SINTECHCOM: Science, Technology and Communication Journal
Vol. 6, No. 3, June 2026, pp. 463-476, DOI: 10.59190/stc.v6i3.401

Comparative evaluation of PCA-based feature extraction and chi-square feature selection for student burnout classification using support vector machine

Mira Salmira*, Junadhi, Rahmiati, Triyani Arita Fitri
Department of Informatics Engineering, Universitas Sains dan Teknologi Indonesia, Pekanbaru 28299, Indonesia

ABSTRACT

Student depression has emerged as a critical mental health issue that adversely affects academic performance, psychological well-being, and quality of life. Early identification of depression risk is essential to enable timely intervention and effective mental health management. This study compares the effectiveness of two feature engineering techniques, principal component analysis (PCA)-based feature extraction and chi-square feature selection, for student depression classification using support vector machine (SVM). The experiments employed the student depression dataset from Kaggle, containing demographic, academic, lifestyle, and psychological attributes. Data preprocessing included data cleaning, label encoding, and feature scaling before feature engineering and classification. PCA was applied to reduce feature dimensionality while preserving the maximum data variance, whereas chi-square selected the most relevant features based on statistical significance. Model performance was evaluated using accuracy, precision, recall, and F1-score. The results demonstrate that PCA consistently outperformed chi-square feature selection. The PCA–SVM model achieved an accuracy of 84.09%, precision of 84.24%, recall of 84.09%, and F1-score of 84.14%, compared with 83.61%, 83.75%, 83.61%, and 83.65%, respectively, for the chi-square–SVM model. These findings indicate that PCA is more effective in reducing feature redundancy while preserving informative patterns, resulting in improved classification performance. Therefore, PCA-based feature extraction is a more suitable feature engineering approach for SVM-based student depression classification and offers a promising solution for intelligent early mental health screening in higher education.

ARTICLE INFO

Article history:

Received Jun 28, 2026
Revised Jun 29, 2026
Accepted Jun 30, 2026

Keywords:

Chi-Square
Feature Engineering
Principal Component
Student Depression
Support Vector Machine

* **Corresponding Author**
E-mail address: 2417052802123@usti.ac.id

1. INTRODUCTION

Student mental health has become an increasingly important concern in higher education institutions worldwide. University students are frequently exposed to various academic and non-academic pressures, including intensive coursework, examinations, project deadlines, financial difficulties, social adaptation, and concerns regarding future careers [1]. The accumulation of these pressures may negatively affect students' psychological well-being and lead to academic burnout. Academic burnout is generally characterized by emotional exhaustion, cynicism toward academic activities, and a reduced sense of academic accomplishment. If left unaddressed, burnout can adversely affect students' learning motivation, academic performance, engagement, and overall quality of life [2, 3].

Recent studies have reported a growing prevalence of burnout among university students. Several investigations have indicated that a substantial proportion of students experience moderate to

high levels of academic burnout due to prolonged academic stress and increasing educational demands. The consequences of burnout extend beyond academic outcomes and may contribute to anxiety, depression, low self-esteem, and even student dropout. Consequently, early identification of burnout risk has become an important issue for educational institutions seeking to provide timely interventions and appropriate psychological support [4-6].

Traditionally, burnout assessment is conducted through questionnaires and manual evaluation by educators, counselors, or mental health professionals. Although such approaches provide valuable insights, they often require considerable time and effort, particularly when dealing with large student populations. Furthermore, manual assessment may be influenced by subjective judgment and may not be suitable for continuous large-scale monitoring. The increasing availability of educational and behavioral data has created opportunities for the application of machine learning techniques to support automated burnout detection and prediction [7, 8].

Machine learning has demonstrated significant potential in educational data analysis and mental health prediction. By learning patterns from historical data, machine learning models can identify complex relationships among multiple factors associated with student well-being. Various machine learning algorithms have been employed for educational prediction tasks, including Decision Trees, Random Forest, Logistic Regression, Naïve Bayes, and Support Vector Machine (SVM) [9]. Among these methods, SVM has gained considerable attention due to its ability to handle high-dimensional data, strong generalization capability, and effectiveness in classification problems involving limited or moderately sized datasets [10, 11].

