A Practical Investigation of Principal Component Analysis (June 2023)

Rijan Ghimire[1] and Sujan Bhattrai[1]

[1]Department of Electronics and Computer Engineering, IOE, Thapathali Campus, Kathmandu 44600, Nepal
Corresponding author: Rijan Ghimire (rijanghimire96@gmail.com)

ABSTRACT Principal Component Analysis (PCA) is a widely adopted technique for dimensionality reduction and data exploration. This paper provides a comprehensive overview of PCA implementation, emphasizing its key concepts and procedural steps. Beginning with centering the dataset, the process involves computing the covariance or correlation matrix, extracting eigenvectors and eigenvalues, selecting principal components, and projecting the data onto these components. By reducing the dimensionality of the dataset, PCA facilitates noise reduction, improved visualization, and enhanced analysis capabilities. By identifying underlying patterns and relationships in complex datasets, PCA empowers effective decision-making processes. This paper serves as a valuable resource for researchers and practitioners seeking to apply PCA in diverse domains such as image processing, data mining, and pattern recognition.

INDEX TERMS data exploration, dimensionality reduction, eigenvectors and eigenvalues, principal component analysis, visualization and analysis capabilities

I INTRODUCTION

Principal Component Analysis (PCA) [1] is a classical method widely adopted for dimensionality reduction. It offers a sequence of best linear approximations to high-dimensional observations. The algorithm for Principal Component Analysis (PCA) is based on finding orthogonal directions that explain the maximum variance in the data. In the context of dimensionality reduction, the objective is to find m orthonormal directions that minimize the representation error. Dimensionality reduction plays a crucial role in analyzing high-dimensional datasets and makes it easier to capture the essential variability within the data while reducing computational complexity and enhancing interpretability. Learning the ways of PCA will enable students to get familiar with dimensionality reduction techniques as we guide to build a groundwork for learning about implementing it in more complex algorithms

In the realm of high-dimensional datasets, it becomes increasingly challenging to analyze and interpret data effectively due to the curse of dimensionality. Dimensionality reduction techniques, such as Principal Component Analysis (PCA), offer a solution by extracting a reduced set of relevant features that capture the essential variability within the data. By reducing the dimensionality, computational complexity is reduced, and the interpretability of the data is enhanced. It offers a sequence of best linear approximations to high-dimensional observations. The algorithm for Principal Component Analysis (PCA) is based on finding orthogonal directions that explain the maximum variance in the data. In the context of dimensionality reduction, the objective is to find m orthonormal directions that minimize the representation error. Gaining knowledge in dimensionality reduction through PCA offers students a valuable understanding of the inner workings of this technique and its practical implications. The hands-on experience of working with real or simulated datasets enables students to delve into the step-by-step process of applying PCA, interpreting the outcomes, and extracting meaningful insights regarding the underlying patterns and structures of the data. By grasping the foundational concepts and techniques of PCA, students acquire a solid foundation that serves as a foundation for exploring more advanced dimensionality reduction algorithms and their practical implementations. This knowledge equips them with the skills necessary to tackle complex data analysis challenges and facilitates their understanding of how to optimize data representation while retaining critical information.

II RELATED WORK

In the field of dimensionality reduction and feature extraction algorithms, Principal Component Analysis (PCA) has been extensively studied. Jolliffe's book ”Principal Component Analysis” provides a comprehensive overview of PCA, encompassing its mathematical foundations, practical implementation, and diverse applications [1]. Shlens' tutorial paper ”A Tutorial on Principal Component Analysis” further delves into the intricacies of PCA, elucidating its algorithm, interpretation of principal components, and its efficiency in dimensionality reduction [2]. Additionally, Turk and Pentland’s paper ”Eigenfaces for recognition” explores the application of PCA for face recognition tasks, showcasing its effectiveness in extracting discriminative facial features [3]. These works collectively contribute to the understanding and evaluation of PCA as a powerful feature extraction technique in various domains.

III METHODOLOGY

A DATASET DESCRIPTION

The performance evaluation of feature extraction algorithms based on Principal Component Analysis (PCA) involved conducting experiments on three datasets. The first dataset was a randomly generated 20x2 array, which served as a synthetic dataset for assessing PCA’s effectiveness in a controlled environment. The second dataset was a crop dataset consisting of five variables: Nitrogen (N), Phosphorus (P), Potassium (K), temperature, and humidity. This dataset contained 400 observations and aimed to analyze PCA’s performance on real-world agricultural data. The 400 observations consist of the data of rice, maize, chickpea, and kidney beans 100 each. Lastly, the well-known scikit-learn iris dataset was used, which includes measurements of petal and sepal length for three types of irises (Setosa, Versicolour, and Virginica) in a 150x4 numpy.ndarray format. The purpose of applying PCA on this dataset was to evaluate its efficacy in feature extraction for iris classification tasks [4–6].

