Journal of Informatics, Computer Science, Data Science And Arificial Intelligence
Volume 1 No. 1 December 2023, ISSN : 3031-9145

JICDA
ISSN : 3031-9145
Journal of Informatics, Computer Science, Data Science, and Artificial Intelligence

PCA IMPACT ON SVM PERFORMANCE: VARIABLE SIGNIFICANCE IN QUESTIONNAIRE DATA

Muhammad Mizan Siregar [1], Irwan Daniel [2], Agus Fahmi Limas Ptr [3]
[1]Universitas Deli Sumatera, mizan.siregar1@gmail.com.
[2]Universitas Deli Sumatera, irwandaniel@gmail.com.
[3]Universitas Deli Sumatera, agusfahmilimasptr@gmail.com.

ARTICLE INFO

ARTICLE HISTORY:
Received :13/12/2023
Revised : 16/12/2023
Accepted : 20/12/2023

Keywords:
Classification
Customer satisfaction analysis
Feature reduction
Principal Component Analysis
Support Vector Machine

ABSTRACT

This research explores the integration of Principal Component Analysis (PCA) with Support Vector Machine (SVM) classification to identify the key factors influencing customer satisfaction using questionnaire data. Leveraging SVM's effectiveness in questionnaire-based classification and PCA's feature reduction capabilities, the study examines the impact of PCA on model performance. From a dataset comprising responses from 100 participants regarding product quality, price, and service level, PCA was employed to reduce the questionnaire features to three components. Results revealed Product Quality as the most significant factor affecting Customer Satisfaction, while Service Level exhibited the lowest influence. Comparative analysis of SVM models with and without PCA integration demonstrated improved performance metrics, including reduced error rates and enhanced accuracy. PCA mitigated overfitting risks, enhancing model generalization and interpretability. The findings underscore the practical significance of PCA in optimizing SVM classification models for customer satisfaction analysis. Future research may focus on optimal component selection and the robustness of PCA-SVM models across diverse domains, advancing machine learning applications in real-world scenarios.

INTRODUCTION

In statistical analysis, questionnaires are commonly used as a data source in research because they enable researchers to gather structured data from respondents, thus ensuring relevance to the research topic (Taherdoost, 2022). In data mining, the utilization of questionnaires as a primary data source is widely prevalent, particularly in identifying features with the most significant impact within the dataset (Kumar et al., 2022). In the context of classification, identifying features or variables with the most significant impact on the dataset can aid in determining the relationship between independent variables (features) and dependent variables (targets) (Cateni et al., 2023). By understanding the most influential features, we can develop a more accurate and efficient model for predicting the dependent variable based on relevant independent variables (Geche et al., 2019).

Principal Component Analysis (PCA) is a technique utilized to identify patterns within data by reducing the number of feature dimensions while preserving the maximum amount of information contained within the dataset (Honest, 2020). Several studies have demonstrated PCA's capability in feature reduction within classification models, as observed in algorithms Multi-Layer Perceptron (MLP) (Zhu et al., 2021), Support Vector Machine (SVM) (Yadav et al., 2019), K-Nearest Neighbors (KNN) (Lubis et al., 2020), Naïve Bayes (Odhiambo Omuya et al., 2021), and C4.5 (Nasution et al., 2018). Previous studies serve as crucial references on how PCA can identify the most important features within a dataset while preserving the core information within the dataset.

As one of the classification algorithms, SVM is a commonly used algorithm to address various problems, particularly those based on questionnaires (Kirchner & Signorino, 2018). In their study on the prediction of student's well-being from stress and sleep, (M & M, 2022) demonstrated that SVM successfully measured the correlation between Perceived Stress Scale (PSS) scores and Pittsburgh Sleep Quality Index (PSQI) global scores with the student's well-being factor. In a study evaluating the quality of piano lesson teachers, (Li, 2021) demonstrated that SVM could classify teacher quality based on questionnaire responses distributed to 500 respondents. Another study demonstrating SVM's capability in classifying questionnaire-based data was conducted by (Nakao et al., 2023), where this algorithm successfully classified hallux valgus based on questions regarding age, sex, height, weight, and foot measurements. A study on predicting students' academic performance based on personal, educational, behavioral, and extra-curricular features successfully demonstrated that SVM could classify data containing 92 questionnaire questions (Dabhade et al., 2021). (Bramhe et al., 2023) utilized a questionnaire dataset obtained from The Financial Opinion Mining and Question Answering Open Challenge, further demonstrating that SVM could classify student performance based on questions regarding attendance, study hours, extracurricular activities, and stress levels. These studies serve as evidence that the SVM algorithm is suitable for use in a questionnaire-based classification model, where the questions in the questionnaire are utilized as classification features.

