  0.3507  0.2939
[12]  0.2612  0.2414  0.1570  0.0941  0.0799  0.0594  0.0526  0.0495
 0.0312  0.0300  0.0274
[23]  0.0243  0.0181  0.0155  0.0082  0.0069  0.0016  0.0007  0.0001
```

Table 7 indicates that the components with tiny eigenvalues exhibit low fluctuation, indicating a minimal impact on the target projection or the diagnosis's outcome values. Using a scree plot, we further display the principle to comprehend the

 https://doi.org/10.59568/JASIC-2024-5-1-10

component importance as follows:

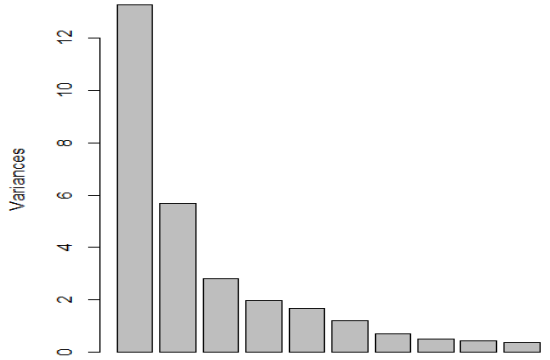


Fig 3: scree plot to visualize the relative importance of principal components

The scree plot above shows a detail of component importance: the y-axis is the eigenvalue which shows the importance of the principal components. The figure shows the first ten principal components and their contribution towards the prediction of the target variable. To further understand this, we used biplots to offer a detailed view. To find out how data and variables are mapped regarding the principal component, we used a biplot, which plots data and the projections of original features. Figure 4, shows that the first two components greatly managed to separate the diagnosis well. PC1 shows a greater influence on discrimination between benign and malignant. However, we want to get a more detailed analysis of what variables are the most influential in the first two components. We also wanted to explain the difference between malignant and benign tumors. So we added the response variable (diagnosis) to the plot and see if we can make better sense of it:

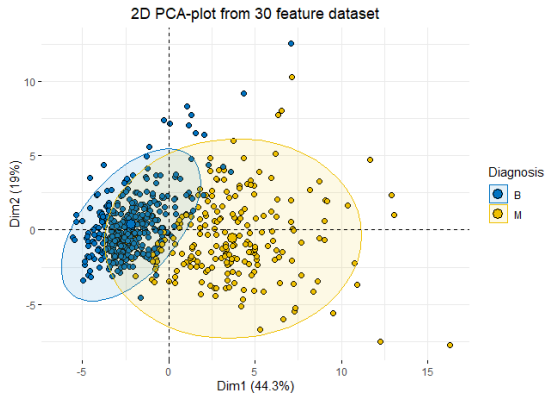


Fig 4: First 2 PCA features

There is a clear distinction between tumors that are malignant and benign in the first two components. This indicates that the data is suitable for use with a classification model, like discriminant analysis. The notable difference found between the 'Malignant' and 'Benign' classifications, according to around 63% of the variation in a 30-dimensional dataset, highlights the possible effectiveness of using just two perspectives in a story. Although these dimensions might produce very accurate estimates, dealing with higher-dimensional data is difficult but also captures a larger degree of variability.

Remarkably, we ascertain that over 60% of the variance can be elucidated by employing solely the initial two components. The variance of each statistic from its average is explained by representing the variables in question as vectors or arrows, where the origin represents the mean value and the data points or sampling identifiers indicate the scores. Notably, the average is positioned at a zero value, serving as the centroid in the data matrix. The length of the arrows directly correlates with variability, offering a proportional depiction.

The angular disposition between two arrows symbolizes the correlation between variables, with acute angles signifying robust positive correlations and greater obtuse angles indicative of negative correlations. To delve deeper into these relationships, corrplots were employed, visually portraying the trajectory of component variability. The ensuing visualization serves to explicate the significance of variables in the overall analysis.