B PROPOSED METHODOLOGY

Principal Component Analysis (PCA) is a statistical technique used to transform a dataset’s features into a set of uncorrelated variables called principal components. These components are derived through an orthogonal transformation, where the first component captures the highest variance, the second component captures the second highest variance, and so on. In our study, the PCA algorithm was implemented from basic Python libraries to perform this transformation. The first step involves data preprocessing, where the dataset is loaded and irrelevant attributes are removed. To encode the crop labels, a mapping dictionary is created, enabling numerical representation. Next, the dataset is converted into a matrix form, facilitating further computations. Exploratory Data Analysis is conducted through a pair plot, visualizing attribute relationships. Standardization is performed by subtracting attribute means, and the resulting standardized dataset is stored. The covariance matrix is computed, revealing attribute relationships. Eigenvalues and eigenvectors of the covariance matrix are calculated and sorted. By multiplying the standardized dataset with the sorted eigenvectors, a change of basis is achieved. The proportion of variance explained by each eigenvalue is determined. Finally, PCA is applied to project the transformed data onto different combinations of principal components, generating scatter plots and 3D visualizations. Through this proposed method, the crop recommendation dataset can be effectively analyzed and visualized using PCA.

C MATHEMATICAL FORMULAE

Let’s assume we have a matrix or dataset with dimensions $m \times n$.
The formula for the mean (μ) is:

$$\mu = \frac{\sum x_{ij}}{m \times n} \tag{1}$$

Where x_{ij} represents the element at the i-th row and j-th column of the matrix.
The formula for the covariance matrix Σ is:

$$\Sigma = \frac{1}{m-1} \cdot (X^T \cdot X) \tag{2}$$

For a square matrix A, the eigenvalues can be obtained by solving the characteristic equation:

$$|A - \lambda I| = 0 \tag{3}$$

Where λ is the eigenvalue and I is the identity matrix of the same size as A.
Given its eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$, the eigenvectors can be found by solving the equation:

$$(A - \lambda_i I)v_i = 0 \tag{4}$$

The proportion of variance (PoV) for each eigenvalue λ_i is given by:

$$\text{PoV}_i = \frac{\lambda_i}{\sum_{j=1}^{n} \lambda_j} \tag{5}$$

Where $\sum_{j=1}^{n} \lambda_j$ represents the sum of all eigenvalues.

D INSTRUMENTATION DETAILS

The experimentation was conducted using **Jupyter Notebook**, a popular interactive coding environment. The code utilized several essential libraries for data analysis and visualization. The **numpy** library was imported to handle numerical operations and manipulate arrays efficiently. The **sklearn.datasets** module was utilized to import the Iris dataset, a commonly used dataset for machine learning tasks. The **matplotlib.pyplot** library allowed for the creation of various plots and visualizations. The **mpltoolkits.mplot3d** module was imported specifically to enable 3D plotting. Finally, the **itertools** library was used to generate combinations for certain analysis tasks. These libraries played a crucial role in performing data analysis and facilitating comprehensive experimentation in the Jupyter Notebook environment.

E SYSTEM BLOCK DIAGRAM

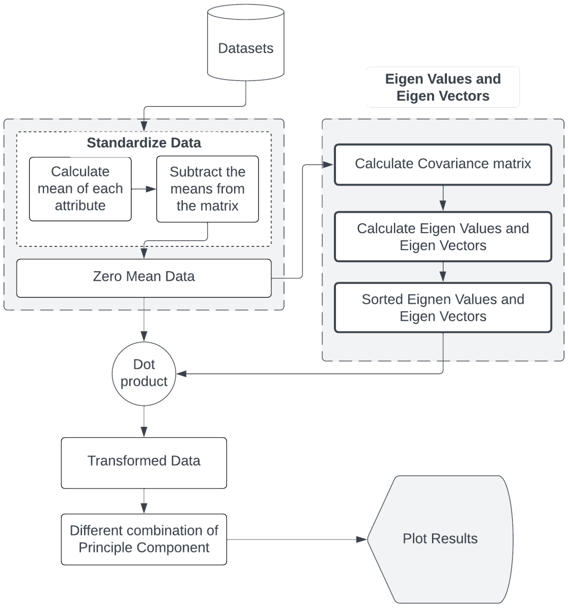
Datasets
Eigen Values and Eigen Vectors
Standardize Data
Calculate mean of each attribute
Subtract the means from the matrix
Zero Mean Data
Calculate Covariance matrix
Calculate Eigen Values and Eigen Vectors
Sorted Eignen Values and Eigen Vectors
Dot product
Transformed Data
Different combination of Principle Component
Plot Results

Figure 1: Block Diagram for Principal Component Analysis

IV EXPERIMENTAL RESULTS

A PROBLEM 1: PCA ON RANDOM DATA

PCA on Random Data showed the effectiveness of Principal Component Analysis (PCA) in reducing the dimensionality of random data. Several 2D plots were generated to visualize the data before and after PCA.