This study employs an 8-question questionnaire addressing factors influencing customer satisfaction, such as product quality, price, and service level. Survey results from 100 respondent answers are classified using SVM and evaluated based on MSE, RMSE, and R2 values. PCA is employed to reduce question features, yielding the most significant factors influencing customer satisfaction from the SVM classification process. Classification evaluation results with and without feature reduction are then compared, determining the most significant factors influencing customer satisfaction among product, product price, and service level.

This research contributes to the existing literature by providing a comprehensive analysis of customer satisfaction factors through the integration of SVM classification and PCA feature reduction techniques applied to questionnaire data. While previous studies have demonstrated the effectiveness of SVM in classifying questionnaire-based data, this research extends the understanding by incorporating PCA to identify the most influential factors affecting customer satisfaction. By utilizing SVM and PCA in tandem, this study aims to offer a more nuanced understanding of the relationship between customer satisfaction and various product and service attributes. Furthermore, by comparing classification results with and without PCA feature reduction, this research seeks to elucidate the added value of PCA in enhancing the accuracy and interpretability of classification models for customer satisfaction analysis. Thus, this study not only advances the theoretical understanding of customer satisfaction but also provides practical insights for businesses to optimize their products and services based on key customer preferences and perceptions.

RESEARCH METHODS

A. Data Collection

This study utilizes a dataset derived from the responses of 100 participants regarding the influence of the variables Product Quality, Product Price, and Service Level on Customer Satisfaction. For each variable, two questions are provided, resulting in a total of 8 questions in the questionnaire. Each of these datasets is then divided into training and testing data with an 80:20 ratio.

Each question in the questionnaire employs a Likert scale, ranging from 1 for the lowest value to 5 for the highest value. Symbol Pi (where i represents the question number) denotes each question. The responses for P1 to P6 are utilized as features in the classification process, while the average responses for P7 and P8 are used as the target variables. Table 1 presents the detailed questions used in the questionnaire.

Table 1. Questionnaire's Questions

Questions		Variable
P1	Please rate the quality of the processed mangrove products you consume.	Product Quality
P2	Have you ever experienced any issues related to the quality of the processed mangrove products you consume?	
P3	How suitable is the price of the processed mangrove products you consume with their quality?	Product Price
P4	How often do you consider the price when purchasing processed mangrove products?	
P5	How would you rate your experience interacting with the seller or producer of mangrove processed products?	Service Level
P6	How important is good customer service in influencing your decision to purchase mangrove processed products?	
P7	How satisfied are you with the mangrove processed products you consume?	Customer Satisfaction
P8	Would you recommend the mangrove processed products you consume?	

B. Classification

This study employs the SVM algorithm in the classification model for training and testing data. SVM is a common technique utilized for classification tasks, renowned for its high accuracy and boasting the added benefit of requiring fewer data samples to mitigate overfitting (Samudra et al., 2023). In this research, the Polynomial kernel function is utilized, which includes γ (gamma constant), r (kernel constant), and p (polynomial degree) to adjust the flexibility of the classification outcomes (Pardede et al., 2023). Table 2 illustrates the detailed parameters utilized in the SVM model.

Table 2. SVM Parameters

Parameters	Value
γ	0.1
r	1
p	2

C. Feature Selection

In this study, PCA was employed to reduce the 8 questionnaire features to 3 features.

PCA functions to identify patterns within the features of the training and testing data by seeking similarities and differences among these features (Nasution et al., 2018).

The feature selection process in this study follows these steps (Basak et al., 2021):

1. Organize the dataset into an m $*$ n matrix, where m represents the number of measurement types and n represents the number of trials.
2. Calculate the mean of the entire dataset for each data dimension and normalize the data by subtracting the corresponding means from the numbers in each column.
3. Compute the Covariance Matrix for the raw dataset using the formula:

   $$cov(x, y) = \frac{1}{n-1} \sum_{i=1}^{n} [(x_i - \bar{x}) * (y_i - \bar{y})] \qquad 1$$

   where ¯x represents the arithmetic mean of data X, ¯y represents the arithmetic mean of data Y, and n is the number of observations. For example, for a 3-dimensional dataset, the covariance matrix will have 3 rows and 3 columns, with values representing the covariance between dimensions.

