American Journal of Engineering Research (AJER)
e-ISSN: 2320-0847 p-ISSN : 2320-0936
Volume-14, Issue-2, pp-91-97
www.ajer.org
Research Paper Open Access

Using PCA for Effective Classification and Visualization of Antimicrobial Peptides

Sophio Barnovi
Assistant Professor of Georgian Technical University
Ivane Makasarashvili
Associate Professor of Georgian Technical University

ABSTRACT:*Antibiotic resistance poses a significant challenge for antimicrobial therapies, prompting interest in antimicrobial peptides (AMPs) as a promising alternative. AMPs are short molecules found in living organisms that serve as a defense mechanism against pathogenic microorganisms.Given the inherent difficulty bacteria face when developing resistance to AMPs, these peptides are considered a viable substitute for traditional antibiotics. This paper explores the application of Principal Component Analysis (PCA) in the classification of AMPs. The study aims to identify AMPs by analyzing their structure and activity. The primary objective is to reduce the dimensionality of complex feature spaces while retaining maximum information. This study examines the methodological benefits of PCA, including its ability to uncover hidden patterns, facilitate data compression, and enhance data visualization. In this study, AMP features were reduced to two- and three-dimensional spaces, enabling the preservation of the majority of the dataset's variance. The visualization results demonstrated a distinct separation between classes, underscoring the effectiveness of PCA in supporting classification tasks. This research provides a crucial foundation for future studies aimed at advancing molecular modeling techniques related to AMP resistance against antibiotics. The use of PCA can significantly enhance the computational efficiency of machine learning models in classification tasks.*
KEYWORDS *antimicrobial peptides (AMP), Principal Component Analysis (PCA), classification, machine learning.*

Date of Submission: 12-02-2025 Date of acceptance: 24-02-2025

I. INTRODUCTION

Reducing the number of variables in a dataset inevitably involves a trade-off with accuracy. However, smaller datasets are easier to study and visualize, enabling faster data analysis without imposing unnecessary computational demands on machine learning algorithms.

Principal Component Analysis (PCA) is a robust statistical method for analyzing and interpreting complex data structures. Historically, the primary limitation of PCA's application was the computational difficulty associated with its implementation. Before advancements in computer technology, large-scale data analysis was virtually infeasible [1, 2].

The development of digital technologies has significantly enhanced the practical utility of PCA and other statistical analysis methods. In many real-world scenarios, datasets often contain a large number of parameters, making optimal processing and effective decision-making both challenging and essential.

In such cases, information compression and compact representation become critical. PCA addresses this need by reducing data dimensionality and uncovering underlying patterns while retaining as much informational content as possible.

This method finds extensive application across diverse scientific and practical domains, including image processing, bioinformatics, economics, finance, and sociological and psychological research. Its primary objective is to reduce multidimensional data to a smaller set of components, preserving the original structural and informational characteristics.

AMPs are diverse molecules with varying physicochemical properties. To classify these peptides, researchers often compute various features such as hydrophobicity, charge, and molecular weight. PCA is then applied to reduce the dimensionality of these feature sets, enabling the identification of patterns that distinguish AMPs from non-AMPs. For instance, in a study focusing on linear cationic AMPs active against Gram-negative

and Gram-positive bacteria, PCA was utilized to map high-dimensional data into lower-dimensional spaces, aiding in the visualization and identification of outliers. Beyond classification, PCA serves as a powerful tool for visualizing the distribution of peptide properties. By projecting high-dimensional data onto principal components, researchers can generate 2D or 3D plots that reveal clusters and relationships within the data. This visualization aids in understanding the variability among peptides and can highlight distinct groups based on their physicochemical characteristics.

II. ADVANTAGES OF THE PRINCIPAL COMPONENT METHOD

Principal Component Analysis (PCA) is a versatile data processing technique that simplifies and enhances data analysis by compressing datasets while retaining critical information. The primary goals and advantages of PCA include the following:

Revealing Hidden Patterns

PCA enables the identification of data features that are not immediately apparent through conventional measurements. This capability allows researchers to uncover the fundamental structure of a process and identify the primary sources of data variability.