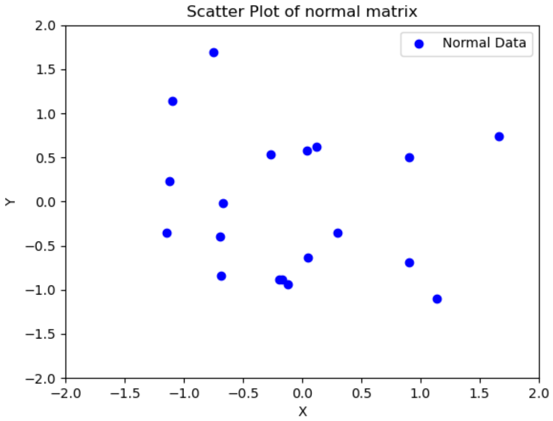

Figure 2: Plot of Normal(Gaussian) data

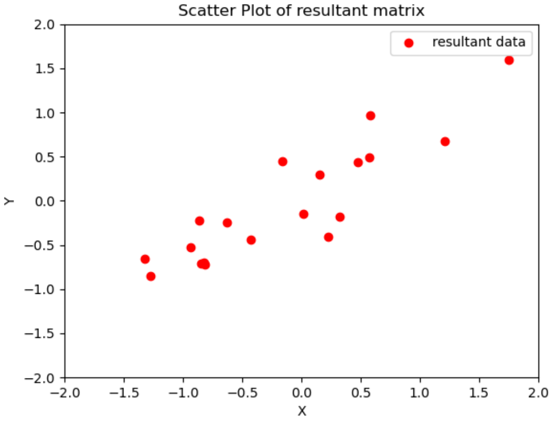

Figure 3: Plot of Resultant data

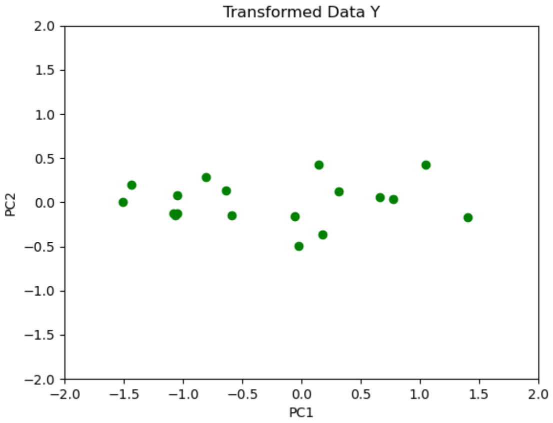

Figure 4: Plot of Transformed data

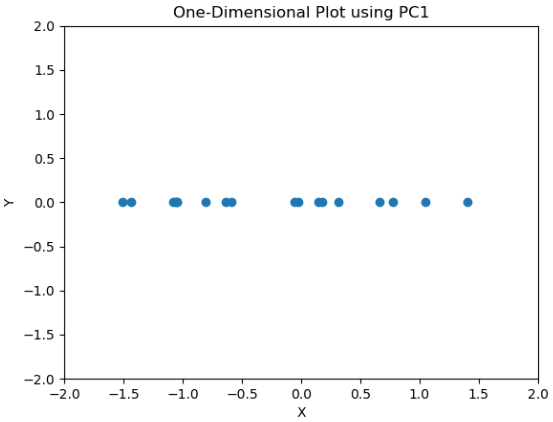

Figure 5: Plot using PC1

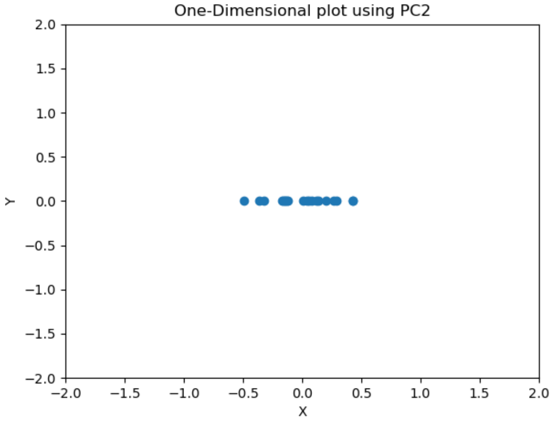

Figure 6: Plot using PC2

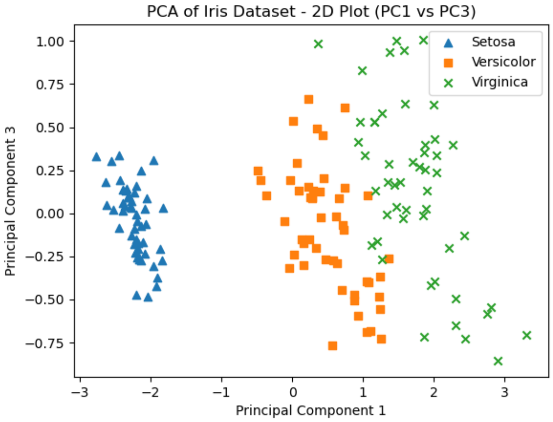

Figure 8: Plot using PC1 and PC3

B PROBLEM 2: PCA ON IRIS DATASET

PCA on the Iris Dataset demonstrated the efficiency of Principal Component Analysis (PCA) in analyzing and visualizing complex datasets. Various 2D and 3D plots were generated to illustrate the impact of PCA on the Iris dataset.