Despite the effectiveness of SVM, the performance of classification models is highly dependent on the quality of input features. Educational and mental health datasets often contain numerous variables that may be irrelevant, redundant, or weakly associated with the target class. The inclusion of such features can increase computational complexity, reduce classification performance, and potentially lead to overfitting. Therefore, feature engineering plays a crucial role in improving model effectiveness by identifying or generating more informative feature representations [12, 13].

Two commonly used approaches in feature engineering are feature extraction and feature selection. Feature extraction transforms original features into a new set of lower-dimensional features while preserving essential information contained in the dataset. One of the most widely adopted feature extraction methods is Principal Component Analysis (PCA) [14], which reduces dimensionality by generating orthogonal principal components that capture the maximum variance of the data. In contrast, feature selection aims to identify and retain the most relevant original features while removing irrelevant or redundant variables. Among various feature selection techniques, Chi-Square has been extensively utilized due to its simplicity, computational efficiency, and ability to measure the statistical relationship between individual features and target classes [15, 16].

Previous studies have successfully applied machine learning methods to mental health prediction and burnout detection. Several researchers have demonstrated that supervised learning algorithms can effectively identify burnout-related patterns among university students. Other studies have shown that feature engineering techniques can significantly influence classification performance by improving data quality and reducing noise. However, most existing studies primarily focus on comparing classification algorithms or evaluating a single feature optimization approach. Comparatively fewer studies have investigated how different feature engineering strategies affect burnout classification performance when the same classification algorithm is employed [17-19].

This limitation creates an important research gap. Although PCA and Chi-Square are among the most frequently used feature optimization techniques, there is still insufficient evidence regarding their comparative effectiveness in student burnout classification using Support Vector Machine. Understanding the strengths and limitations of these approaches is essential for developing more accurate and computationally efficient burnout prediction systems. Furthermore, identifying the most suitable feature processing strategy may help educational institutions implement machine learning solutions that support early intervention and student well-being management [20, 21].

To address this gap, this study presents a comparative evaluation of PCA-based feature extraction and Chi-Square feature selection for student burnout classification using Support Vector Machine. The research utilizes the Student Mental Health and Burnout Dataset and investigates the impact of both feature optimization techniques on classification performance. The models are

evaluated using Accuracy, Precision, Recall, and F1-Score metrics to provide a comprehensive assessment of predictive capability.

The main contributions of this study are threefold. First, it develops a machine learning-based framework for student burnout classification using Support Vector Machine. Second, it provides a systematic comparison between PCA-based feature extraction and Chi-Square feature selection within the same experimental setting. Third, it identifies the feature optimization approach that offers superior classification performance for student burnout prediction. The findings are expected to contribute to the growing body of research on educational data mining and mental health analytics while providing practical insights for the development of intelligent student support systems.

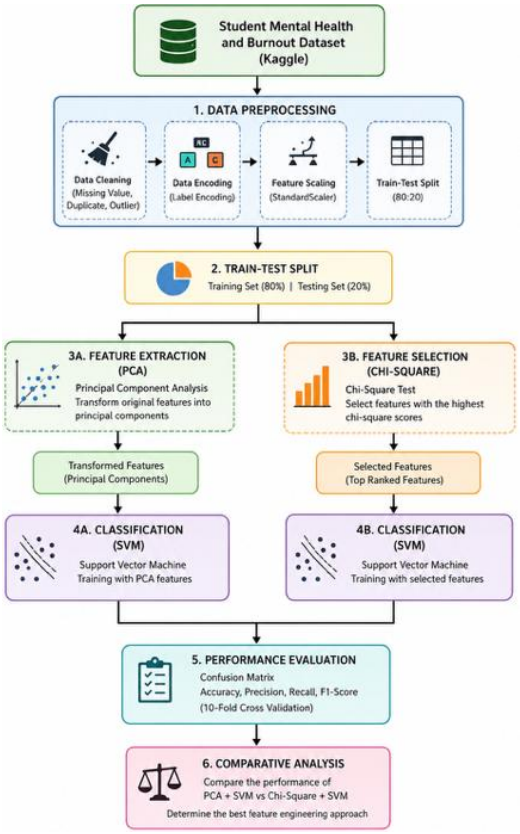

Figure 1. Proposed classification workflow.