4. Calculate the Eigenvectors and Eigenvalues of the Covariance Matrix.
5. Select the components (eigenvectors) and form the feature vector. Arrange the eigenvectors by eigenvalue, from highest to lowest, to determine their significance. Form the feature vector by selecting the eigenvectors to retain and constructing a matrix with these eigenvectors as columns.
6. Compute the transpose of the feature vector and multiply it with the normalized original dataset.

When performing PCA, both cumulative variance and component variance are important measures, but they serve different purposes (Chen et al., 2021):

1. Cumulative Variance
   Cumulative variance shows the total amount of variance explained by each principal component, summed up across all components. It helps you understand how much of the original variability in the data is captured by the principal components. Cumulative variance is useful for determining how many principal components to retain in your analysis. You typically want to retain enough components to explain a high percentage of the total variance in the data.
2. Component Variance
   Component variance represents the variance explained by each individual principal component. It provides insight into the contribution of each component to the overall variability in the data. Component variance helps you understand the relative importance of each principal component in capturing the underlying

structure of the data.

In this study, we utilized component variance values to identify the questionnaire features most significant to Customer Satisfaction.

D. Model Configuration

This study utilizes the Orange Data Mining application to construct a classification model, leveraging widgets as depicted in Figure 1. The widgets utilized in this study are as follows:

1. File

   This widget is employed to read respondent answer files in XLSX format.

2. Data Sampler

   A widget designed to split the data into training and testing sets with an 80:20 ratio.

3. PCA

   This widget offers the PCA option, with three components chosen to reduce dataset features.

4. Select Column

   This widget is used to filter the features that will be used in the classification process. In this study, the Select Columns widget is employed to filter PC1, PC2, and PC3 as the classification features for the training and testing data.

5. Data Table

   Utilized to display the training and testing data used in the classification process.

6. SVM

   A learner widget containing pre-defined SVM parameter configurations.

7. Test and Score

   This widget functions to display the model's performance in terms of MSE, RMSE, and R2 values.

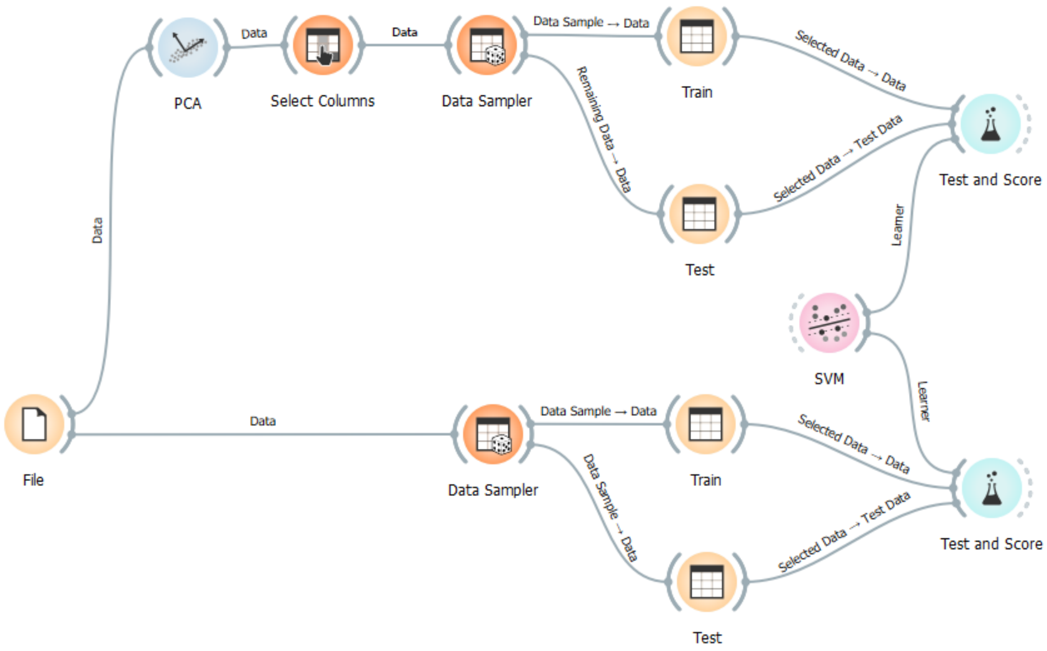

Figure 1. Model Configuration

The model in this study performs classification, feature selection, and performance evaluation using the following steps:

1. The model reads a dataset containing 100 respondent answers for each of the 8 question features and 1 target variable, Customer Satisfaction.
2. For the classification process, two datasets are utilized: the original dataset (without feature selection) and the PCA dataset (after feature selection).
3. The SVM algorithm is employed to train 80 previously divided data points for each of the original and PCA datasets.
4. The model's performance for the training data process is evaluated using MSE, RMSE, and R2 values.
5. The SVM algorithm is used to test 20 previously divided data points for each of the original and PCA datasets.
6. The model's performance for the testing data process is evaluated using MSE, RMSE, and R2 values.
7. The results of both training and testing data evaluations are then compared to assess whether PCA has successfully improved the classification performance of the SVM model.

RESULTS AND DISCUSSION

A. Results

From the feature selection process using PCA, 3 features are generated, namely PC1, PC2, and PC3, which are used as the comparative dataset for feature selection results. The result of the feature selection process using PCA is the component variance values for each questionnaire feature, as depicted in Figure 2.

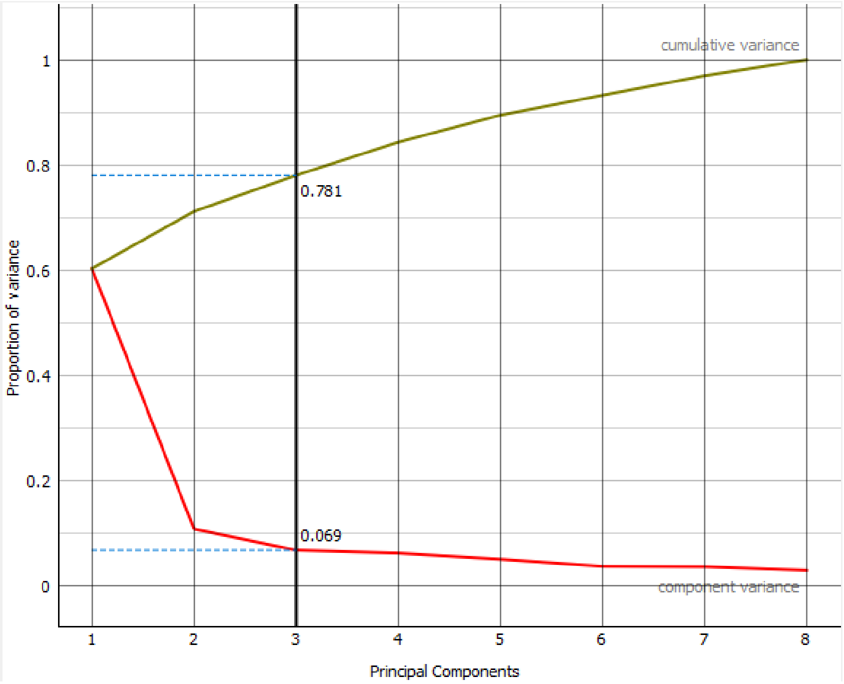

Figure 2. Component Variance Result

From Figure 2, it is evident that the questionnaire feature with the highest component variance value is P1 (component 1), with a value of 0.6. The figure also indicates that the questionnaire feature with the lowest component variance value is P8 (component 8).

From the classification model results for each original dataset and PCA, the metrics of MSE, RMSE, and R2 are obtained as shown in Table 2.

Table 2. Metrics Evaluation

Dataset	Process	MSE	RMSE	R2
Without PCA	Training	0.004	0.062	0.994
	Testing	0.008	0.087	0,988
With PCA	Training	0.003	0.056	0.995
	Testing	0.003	0.051	0.996

B. Discussion

The results of feature selection using PCA indicate that P1 (component 1) has the highest component variance value. This suggests that P1 is the most significant feature among all questionnaire questions. Based on the previous question design, P1 is part of the Product Quality factor; therefore, it can be concluded that Product Quality is the most significant factor influencing Customer Satisfaction in this study. On the contrary, P8 (component 8) has the lowest component variance value. This indicates that P8 is the least significant feature among all questionnaire questions. P8 belongs to the Service Level factor; hence, it can be inferred that the Service Level has the lowest influence on Customer Satisfaction.