1 2D PLOTS OF IRIS DATASET

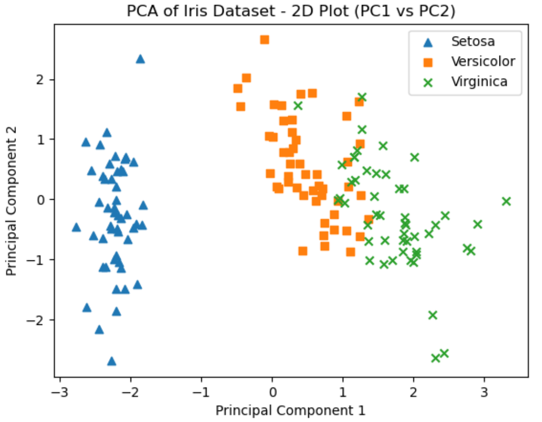

Figure 7: Plot using PC1 and PC2

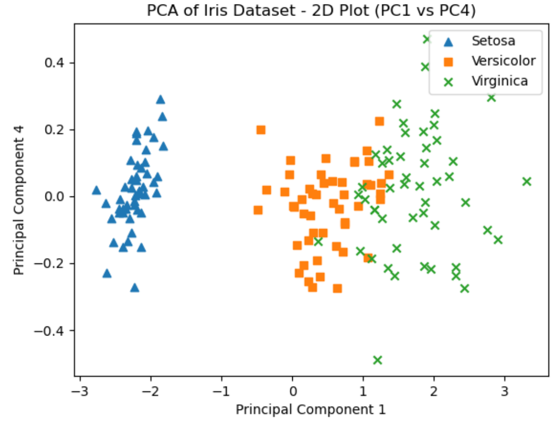

Figure 9: Plot using PC1 and PC4

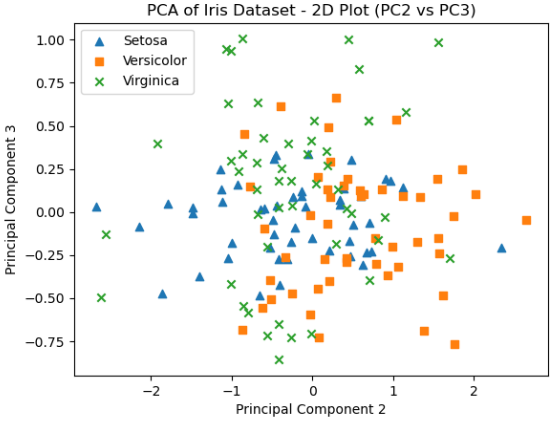

Figure 10: Plot using PC2 and PC3

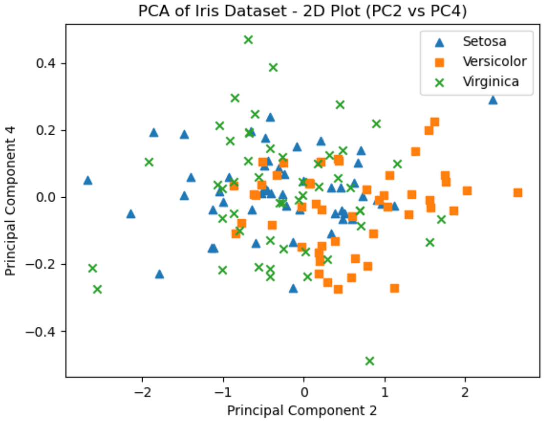

Figure 11: Plot using PC2 and PC4

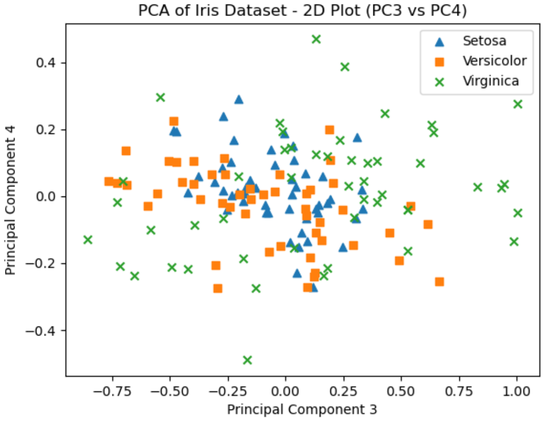

Figure 12: Plot using PC3 and PC4

2 3D PLOTS OF IRIS DATASET

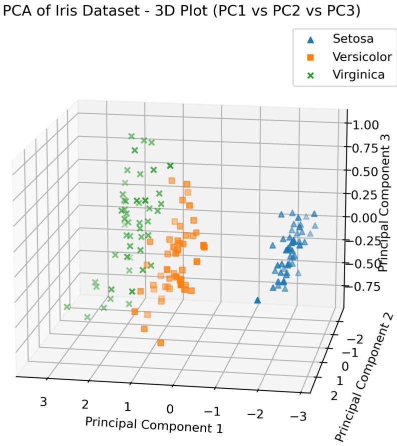

Figure 13: 3D Plot using PC1, PC2 and PC3

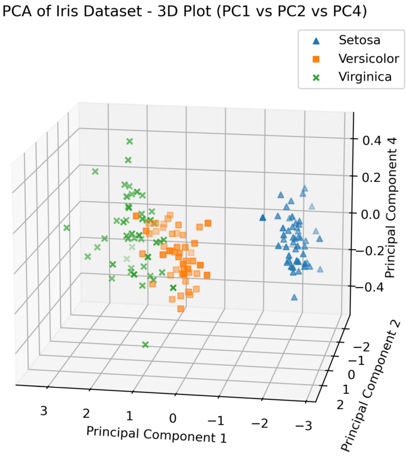