2. RESEARCH METHODS

This study adopted an experimental research approach to investigate the effectiveness of two feature engineering techniques, namely PCA-based feature extraction and Chi-Square feature selection, in improving student burnout classification using the Support Vector Machine (SVM) algorithm. The proposed methodology consists of six sequential stages: dataset collection, data preprocessing, feature engineering, model development, performance evaluation, and comparative analysis. The study began with the collection of the Student Mental Health and Burnout Dataset obtained from the Kaggle platform. The collected data were subsequently subjected to preprocessing procedures, including data cleaning, encoding, feature scaling, and dataset partitioning, to enhance data quality and ensure compatibility with machine learning algorithms. Following preprocessing, two feature engineering strategies were implemented independently. The first strategy employed Principal Component Analysis (PCA) to transform the original features into a reduced set of principal components, while the second strategy utilized the Chi-Square method to select the most relevant features associated with the target class. The resulting feature sets were then used to train SVM classification models. Finally, the performance of each model was evaluated using Accuracy, Precision, Recall, and F1-Score metrics, and a comparative analysis was conducted to identify the

most effective feature optimization approach for student burnout classification. Figure 1 illustrates the overall research framework adopted in this study.

2.1. Dataset

The dataset used in this study was the Student Depression Dataset, obtained from the Kaggle platform. The dataset contains demographic, academic, lifestyle, and psychological factors that are associated with students' mental health conditions. It comprises 27,901 student records and 18 attributes, including one target variable and seventeen predictor variables. The predictor variables include gender, age, city, profession, academic pressure, work pressure, cumulative grade point average (CGPA), study satisfaction, job satisfaction, sleep duration, dietary habits, degree, suicidal thoughts, work/study hours, financial stress, and family history of mental illness. The target variable is Depression, which indicates whether a student exhibits depressive symptoms. Prior to model development, data preprocessing procedures were conducted to handle missing values, transform categorical attributes into numerical representations, standardize feature values, and prepare the dataset for machine learning analysis. The dataset was selected because it provides comprehensive information regarding students' academic conditions, lifestyle patterns, and psychological factors, making it suitable for evaluating the effectiveness of feature engineering techniques in mental health classification tasks. Table 1 presents the attributes used in this study.

Table 1. Dataset attributes.

Attribute	Data type
Gender	Categorical
Age	Numerical
City	Categorical
Profession	Categorical
Academic pressure	Numerical
Work pressure	Numerical
CGPA	Numerical
Study satisfaction	Numerical
Job satisfaction	Numerical
Sleep duration	Categorical
Dietary habits	Categorical
Degree	Categorical
Suicidal thoughts	Categorical
Work/study hours	Numerical
Financial stress	Numerical
Family history of mental illness	Categorical
Depression	Target variable

2.2. Data Preprocessing

Data preprocessing was conducted to improve dataset quality and prepare the data for machine learning analysis. Raw datasets often contain missing values, categorical attributes, and features with different numerical scales that may negatively affect classification performance. Therefore, several preprocessing procedures were performed before applying feature engineering techniques and developing the Support Vector Machine (SVM) classification model. The preprocessing stage consisted of data cleaning, data encoding, and feature scaling.

2.2.1. Data Cleaning

The first preprocessing step involved examining the dataset for missing values and inconsistencies. Based on the initial inspection, the dataset contained missing values in the Financial Stress attribute, while the remaining attributes were complete. To maintain data integrity and avoid information loss, missing values were replaced using an appropriate imputation strategy. This process ensured that all records could be utilized during model training and evaluation. In addition, the dataset was examined for duplicate records and invalid observations to improve data consistency and reliability. After the cleaning process, the dataset was considered suitable for further analysis.