Describing Processes with Fewer Variables

PCA reduces high-dimensional datasets to a smaller number of principal components, preserving the most significant information within the data.

- **Data Compression:** By reducing dataset size, PCA facilitates simplified modeling. The reduced components often capture the essence of the process more effectively than the original variables.

Analyzing the Interdependence of Variables

PCA helps identify statistical relationships between variables and principal components. This insight enhances understanding of how specific features influence the process. Such analysis is crucial for optimizing processes and making informed decisions.

Enhancing Prediction Models

PCA supports the development of more accurate predictive models.

- **Regression Models:** Regression models based on principal components are particularly effective in cases where multicollinearity among variables poses challenges. PCA simplifies these models, improving both reliability and accuracy.
- PCA contributes to the simplicity and robustness of models, which is especially critical in the context of complex datasets.

Practical Importance of PCA in Classification Tasks

PCA serves as a vital preprocessing tool in classification problems, offering several benefits:

- **Increased Classification Efficiency:** PCA reduces data complexity and emphasizes the primary sources of variation, making classification algorithms more efficient.
- **Improved Data Visualization:** PCA projects data into two or three dimensions, simplifying representation and facilitating a clearer distinction between classes.
- **Noise Reduction:** By filtering out less significant components, PCA minimizes noise and enhances data quality.
- **Prevention of Overfitting:** PCA reduces the number of variables, enabling models to generalize more effectively and avoid overfitting.

In summary, PCA is an ideal method for data analysis, particularly when simplicity, accuracy, and efficiency are paramount. It is widely applicable across various domains and proves indispensable for uncovering insights in complex datasets.

III. PRINCIPAL COMPONENT ANALYSIS IN THE AMP CLASSIFICATION PROBLEM

Principal Component Analysis (PCA) is an unsupervised learning method that transforms data without relying on class labels. Its primary objective is to identify the principal directions of variation in the data, thereby reducing dimensionality and increasing simplicity.

Visualizing PCA results often yields valuable insights, particularly regarding class separation. For instance, when data is reduced to two dimensions and displayed graphically, the separation of classes can be assessed based on how distinctly they are represented in the principal component space. After PCA transformation, classes are often more clearly delineated within the reduced-dimensional space.

In the specific case of antimicrobial peptides (AMPs) and non-antimicrobial peptides (NAMPs), PCA can redistribute NAMPs, highlighting their dominance in the dataset. This redistribution is particularly useful for developing models that address imbalanced class distributions [5].

In our study, dimensionality reduction using PCA was applied to condense the features to 12 principal components for analysis and visualization. The PCA algorithm evaluates the features and reduces them to a specified number of dimensions. By setting the number of components (n_components) to 2 or 3, the algorithm transforms the data and returns an array with the same number of rows but only two or three columns, representing the reduced dimensions.

Heatmaps and Correlation Analysis

Heatmaps are an effective tool for visualizing data, providing a concise summary of relationships between variables. Specifically, heatmaps display correlationsa statistical measure of the strength and direction of the relationship between two or more variables.

Heat maps play a crucial role in machine learning by offering a clear and intuitive visualization of complex, multidimensional data. Through color-coded representations of numerical values, heat maps effectively highlight trends, variations, correlations, and distributions that might otherwise remain unnoticed in tables or conventional charts.

Figure 1 presents a correlation matrix, which illustrates the correlation coefficients between variables. The correlation coefficient quantifies the degree of association between two variables, offering insight into their interdependence. Such visualization aids in identifying patterns and understanding the underlying relationships in the data.