Figure 14: 3D Plot using PC1, PC2 and PC4

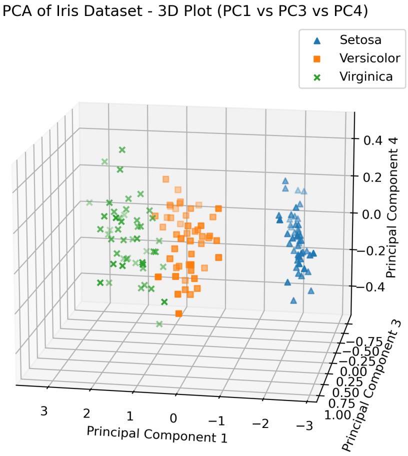

Figure 15: 3D Plot using PC1 ,PC3 and PC4

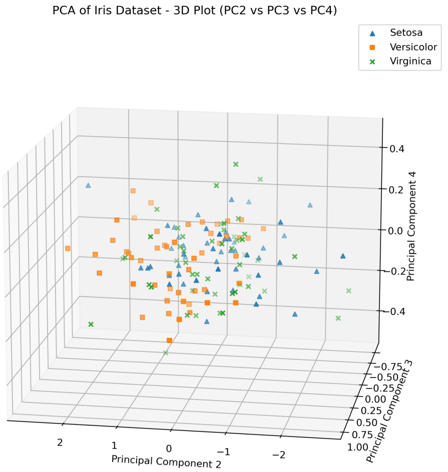

Figure 16: 3D Plot using PC2, PC3 and PC4

C PROBLEM 3: PCA ON CROP DATASET

PCA on the Crop Dataset revealed valuable insights into the relationship between different crop types based on their attributes. The dataset consisted of four crop types, namely rice, maize, chickpeas, and kidney beans, each having 100 data points. Various 2D and 3D plots were generated to illustrate the impact of PCA on the Crop dataset.