2.2.2. Data Encoding

Several attributes in the Student Depression Dataset were categorical, including Gender, City, Profession, Sleep Duration, Dietary Habits, Degree, Have You Ever Had Suicidal Thoughts?, and Family History of Mental Illness. Since Support Vector Machine (SVM) requires numerical input, categorical variables were transformed into numerical representations using the Label Encoding technique.

Label Encoding assigns a unique integer value to each category within a feature. The transformation can be expressed as:

$$LE(x_i) = k \tag{1}$$

where: $LE(x_i)$ denotes the encoded value of category; and k represents a unique numerical label assigned to each category.

Through this process, all categorical attributes were converted into numerical form, enabling them to be processed by machine learning algorithms.

2.2.3. Feature Scaling

After the encoding stage, feature scaling was performed to normalize the numerical attributes and reduce differences in feature magnitudes. The dataset contains variables with varying ranges, such as Age, CGPA, Academic Pressure, Work Pressure, Study Satisfaction, Work/Study Hours, and Financial Stress. Without scaling, features with larger values could dominate the learning process and negatively influence classification performance. In this study, feature scaling was performed using the StandardScaler method. Standardization transforms each feature to have a mean of zero and a standard deviation of one. The standardized value is calculated using:

$$z = \frac{x-\mu}{\sigma} \tag{2}$$

where: z is the standardized value; x is the original feature value; μ is the mean of the feature; and σ is the standard deviation of the feature.

Feature scaling is particularly important because both Principal Component Analysis (PCA) and Support Vector Machine (SVM) are sensitive to differences in feature scales. Standardization ensures that all features contribute proportionally during model training and prevents attributes with larger numerical ranges from disproportionately affecting the classification process. The output of the preprocessing stage was a clean, numerical, and standardized dataset that was subsequently used for PCA-based feature extraction, Chi-Square feature selection, and SVM classification.

2.3. Feature Engineering

Feature engineering was performed to improve data representation and investigate the impact of different feature optimization techniques on classification performance. In this study, two feature engineering approaches were evaluated: PCA-based feature extraction and Chi-Square feature selection. Both techniques were applied to the preprocessed dataset before classification using Support Vector Machine (SVM). The objective of feature engineering is to reduce irrelevant information, minimize redundancy among features, and improve the effectiveness of the classification model. The resulting feature sets were subsequently used as inputs for the SVM classifier, and their performances were compared using several evaluation metrics.

2.3.1. PCA-Based Feature Extraction

Principal Component Analysis (PCA) was employed as the feature extraction technique to transform the original features into a smaller set of uncorrelated variables called principal components. PCA aims to preserve most of the variance contained in the original dataset while reducing dimensionality. The PCA process begins by computing the covariance matrix of the standardized dataset. The covariance matrix is calculated as follows:

$$Cov(X) = \frac{1}{n-1}(X - \bar{X})^T(X - \bar{X})z = \frac{x-\mu}{\sigma} \tag{3}$$

where: X represents the feature matrix; $\bar{X}$ denotes the mean vector; and n is the number of observations.

Subsequently, eigenvalues and eigenvectors are obtained from the covariance matrix by solving:

$$Cov(X)v = \lambda v \tag{4}$$

where: λ is the eigenvalue; and v is the eigenvector.

The eigenvectors corresponding to the largest eigenvalues are selected as principal components because they capture the highest variance in the dataset. The original data are then projected onto the selected principal components according to:

$$Z = XW \tag{5}$$

where: Z represents the transformed feature matrix,; X is the standardized dataset; and W is the matrix of selected eigenvectors.

The transformed features generated by PCA were subsequently used as input for SVM classification. PCA is expected to reduce feature redundancy and improve computational efficiency while retaining important information from the original dataset.