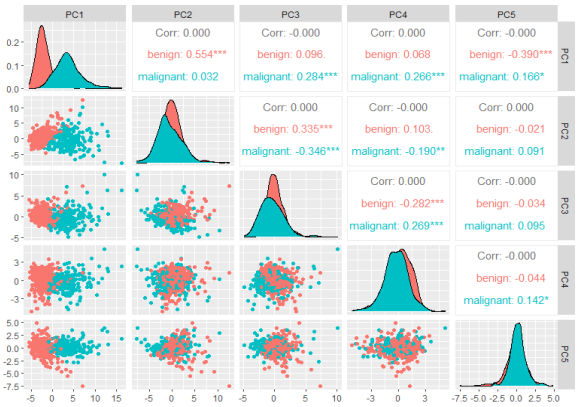


Fig 5: a correlation plot of the first five principal components

The figure above shows the application of PCA and determining the importance of components using the bc_data_corr1 dataset where highly correlated variables were removed. The results show how the first component performs very well on the data. The importance decreases as we move components from PC2 to PC5.

PCA Using our normalized dataset: We also used our normalized dataset to perform the PCA as follows. We do this as a final analysis to determine which features are more important and create a subset of the original dataset that we can use for the next steps in the modeling phase. The Summary of In the PCA on the normalized dataset, a test of the

 https://doi.org/10.59568/JASIC-2024-5-1-10

claim that the five parts are sufficient results in a mean item level of complexity 2.2. The fit is based on off-diagonal values of 0.99, with an experimental chi-square of 912.23 and a probability of less than 3e-64. The root means square of the residuals (RMSR) is 0.04.

The total weight of the main components is shown by the SS loadings: PC1 weighs 13.28, PC2 weighs 5.69, PC3 weighs 2.82, PC4 weighs 1.98, and PC5 weighs 1.65. The overall difference

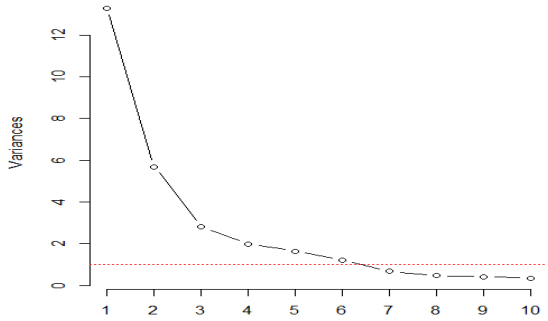
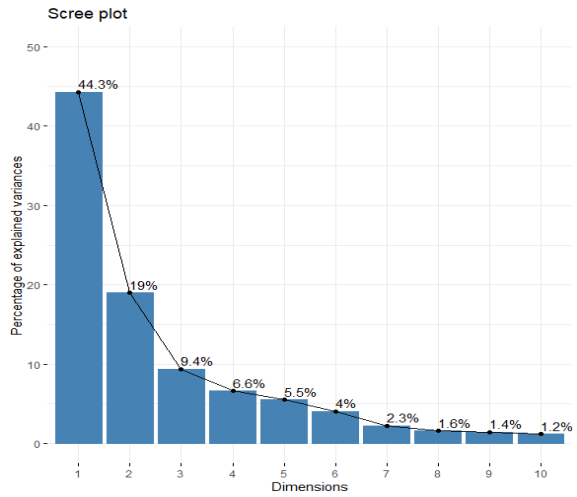


Fig 6: PCA using cumulative variance

According to the figure, 85% of the total variation explained is covered by the first five PCs. The analysis's conclusions showed that the data's significant variance can be explained by principal component one. According to the findings, the first six main components account for the majority of variance. Using the scree plot with the cut-off line, we were able to further illustrate the feature extraction.



Scree plot showing the proportion of variations explained in Figure 7

The statistical importance of the main elements and the variance that is explained are shown in Figure 7, where PC1 makes up 44.3% of the total variance described and the first five elements account for 85% of the variation explained. The first ten primary components explain 95% of the variance.

Discussion:

The exploration and preparation of the dataset for breast cancer diagnosis showcased a meticulous process aimed at ensuring the data's quality and relevance for subsequent modeling phases. By employing widely used datasets, such as the Wisconsin Breast Cancer Database (WBCD), and following established guidelines for data science (Masters, 2020), the study laid a robust foundation for meaningful analysis.