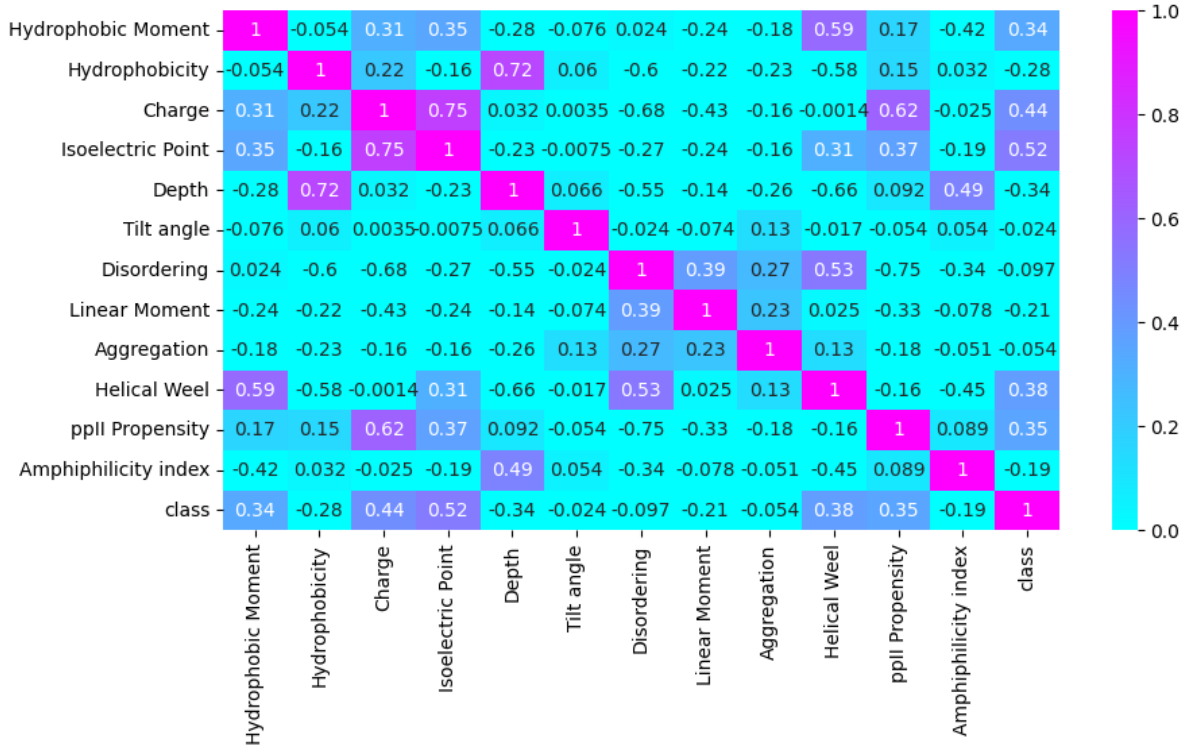

Fig.1. Heatmap illustrating feature correlations

Before applying Principal Component Analysis (PCA), the data is standardized to ensure each feature has unit variance. This is achieved using the StandardScaler preprocessing tool. The scaling process involves calling the fit and transform methods, which compute the required parameters and apply the transformation, respectively.

Learning and applying the PCA transformation is straightforward and follows a similar process to other preprocessing techniques. A PCA object is created, and the principal components are determined by invoking the fit method. To reduce the dimensionality of the data, the desired number of components must be specified when initializing the PCA object.

In this analysis, the data features are reduced to two and three dimensions. The PCA coefficients are then visualized on a heatmap, providing an intuitive representation of the relationships between components.

This is a reasonable compression ratio, and you can see how this size reduction can speed up a classification algorithm. results from the above system are given in graphical forms. Some results are shown in Fig. 1.

Now that we have reduced the original dataset to just two dimensions, let's use a scatter plot to display the data. In this case, we will use Seaborn's scatterplot() function. First, we will convert the reduced data into a DataFrame and add a species column, which will determine the colors of the points [19, 15]. The results are visualized in Figures 2–6.