2.3.2. Chi-Square Feature Selection

Chi-Square was employed as the feature selection technique to identify the most relevant features associated with the target class. Unlike PCA, Chi-Square does not generate new features but selects the most informative attributes from the original dataset. The Chi-Square statistic measures the degree of dependency between a feature and the target variable. The score is computed as:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \tag{6}$$

where: O_i is the observed frequency; and E_i is the expected frequency.

A higher Chi-Square score indicates a stronger relationship between the feature and the target class. Therefore, features with larger scores are considered more informative for classification purposes. After calculating the Chi-Square score for each feature, all features were ranked according to their scores. The top-ranked features were then selected and used as input variables for the SVM classifier. This process reduces the influence of irrelevant attributes while maintaining the interpretability of the original features. The performance of the Chi-Square-selected feature set was subsequently compared with the PCA-transformed feature set to determine the most effective feature engineering approach for student depression classification.

2.4. Support Vector Machine Classification

Support Vector Machine (SVM) was employed as the classification algorithm in this study due to its ability to handle high-dimensional data and its strong generalization capability in binary classification problems. SVM aims to identify an optimal hyperplane that maximizes the margin between different classes, thereby improving the classifier's ability to distinguish between students with and without depression. Given a training dataset consisting of feature vectors and class labels, SVM constructs a decision boundary represented by the following hyperplane:

$$f(x) = w^T x + b \tag{7}$$

where: x represents the input feature vector; w denotes the weight vector; and b is the bias term.

The optimal hyperplane is determined by maximizing the margin between the nearest data points from different classes, known as support vectors. This optimization problem can be formulated:

$$\min \frac{1}{2} \| w \|^2 \tag{8}$$

subject to:

$$y_i(w^T x_i + b) \geq 1 \tag{9}$$

where: x_i denotes the training samples; y_i represents the class labels; and $\| w \|$ is the Euclidean norm of the weight vector.

The objective of this optimization process is to maximize the separation margin while minimizing classification errors. In this study, SVM was applied to two different feature-engineering scenarios:

- PCA + SVM, where the transformed principal components obtained from Principal Component Analysis were used as input features.
- Chi-Square + SVM, where the selected features obtained from the Chi-Square method were used as input features.

The same classification algorithm was employed in both scenarios to ensure a fair comparison between the two feature engineering approaches. The resulting models were trained using the processed training dataset and subsequently evaluated on the testing dataset. The classification performance of each model was measured using Accuracy, Precision, Recall, and F1-Score. The comparison of these performance metrics was used to determine the most effective feature engineering technique for improving student depression classification using Support Vector Machine.

2.5. Performance Evaluation

Performance evaluation was conducted to assess the effectiveness of the proposed classification models in predicting student depression. The evaluation process aimed to compare the performance of the two feature engineering approaches, namely PCA-based feature extraction and Chi-Square feature selection, when used in conjunction with the Support Vector Machine (SVM) classifier.

The evaluation was performed using a confusion matrix, which provides a summary of classification results by comparing the predicted labels with the actual labels. A confusion matrix consists of four possible outcomes: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN), as presented in Table 2.

Table 2. Confusion matrix.

Actual class	Predicted positive	Predicted negative
Positive	True positive (TP)	False negative (FN)
Negative	False positive (FP)	True negative (TN)

where: true positive (TP) represents correctly predicted positive instances; true negative (TN) represents correctly predicted negative instances; false positive (FP) represents negative instances incorrectly classified as positive; and false negative (FN) represents positive instances incorrectly classified as negative.

Based on the confusion matrix, four evaluation metrics were calculated: Accuracy, Precision, Recall, and F1-Score.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{10}$$

Precision measures the reliability of positive predictions:

$$Precision = \frac{TP}{TP+FP} \tag{11}$$

Recall evaluates the ability of the model to identify relevant instances:

$$Recall = \frac{TP}{TP+FN} \tag{12}$$

F1-Score provides a balanced assessment of Precision and Recall:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (13)$$

The values of Accuracy, Precision, Recall, and F1-Score obtained from the PCA + SVM and Chi-Square + SVM models were compared to determine the most effective feature engineering approach for student depression classification. The model achieving the highest evaluation scores was considered to provide superior classification performance.