1 PAIRPLOT OF CROP DATASET BEFORE PERFORMING PCA

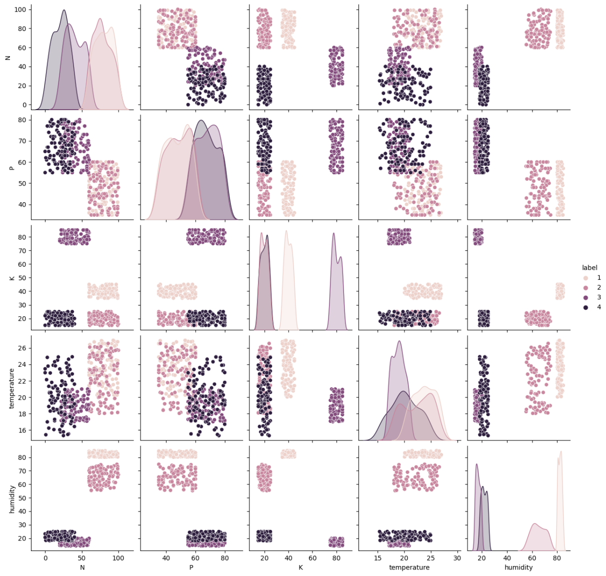

Figure 17: Sns Pairplot

2 2D PLOTS OF CROP DATASET

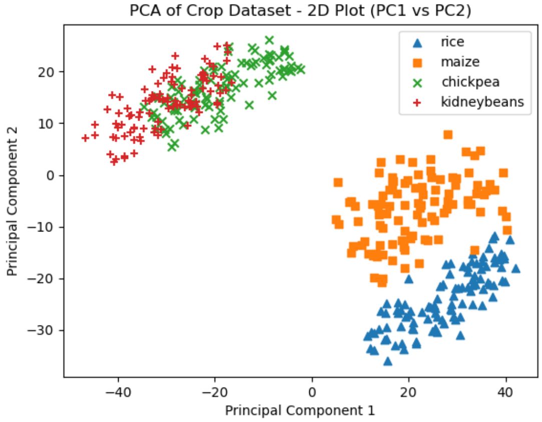

Figure 18: Plot using PC1 and PC2

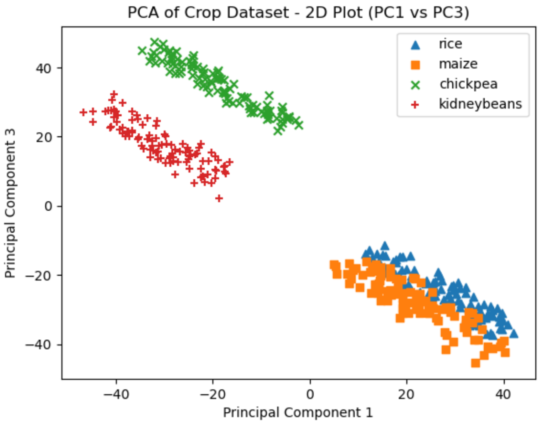

Figure 19: Plot using PC1 and PC3

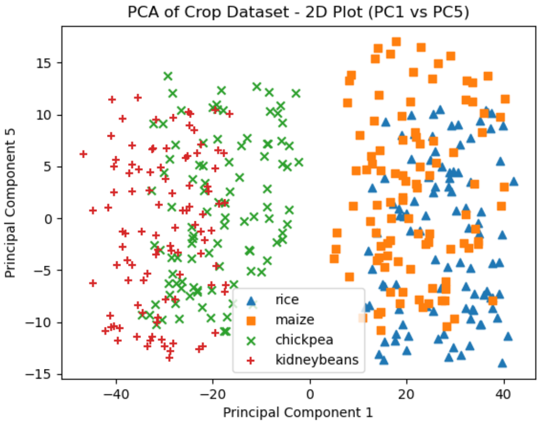

Figure 20: Plot using PC1 and PC5

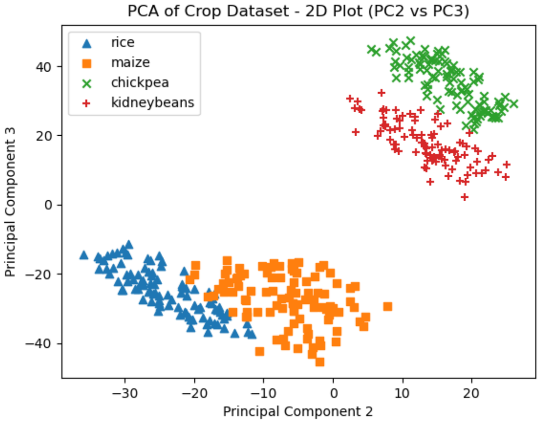

Figure 21: Plot using PC2 and PC3

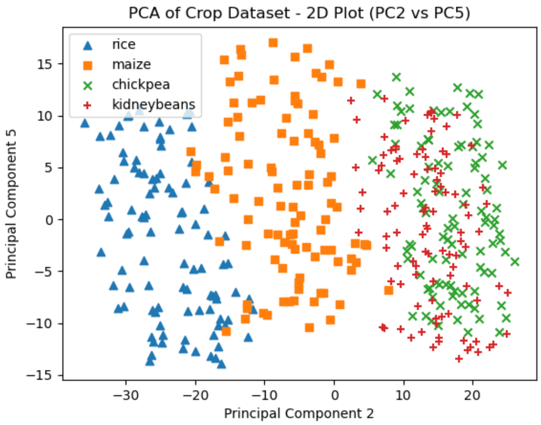

Figure 22: Plot using PC2 and PC5

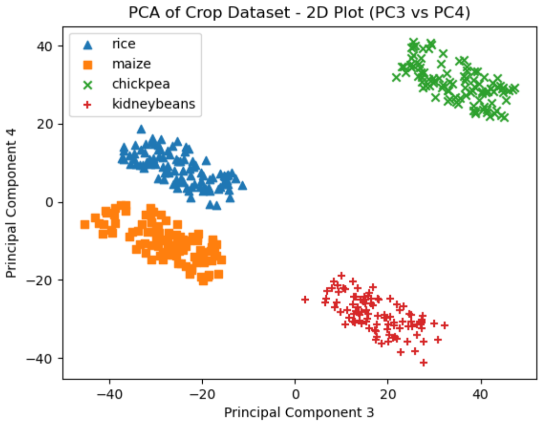

Figure 23: Plot using PC3 and PC4

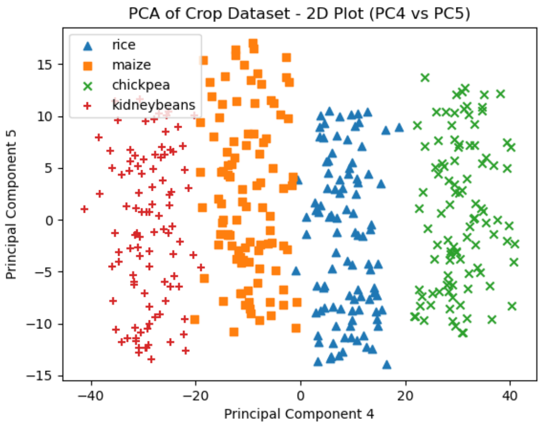

Figure 24: Plot using PC4 and PC5

3 3D PLOTS OF CROP DATASET

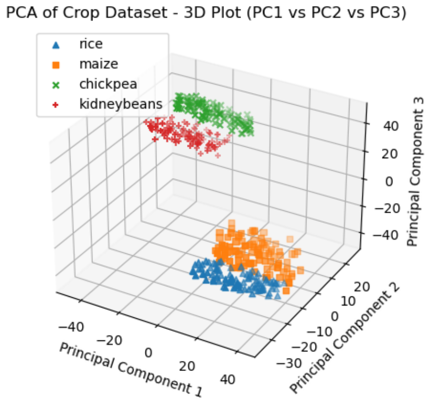

Figure 25: 3D Plot using PC1 , PC2 and PC3

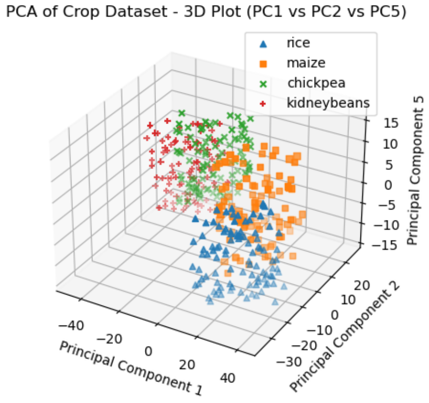

Figure 26: 3D Plot using PC1, PC2 and PC5

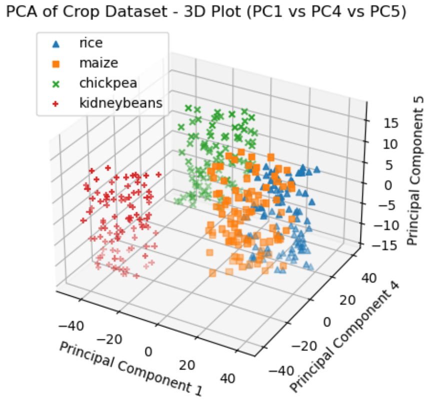

Figure 27: 3D Plot using PC1, PC4 and PC5

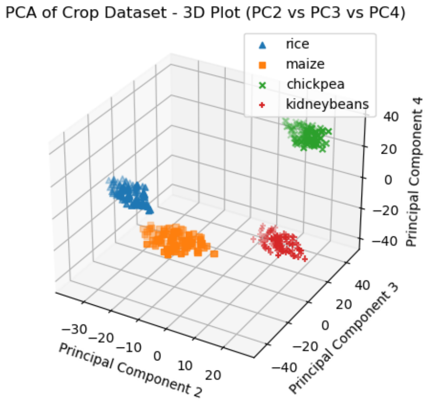

Figure 28: 3D Plot using PC2, PC3 and PC4

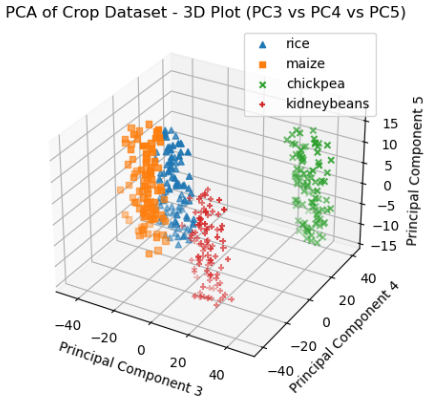

Figure 29: 3D Plot using PC2, PC3 and PC4

V DISCUSSION AND ANALYSIS

In IV A initially, the random data points were scattered across the plot with no apparent structure. However, after applying PCA, the data points were transformed into a new coordinate system aligned with the principal components.

In IV B initially, the dataset contained multiple features, making it difficult to visualize patterns. By applying PCA, the data was transformed into a lower-dimensional space, preserving the most significant variations. This transformation allowed for more accessible visualization. The 2D plots showed the distribution of the data points in reduced dimensions, revealing clusters and patterns that were not apparent in the original dataset. The 3D plots provided an even more comprehensive representation of the data, allowing for the examination of relationships between three principal components.