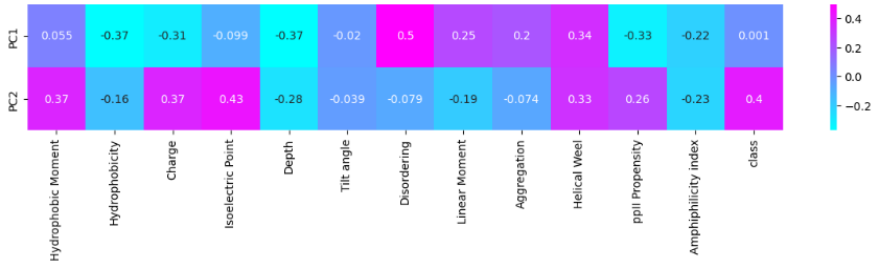

Fig. 2. The graph illustrates the data projected onto the first two principal components

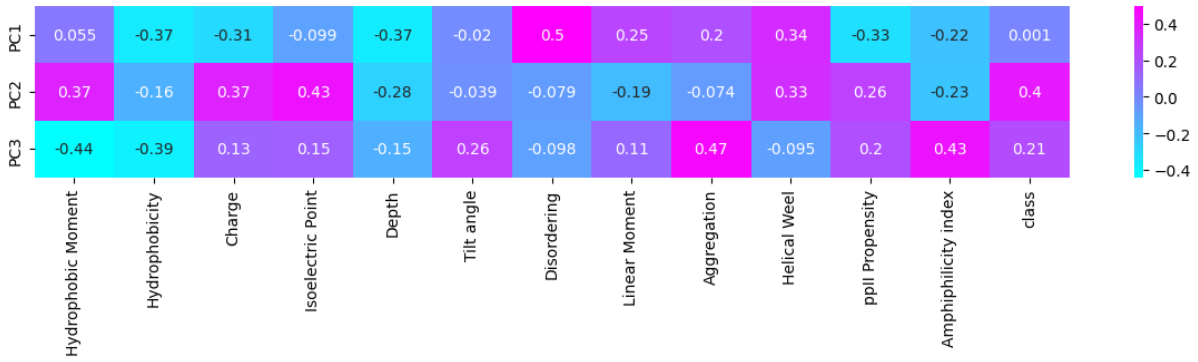

Fig. 3.The graph illustrates the data projected onto the first three principal components

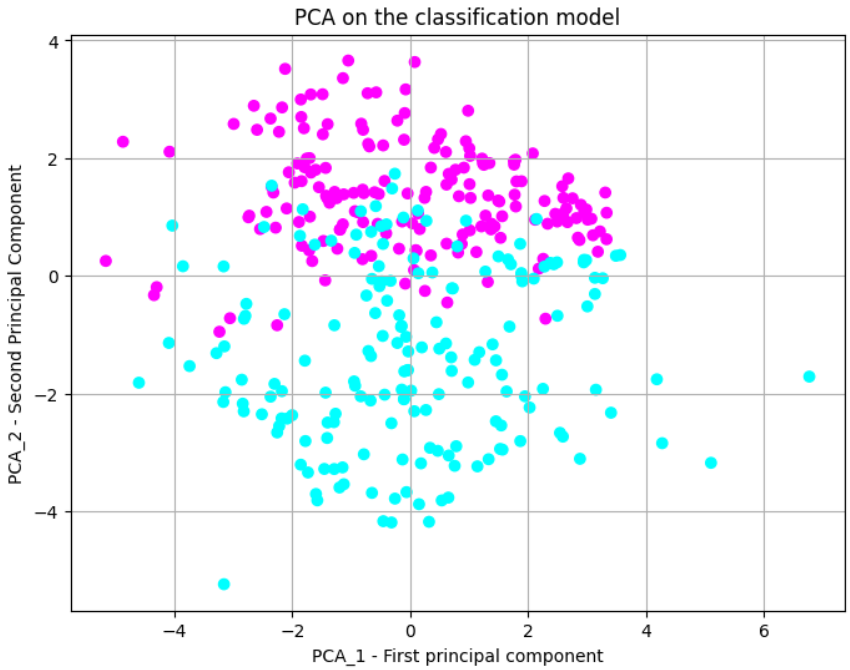

Fig. 4. Two-dimensional scatter plot of the peptide data set with the first two principalcomponents