3. RESULTS AND DISCUSSIONS

3.1. Exploratory Data Analysis

Before model development, exploratory data analysis (EDA) was conducted to understand the characteristics of the dataset and identify relationships between features and the target variable. Correlation analysis revealed that several attributes exhibited a strong association with student depression. The correlation matrix showed that Have You Ever Had Suicidal Thoughts? had the highest positive correlation with depression, followed by Academic Pressure, Financial Stress, and Work/Study Hours. Conversely, Age demonstrated a negative correlation with the target variable. These findings indicate that psychological and academic-related factors play a significant role in predicting student depression.

Table 3. Top features correlated with depression.

Feature	Correlation
Have you ever had suicidal thoughts?	0.546
Academic pressure	0.475
Financial stress	0.364
Work/study hours	0.209
Age	-0.226

The results suggest that students experiencing high academic pressure, financial difficulties, and suicidal thoughts are more likely to exhibit depressive symptoms. These findings are consistent with previous studies highlighting the influence of academic and psychological stressors on student mental health.

3.2. Feature Engineering Results

3.2.1. PCA-Based Feature Extraction

Principal Component Analysis (PCA) was applied to reduce the dimensionality of the dataset while preserving most of the information contained in the original features. PCA transforms the original feature space into a set of orthogonal principal components that capture the maximum variance within the data. To determine the optimal number of principal components, cumulative explained variance analysis was conducted. The cumulative explained variance indicates the proportion of total dataset variance retained as additional components are included. Figure 2 presents the cumulative explained variance obtained from the PCA transformation.

Based on Figure 2, the cumulative explained variance increased steadily with the addition of principal components. The analysis shows that the cumulative explained variance exceeded the 95% threshold when 15 principal components were retained. Specifically, the selected components preserved approximately 98% of the total variance contained in the original dataset. Therefore, 15 principal components were selected for subsequent classification experiments. The dimensionality reduction process decreased feature redundancy while maintaining most of the information necessary for classification. The resulting transformed feature set was subsequently used as input for the Support Vector Machine (SVM) classifier.

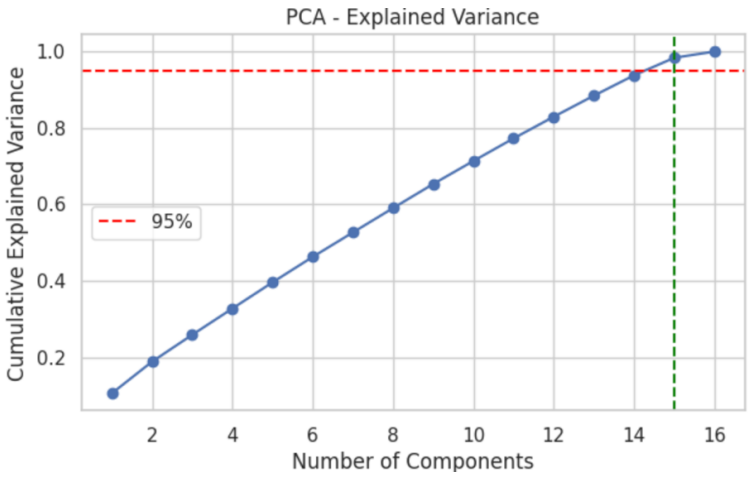

Figure 2. Cumulative explained variance of principal components.

Table 3. PCA dimensionality reduction results.

Description	Value
Original features	17
Selected principal components	15
Cumulative explained variance	98%
Variance threshold	95%

The results indicate that PCA successfully reduced the dimensionality of the dataset while preserving a substantial amount of information, making it suitable for the subsequent classification process.