The initial dataset, comprising various attributes such as radius, texture, perimeter, and more, underwent thorough preprocessing steps within the RStudio environment. This included rearranging features, removing unnecessary columns like the ID, and ensuring proper labeling of the diagnosis variable (Sultan, 2023). a thorough examination of the data structure, the normalization procedures, and a malignant or benign response variable equilibrium check provided essential insights into the characteristics of the dataset (Nwanganga, 2020).
One of the significant challenges addressed during the data preparation phase was the identification and handling of multicollinearity. The study recognized the importance of examining correlations among features to ensure the robustness of the subsequent machine-learning models (Sultan 2023). The use of correlation coefficients from Pearson and visual aids such as scatter graphs and correlation plots allowed for a complete understanding of feature correlations (Masters, 2020).

Principal Component Analysis (PCA), one of the dimensionality reduction approaches introduced, showed how to strategically address the dataset's high dimensionality (Kantardzic, 2020). The study effectively decreased the number of indicators while keeping a significant amount of the original variance by methodically converting the data set to a lower-dimensional subspace. The screeplot and biplots provided valuable insights into the importance of principal components, offering a roadmap for subsequent modeling steps (Sultan, 2015).

The discussion of results underscores the significance of these preparatory steps in shaping the subsequent phases of the study. The dimensionality reduction not only addresses computational challenges but also enhances the interpretability of the dataset, crucial for effective modeling (Kantardzic, 2020). The choice of PCA as a feature extraction method aligns with its suitability for datasets with redundant and correlated features, as identified through correlation checks (Nwanganga, 2020).

The examination of eigenvalues further emphasized the importance of each principal component in contributing to the dataset's variability (Masters, 2020). A clear grasp of the declining returns as a

 https://doi.org/10.59568/JASIC-2024-5-1-10

percentage of variance explained by other variables was made possible by the visual representation of component importance using scree-plots (Akampurira, 2022).

IV Conclusion

The researcher was able to produce data from the data preparation and exploration phase that can be accessible from any data modeling program, such as IBM SPSS, Stata, Excel, R, etc. Effective data cleaning allowed us to generate clean data free of unsalvageable things. By reorganizing and rescaling our features, we produced subsets of data that we used in the following steps. We were able to reduce our highly dimensional dataset of 30 predicting variables to 22 predicting variables with correlational removal to remove highly correlated features. We also managed to use PCA for feature extraction and reduced our data dimensionality by at least 24% maintaining the reliability of the predicted features. The resulting datasets were used in the next phase of model building and evaluation.

Acknowledgment: Several people contributed invaluable contributions that enabled this work to be completed. We would like to express our appreciation to Kareyo Margaret, Elly Gamukama, and Semalulu Paul for their committed work and assistance during the research process. Their knowledge and dedication significantly raised the caliber of this endeavor.

V References

[1]Abuassba, A. O. M. (2017), “Improving Classification Performance through an Advanced Ensemble Based Heterogeneous Extreme Learning Machines”.

[2] AfefRania, L., et al. (2018), “Comparison Study for Computer Assisted Detection and Diagnosis ‘CAD’ systems Dedicated to Prostate Cancer Detection Using MRImp Modalities”.

[3] Ahmad, L. G., et al.using three machine learning techniques for predicting. American Cancer Society. (2020). Cancer Facts and Figures, Atlanta, GA: American Cancer Society.

[4] American Cancer Society. (2020), “Cancer Facts and Figures. Atlanta, GA: American Cancer Society”.

[5] American Joint Committee on Cancer. (2017). Breast. In: AJCC Cancer Staging Manual. 8th Ed. New York, NY: Springer.

[6]Arunachalam, A. (2017). Combining Heterogeneous Ensemble Learners into a Single Meta-Learner in an Amateur Way.

[7] Asri, H. M. (2016). Using machine learning algorithms for breast cancer risk prediction and diagnosis.

[8] Bashir, S. Q. (2015). Heterogeneous classifiers fusion for dynamic breast cancer diagnosis using weighted vote-based ensemble. Black, improving early detection of breast cancer in sub-Saharan Africa: why mammography may not be the way forward.

[9] Bostock, M., et al. (2016). "Introducing Data Science: Big data, machine learning, and more, using Python tools."

[10] Bowles, M. (2015), “Machine Learning in Python Essential Techniques for Predictive Analysis”.

[11] Chhatwal, J., et al. (2010). Optimal Breast Biopsy Decision-Making Based on Mammographic Features and Demographic Factors.