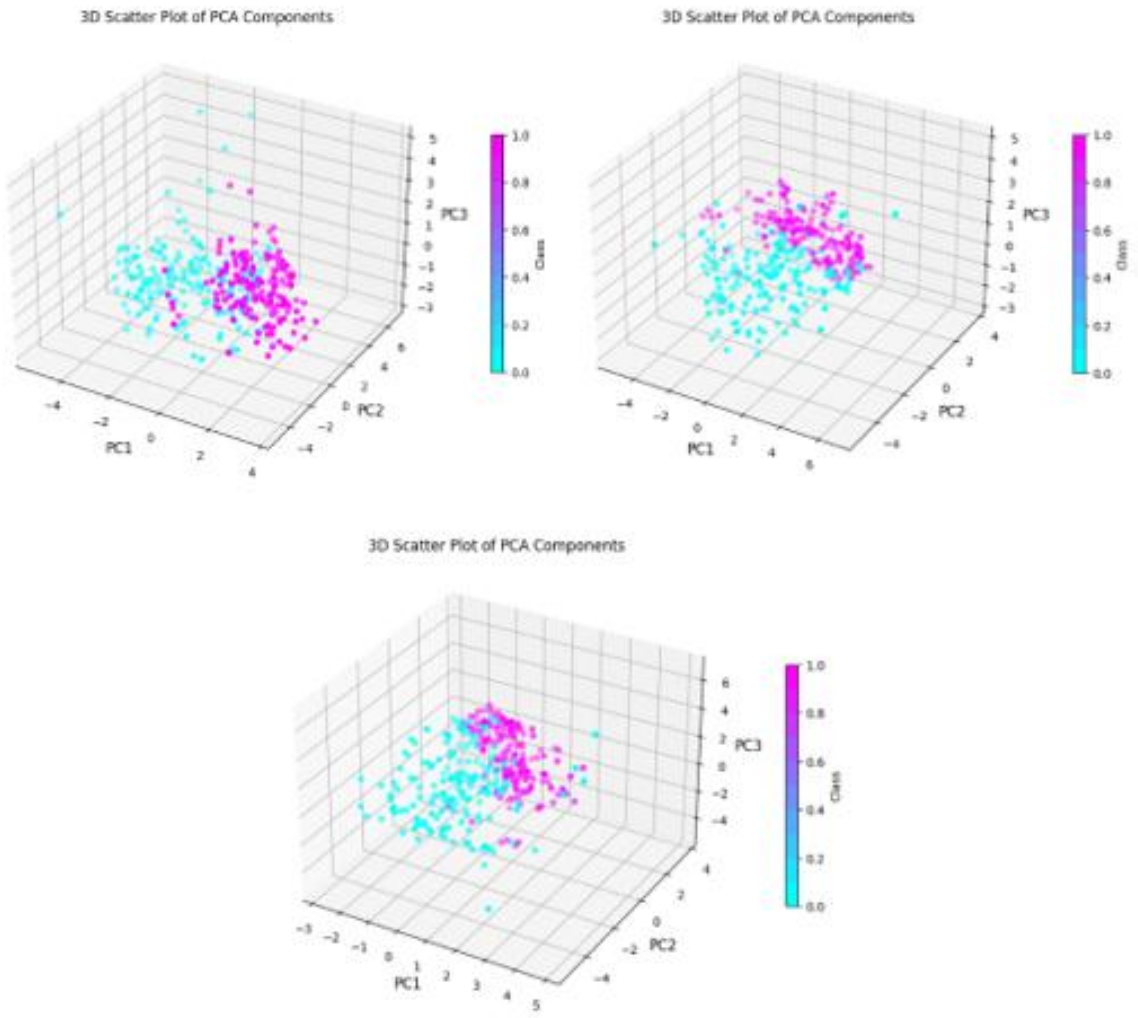

Fig. 5.Three-dimensional scatter plot of the peptide dataset, visualized using the first two principal components

It is important to emphasize that PCA is an unsupervised learning method, meaning it does not utilize class information when identifying the principal components. Instead, it relies solely on the correlations within the data. The figure demonstrates that in both the two-dimensional and three-dimensional spaces, the two classes are relatively well-separated, with only a few overlapping points.