Finally from IV C we find by applying PCA and generating 2D and 3D plots, the data's inherent patterns and clusters were visually represented. The plots showcased the separation and grouping of crop types based on their attribute similarities. In the 2D plots, distinct clusters corresponding to each crop type were observed, indicating the presence of characteristic attributes unique to each crop. The 3D plots provided a more comprehensive view, enabling the examination of interrelationships among multiple attributes.

VI CONCLUSION

Based on the results and analysis presented in the previous sections, we can draw several conclusions. Firstly, Principal Component Analysis (PCA) demonstrated its effectiveness in feature extraction and dimensionality reduction across all three datasets. In Problem 1, PCA was able to capture the underlying structure of the randomly generated data, showcasing its ability to reveal patterns in synthetic datasets. In Problem 2, PCA successfully reduced the dimensionality of the iris dataset while preserving its discriminatory information, as evidenced by the 2D and 3D plots. Highlights PCA's use in facilitating visualization and potentially improving classification tasks. In Problem 3, PCA showcased its capability in extracting essential features from the crop dataset, enabling meaningful representations and potentially aiding in further analysis.

The practical implementations confirmed that Principal Component Analysis (PCA) is a valuable tool for feature extraction, providing insights into the intrinsic structure of data and reducing its dimensionality. Despite its limitations, such as linearity assumptions and sensitivity to outliers, PCA showcased its potential in various domains including data analysis, and pattern recognition. Future research focusing on alternative feature extraction techniques and larger datasets can further enhance our understanding and applicability of PCA, contributing to our overall learning experience in the field.