The comparison of model performance metrics between SVM models with and without PCA reveals notable improvements across various indicators. The reduced MSE and RMSE values in both training and testing datasets with PCA integration signify enhanced model accuracy and precision. Additionally, the higher R2 values obtained with PCA suggest better model fit and increased variance explained by the predictors, indicating improved generalization ability.

The decrease in MSE and RMSE values in the testing dataset with PCA indicates a reduction in model overfitting. By reducing the number of features and capturing the essential variance within the data, PCA helps mitigate the risk of overfitting, ensuring that the model's predictions generalize well to unseen data instances. This is further supported by the consistent performance improvements observed in both training and testing datasets with PCA integration.

The enhanced performance of SVM classification models with PCA integration also enhances the interpretability of the model. By reducing the dimensionality of the feature space while retaining essential information, PCA simplifies the model's complexity and facilitates the identification of the most influential features driving classification outcomes. This increased interpretability enables stakeholders to gain deeper insights into the underlying relationships within the data and make more informed decisions based on the model's predictions.

The findings highlight the practical significance of incorporating PCA into SVM classification models for various applications, including customer satisfaction analysis, medical diagnosis, and financial forecasting. The observed improvements in model accuracy and generalization ability underscore the potential of PCA as a valuable preprocessing technique for enhancing the performance of SVM classifiers. Future research directions may involve investigating the optimal number of principal components and exploring the robustness of PCA-SVM models across different datasets and problem domains.

CONCLUSION

This study has investigated the impact of Principal Component Analysis (PCA) on Support Vector Machine (SVM) classification, focusing on identifying the most significant factors influencing customer satisfaction from questionnaire data. By integrating SVM classification with PCA feature reduction techniques, this research contributes to a comprehensive analysis of customer satisfaction factors and extends the understanding of their relationship with product and service attributes. The results of feature selection using PCA revealed that Product Quality, represented by the questionnaire feature P1, emerged as the most significant factor influencing Customer Satisfaction. Conversely, Service Level, represented by the feature P8, exhibited the lowest influence on Customer Satisfaction. This highlights the importance of prioritizing product quality in optimizing customer satisfaction strategies. Comparative analysis of SVM models with and without PCA integration demonstrated notable improvements in model performance metrics. The reduced error rates, enhanced model accuracy, and improved generalization ability observed with PCA integration underscore its effectiveness in optimizing SVM classification models. Furthermore, PCA mitigated overfitting risks by reducing the dimensionality of the feature space and capturing essential variance within the data. The enhanced interpretability facilitated by PCA simplifies model complexity and enables stakeholders to gain deeper insights into underlying data relationships. This provides valuable insights for businesses to optimize product and service offerings based on key customer preferences and perceptions. The practical significance of incorporating PCA into SVM classification models extends to various domains beyond customer satisfaction analysis. Future research directions may explore optimal principal component selection and the robustness of PCA-SVM models across diverse datasets and problem domains, further advancing the application of machine learning methodologies in real-world scenarios.

REFERENCES

Basak, H., Roy, A., Lahiri, J. B., Bose, S., & Patra, S. (2021). *SVM and ANN based Classification of EMG signals by using PCA and LDA. I.* http://arxiv.org/abs/2110.15279

Bramhe, M. V, Gohade, S., Kapse, R., Thamke, S., & Mehar, A. (2023). Student Performance Prediction Using Regression Analysis & Feature- Based Opinion Mining on Student Feedback. *Journal of Harbin Engineering University*, *44*(7), 360–366.

Cateni, S., Colla, V., & Vannucci, M. (2023). Improving the Stability of the Variable Selection with Small Datasets in Classification and Regression Tasks. *Neural*

Processing Letters, *55*(5), 5331–5356. https://doi.org/10.1007/s11063-022-10916-4

Chen, J., Gong, F., Xiang, S., & Yu, T. (2021). Application of principal component analysis in evaluation of epidemic situation policy implementation. *Journal of Physics: Conference Series*, *1903*(1). https://doi.org/10.1088/1742-6596/1903/1/012056

Dabhade, P., Agarwal, R., Alameen, K. P., Fathima, A. T., Sridharan, R., & Gopakumar, G. (2021). Educational data mining for predicting students' academic performance using machine learning algorithms. *Materials Today: Proceedings*, *47*(xxxx), 5260–5267. https://doi.org/10.1016/j.matpr.2021.05.646