When selecting principal components, it is advisable to prioritize axes that preserve the maximum variance, as this minimizes information loss.

The variance coefficients for the first two and three components are as follows:

- For the first two components: array ([0.28161919, 0.26032414]).
- For the first three components: array ([0.28161919, 0.26032414, 0.09579407]).

These results indicate that the first two components account for approximately **54%** of the dataset's variance, while the first three components account for approximately **63%**.

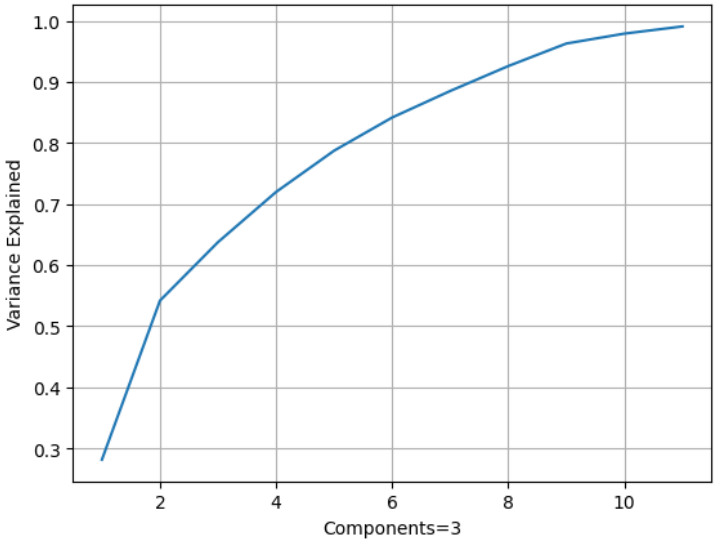

Fig. 6. Explained variance as a function of the number of dimensions

IV. CONCLUSION

The application of PCA in AMP research provides a robust framework for simplifying complex datasets, improving classification accuracy, and facilitating meaningful visualizations. By capturing the most significant variations in peptide properties, PCA aids in the effective analysis and interpretation of data, thereby advancing the understanding and development of antimicrobial peptides. Principal Component Analysis (PCA) is a powerful technique for analyzing and reducing the dimensionality of data while retaining the most significant information. This study demonstrated that PCA is an effective method for handling high-dimensional data in classification tasks. Key findings include:

1. PCA successfully reduced 12 features to 2 and 3 dimensions, retaining approximately 54% and 63% of the data variance, respectively.
2. The 2D and 3D visualizations revealed that the classes (AMP and NAMP peptides) are well separated in the reduced spaces, with only minor overlaps observed.
3. Correlation analysis highlighted significant relationships between features, offering deeper insights into the data structure.
4. The unsupervised nature of PCA, which does not rely on class labels, proved effective in uncovering the natural structure of the data.

These findings underscore the importance of PCA as a data preprocessing tool. It enhances the efficiency of classification models, simplifies data interpretation, and provides a robust framework for analyzing complex, high-dimensional datasets.