3.2.2. Chi-Square Feature Selection

The Chi-Square method was employed to identify the most relevant features associated with the target variable. Unlike PCA, which transforms the original feature space into a new set of components, Chi-Square retains the original features and ranks them according to their statistical association with the class label. The Chi-Square score was calculated for each feature, and features with higher scores were considered more informative for classification. Figure 3 illustrates the distribution of Chi-Square scores across all features in the dataset.

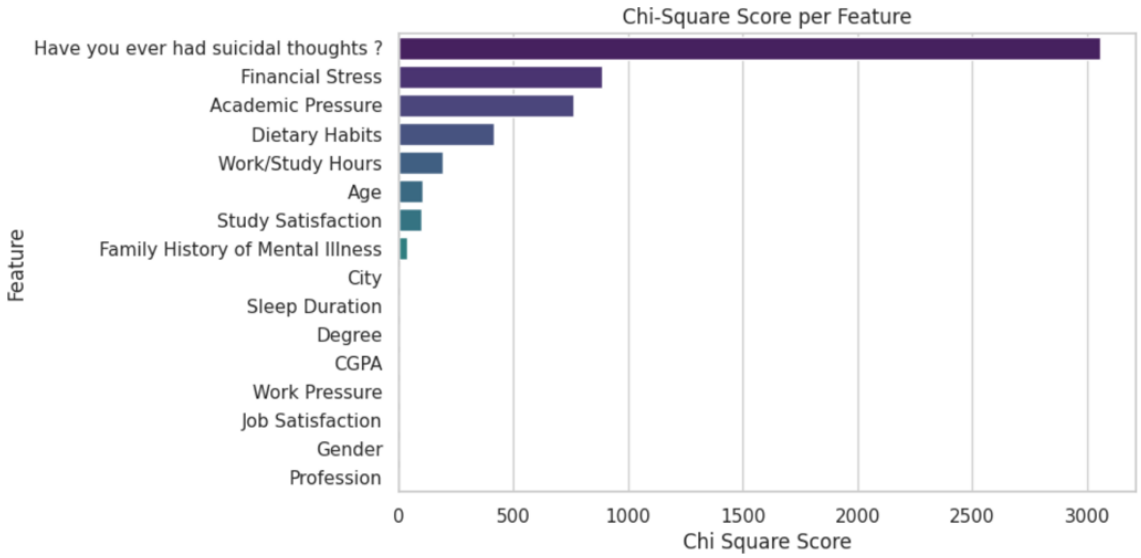

Figure 3. Chi-square scores of features.

Based on Figure 3, the feature Have You Ever Had Suicidal Thoughts? obtained the highest Chi-Square score, significantly outperforming all other variables. This result indicates that suicidal thoughts have the strongest statistical relationship with depression status and represent the most influential predictor in the dataset. Other highly ranked features include Financial Stress, Academic Pressure, and Dietary Habits, suggesting that psychological, financial, and academic-related factors contribute substantially to depression among students. In contrast, variables such as Gender, Profession, Job Satisfaction, and Work Pressure obtained relatively low Chi-Square scores, indicating a weaker association with the target class. The top-ranked features identified by the Chi-Square method are presented in Table 4.

Table 4. Top features selected by chi-square.

Feature	Chi-square score
Have you ever had suicidal thoughts?	3057.29
Financial stress	890.07
Academic pressure	764.37
Dietary habits	416.24
Work/Study Hours	196.54

The findings reveal that mental health indicators and academic-related pressures are the dominant factors associated with student depression. These results are consistent with previous studies reporting that suicidal ideation, financial burden, and academic stress are among the strongest predictors of psychological distress in university students. The selected features were subsequently used as input variables for the Support Vector Machine classifier and compared with the PCA-transformed feature set in the classification experiments.

3.3. Classification Performance Comparison

To evaluate the effectiveness of the proposed feature engineering approaches, Support Vector Machine (SVM) classifiers were trained using the feature sets generated by PCA and Chi-Square. Model performance was assessed using Accuracy, Precision, Recall, and F1-Score.