Geche, F., Mulesa, O., Hrynenko, V., & Smolanka, V. (2019). Search for impact factor characteristics in construction of linear regression models. *Technology Audit and Production Reserves*, *3*(2(47)), 20–25. https://doi.org/10.15587/2312-8372.2019.175020

Honest, N. (2020). A survey on Feature Selection Techniques. *GIS SCIENCE JOURNAL*, *7*(6), 353–358. https://www.researchgate.net/publication/344121693

Kirchner, A., & Signorino, C. S. (2018). Using Support Vector Machines for Survey Research. *Survey Practice*, *11*(1), 1–14. https://doi.org/10.29115/sp-2018-0001

Kumar, M., Sharma, C., Sharma, S., Nidhi, N., & Islam, N. (2022). Analysis of Feature Selection and Data Mining Techniques to Predict Student Academic Performance. *2022 International Conference on Decision Aid Sciences and Applications, DASA 2022*, *March*, 1013–1017. https://doi.org/10.1109/DASA54658.2022.9765236

Li, T. (2021). Application of slack variable-optimized SVM on piano teaching reform. *Journal of Physics: Conference Series*, *1941*(1), 012084. https://doi.org/10.1088/1742-6596/1941/1/012084

Lubis, A. H., Sihombing, P., & Nababan, E. B. (2020). Analysis of Accuracy Improvement in K-Nearest Neighbor using Principal Component Analysis (PCA). *Journal of Physics: Conference Series*, *1566*(1), 2–10. https://doi.org/10.1088/1742-6596/1566/1/012062

M, S. S., & M, J. (2022). Prediction of Student's Wellbeing from Stress and Sleep Questionnaire data using Machine Learning Approach. *2022 IEEE Bombay Section Signature Conference (IBSSC)*, 1–6. https://doi.org/10.1109/IBSSC56953.2022.10037549

Nakao, H., Imaoka, M., Hida, M., Imai, R., Nakamura, M., Matsumoto, K., & Kita, K. (2023). Determination of individual factors associated with hallux valgus using SVM-RFE. *BMC Musculoskeletal Disorders*, *24*(1), 1–7. https://doi.org/10.1186/s12891-023-06303-2

Nasution, M. Z. F., Sitompul, O. S., & Ramli, M. (2018). PCA based feature reduction to improve the accuracy of decision tree c4.5 classification. *Journal of Physics: Conference Series*, *978*(1), 0–6. https://doi.org/10.1088/1742-6596/978/1/012058

Odhiambo Omuya, E., Onyango Okeyo, G., & Waema Kimwele, M. (2021). Feature Selection for Classification using Principal Component Analysis and Information Gain. *Expert Systems with Applications*, *174*(November 2020), 114765. https://doi.org/10.1016/j.eswa.2021.114765

Pardede, D., Wanayumini, & Rosnelly, R. (2023). A Combination Of Support Vector Machine And Inception-V3 In Face-Based Gender Classification. *International Conference on Information Science and Technology Innovation (ICoSTEC)*, *2*(1), 34–39. https://doi.org/10.35842/icostec.v2i1.30

Samudra, J. T., Rosnelly, R., & Situmorang, Z. (2023). Comparative Analysis of SVM and Perceptron Algorithms in Classification of Work Programs. *MATRIK : Jurnal Manajemen, Teknik Informatika Dan Rekayasa Komputer*, *22*(2), 285–298. https://doi.org/10.30812/matrik.v22i2.2479

Taherdoost, H. (2022). Designing a Questionnaire for a Research Paper: A Comprehensive Guide to Design and Develop an Effective Questionnaire. *Asian Journal of Managerial Science*, *11*(1), 8–16. https://doi.org/10.51983/ajms-2022.11.1.3087

Yadav, A., Jamir, I., Jain, R. R., & Sohani, M. (2019). Breast Cancer Prediction using SVM with PCA Feature Selection Method. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, *5*(2), 969–978. https://doi.org/10.32628/cseit1952277

Zhu, X., Dong, H., Salvo Rossi, P., & Landrø, M. (2021). Feature Selection Based on Principal Component Regression for Underwater Source Localization by Deep Learning. *Remote Sensing*, *13*(8), 1486. https://doi.org/10.3390/rs13081486