REFERENCES

[1]. Barnovi S., Ckhaidze M., Mchedlishvili N., Recognition of Antimicrobial Peptides by Neural Networks, American Journal of Engineering Research (AJER), vol. 10(5), 2021, pp. 163-169.
[2]. Vishnepolsky B., Pirtskhalava M. Prediction of Linear Antimicrobial Peptides Based on Characteristics Responsible for Their Interaction with the Membranes. Journal of Chemical Information and Modeling, 2014, 54, 5, 1512–1523.
[3]. Lee Y., Lee W., Fulan M., Ferguson L., Wong GCL. What can machine learning do for antimicrobial peptides, and what can antimicrobial peptides do for machine learning? Interface Focus 7, 2017.
[4]. Fang C., Moriwaki Y., Li Caihong L., Shimizu K. prediction of Antifungal peptides by Deep Learning with Character Embedding. IPSJ Transactions on Bioinformatics Vol.12, 21-29, 2019.
[5]. Hamid M., Friedberg I., Identifying antimicrobial peptides using word embedding with deep recurrent neural networks. Bioinformatics, Vol. 35, Issue 12, 2019, 2009-2016 p.
[6]. Veltri D., Kamath U,Shehu A.Deep learning improves antimicrobial peptide recognition. Bioinformatics, Vol 34, Issue 16, 2018, 2740-2747 p.
[7]. Youmans M., Spainhour C., Qui P. Long short term memory recurrent neural networks for antibacterial peptide identification, IEEE,multi 2017.
[8]. Guangshun Wang, Antimicrobial Peptides, Boston, USA, 2017.
[9]. Su X., Xu J., yin Y., Quan X. Antimicrobial peptide identification multi-scale convolutional network. BMC, 2019.
[10]. Chuncai Z,Chuncai Z. Design, Synthesis and Applications of Antimicrobial Peptides and Antimicrobial Peptide-Mimetic Copolymers[J]. Progress in Chemistry, 2018, 30(7): 913-920.
[11]. Wang G., Li X., Wang Z. The Antimicrobial peptides database as a tool for research and education. Nucleic Acids Research, 2015, Vol. 44, pp 1087-1093.

[12]. Gogoladze G., Grigolava M., Vishnepolsky B., Chubinidze M., Duroux P., Lefranc M., Pirtskhalava M. DBAASP: Database of Antimicrobial Activity and Structure of Peptides. FEMS MicrobioL Lett, 2014.
[13]. Zhang L., Gallo R. Antimicrobial peptides. Current Biology 26, 2016, R1–R21.
[14]. Wang S., Zeng X., Yang Q., Qiao S. Antimicrobial Peptides as Potential Alternatives to Antibiotics in Food Animal Industry. Int Mol Sci, V.17(5), 2016.
[15]. Géron A. Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly Media, Inc, 2019, 799p.
[16]. Seaung Lok Ham, Boosted-PCA for binary classification problems, 2024.
[17]. A reviewm, S. Aslam, T. Rabie, Principal Component Analysis in Image Classification, 2023.
[18]. Sidharth Prasad Mishra, Uttam Sarkar, Subhash Taraphder, Sanjay Datta, Devi, Prasanna Swain, Reshma Saikhom, Sasmita Panda and Menalsh Laishram
[19]. Paul J. Deitel, Harvey Deitel - Intro to Python for Computer Science and Data Science_ Learning to Program with AI, Big Data and The Cloud-Pearson, 2020.
[20]. Perry Xiao - Artificial Intelligence Programming with Python from Zero to Hero-John WIley, 2022.
[21]. Sean R McWhinney, Jaroslav Hlinka, Eduard Bakstein Principal Component Analysis as an Efficient Method for Capturing Multivariate Brain Signatures of Complex Disorders—ENIGMA Study in People with Bipolar Disorders and Obesity, 2024.
[22]. Jon Ahlinder, David Hall, Mari Suontama, Mikko J Sillanpää, Principal Component Analysis Revisited: Fast Multitrait Genetic Analysis Using PC, 2023.
[23]. Geonseok Lee, Tianhui Wang, Dohyun Kim, Myong Kee Jeong, Sparse Group Principal Component Analysis Using Elastic-Net Regularization, 2024.
[24]. Peijun Sang, Dehan Kong, Shu Yang, Functional Principal Component Analysis with Informative Covariates, 2023.
[25]. Guo Li, Yi Qin, An Exploration of the Application of Principal Component Analysis in Big Data Processing, 2023.