REFERENCES

[1] I. T. Jolliffe, *Principal component analysis for special types of data*. Springer, 2002. [Online]. Available: https://link.springer.com/chapter/10.1007/0-387-22440-8_13

[2] J. Shlens, "A tutorial on principal component analysis," *arXiv preprint arXiv:1404.1100*, 2014. [Online]. Available: https://arxiv.org/abs/1404.1100

[3] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of cognitive neuroscience*, vol. 3, no. 1, pp. 71–86, 1991. [Online]. Available: https://direct.mit.edu/jocn/article/3/1/71/3025/Eigenfaces-for-Recognition

[4] E. Anderson, "The species problem in iris," *Annals of the Missouri Botanical Garden*, vol. 23, no. 3,

pp. 457–509, 1936. [Online]. Available: https://www.jstor.org/stable/2394164

[5] A. Chandramouli and G. Deepak, "Ontofusioncrop: An ontology centric approach for crop recommendation based on bagging and semantic alignment," in *Digital Technologies and Applications: Proceedings of ICDTA'22, Fez, Morocco, Volume 2*. Springer, 2022, pp. 210–219. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-031-02447-4_22

[6] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of eugenics*, vol. 7, no. 2, pp. 179–188, 1936. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1111/j.1469-1809.1936.tb02137.x

Rijan Ghimire is currently pursuing his undergraduate degree in Electronics, Communication, and Information at IOE, Thapathali Campus. His research interests encompass various areas, including data mining, network communications, and optimization theory and technology.(THA076BEI022)

Sujan Bhattrai is currently pursuing his undergraduate degree in Electronics, Communication, and Information at IOE, Thapathali Campus. His research interests cover various areas, including data mining, deep learning, operating system, robotics, and power electronics.(THA076BEI037)