[12] Cortes, C., et al. (1995). Support-vector networks. Machine Learning. Das, S. A. (2019). Big data in healthcare: management, analysis, and prospects.

[13] Dhahri, H., et al. (2019). Automated Breast Cancer Diagnosis Based on Machine Learning Algorithms.

[14] Elter, M. (2007). The prediction of breast cancer biopsy outcomes using two CAD approaches that both emphasize an intelligible decision process.

[15] Faure, C. A. (2017). Empirical and fully Bayesian approaches for the identification of vibration sources from transverse displacement measurements. Mechanical Systems and Signal Processing.

[16] Forsyth, A., et al. (2018). Machine learning methods to extract documentation of breast cancer symptoms from electronic health records. Frankenfield, J. (2020). An introduction to Machine learning.

[17] Hazra, A., et al. (2016). "Study and Analysis of Breast Cancer Cell Detection using Naïve Bayes, SVM and Ensemble Algorithms." International Journal of Computer Applications.

[18] Jemal, A., et al. (2011). Global cancer Statistics. Jordan, M. I., et al. (2015). Machine learning: Trends, perspectives, and prospects.

[19] Khairunnahar, L. H. (2019). Classification of malignant and benign tissue with logistic regression. Li, W., et al. (2017). Extraction of modal parameters for identification of time-varying systems using data-driven stochastic subspace identification. Journal of Vibration and Control.

[20] Liew, X. Y., et al. (2021). A Review of Computer-Aided Expert Systems for Breast Cancer Diagnosis.

[21] Mustafa, M., et al. (2016). Breast cancer: Detection markers, prognosis, and prevention. IOSR Journal of Dental and Medical sciences.

[22] Nematzadeh, Z., et al. (2015). Comparative studies on breast cancer classifications with k-fold cross validations using machine learning techniques. Noske, A. A. (2020). Risk stratification in luminal-type breast cancer: Comparison of Ki-67 with EndoPredict test results.

[23] Nwanganga, F., et al. (2020). Practical machine learning in R. Quinlan, J. (1996). Improved Use of Continuous Attributes in C4.5. Journal of Artificial Intelligence Research.

[24] Ricvan, D. N., et al. (2018). Diagnostic Accuracy of Different Machine Learning Algorithms for Breast Cancer Risk Calculation: a Meta-Analysis.

[25] Rokach, L. (2010). Ensemble-based classifiers. Salama, G. A. (2012). Breast cancer diagnosis on three different datasets using multi-classifiers. Generating concise and accurate classification rules for breast cancer diagnosis. Artificial Intelligence Medicine.

[26] Shen, R., et al. (2015). Intelligent breast cancer prediction model and clinical features: A comparative investigation in machine learning paradigm.

[27] Ting, F., et al. (2019). Convolutional neural network improvement for breast cancer classification. Expert Systems with Applications.

[28] Toğaçar, M., et al. (2020). Application of breast cancer diagnosis based on a combination of convolutional neural networks, ridge regression and linear discriminant analysis using invasive breast cancer images processed with autoencoders.

[29] Trieu, P. T. (2019). Improvement of cancer detection on mammograms via BREAST test sets. Uhlig, J., et al. (2019). Discriminating malignant and benign clinical T1 renal masses on computed tomography, A pragmatic radiomics and machine learning approach.

[30]Vogelstein, J. T., et al. (2021). Supervised dimensionality reduction for big data. Nat Commun 12, 2872. https://doi.org/10.1038/s41467-021-23102-2

[31] Vrigazova, B., et al. (2019). Optimization of the ANOVA procedure for support vector machines. International Journal of Recent Technology and Engineering.

[32] Wu, M., et al. (2019). Prediction of molecular subtypes of breast cancer using BI-RADS features based on a “white box” machine learning approach in a multi-modal imaging setting.

[33] Yaghoubi, V., et al. (2017). Automated Modal Parameter Estimation Using Correlation Analysis and Bootstrap Sampling. Mechanical Systems and Signal Processing.

[34] Zonno, G., et al. (2017). Laboratory evaluation of a fully automatic modal identification algorithm using an automatic hierarchical clustering approach. Procedia Engineering.