Table 5. Performance comparison of PCA + SVM and chi-square + SVM.

Method	Accuracy	Precision	Recall	F1-score
PCA + SVM	0.8409	0.8424	0.8409	0.8414
Chi-square + SVM	0.8361	0.8375	0.8361	0.8365

The results presented in Table 5 indicate that both feature engineering approaches achieved comparable classification performance. However, the PCA-based model consistently outperformed the Chi-Square-based model across all evaluation metrics. The PCA + SVM model achieved an Accuracy of 84.09%, Precision of 84.24%, Recall of 84.09%, and F1-Score of 84.14%. In comparison, the Chi-Square + SVM model obtained an Accuracy of 83.61%, Precision of 83.75%, Recall of 83.61%, and F1-Score of 83.65%. To facilitate visual comparison, the classification performance of both models is illustrated in Figure 4.

The results demonstrate that PCA achieved slightly better performance than Chi-Square in all evaluation metrics. Specifically, PCA improved Accuracy by 0.48%, Precision by 0.49%, Recall by 0.48%, and F1-Score by 0.49% compared to Chi-Square. Although the performance improvement is relatively small, the findings suggest that PCA was more effective in preserving informative patterns within the dataset. By transforming correlated features into a set of orthogonal principal components, PCA was able to reduce redundancy while maintaining important information for classification.

Consequently, the SVM classifier was able to construct a more discriminative decision boundary. In contrast, Chi-Square evaluates each feature independently based on its statistical relationship with the target variable. While this approach successfully identifies important features, it may overlook interactions among variables, potentially leading to a slight reduction in classification performance. The highest F1-Score was achieved by the PCA + SVM model (84.14%), indicating a

better balance between Precision and Recall. Since F1-Score considers both false positives and false negatives, this result suggests that PCA provides a more robust feature representation for student depression classification.

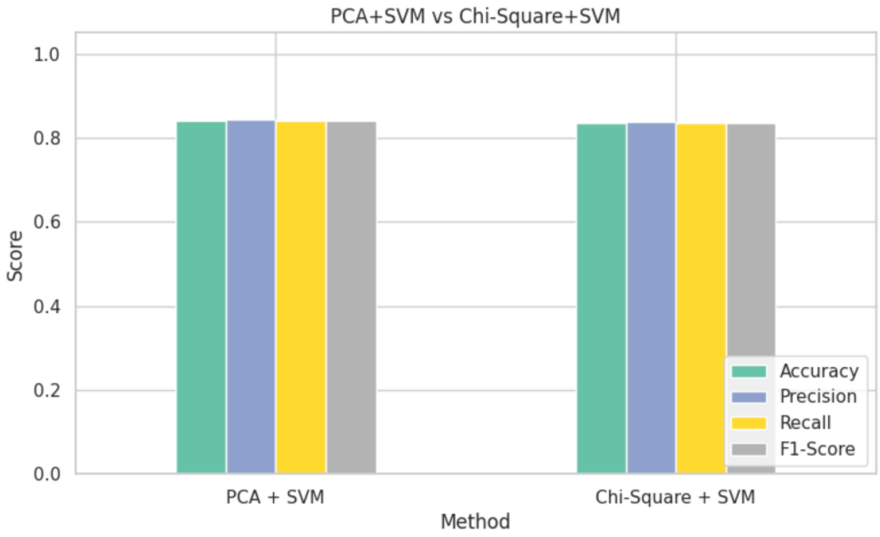

Figure 4. Performance comparison between PCA + SVM and chi-square + SVM.

Overall, the experimental results demonstrate that PCA-based feature extraction is slightly more effective than Chi-Square feature selection when combined with Support Vector Machine for student depression classification. Therefore, PCA can be considered the preferred feature engineering technique for this dataset.

3.4. Discussion

The experimental results demonstrate that both PCA-based feature extraction and Chi-Square feature selection were effective in supporting student depression classification using Support Vector Machine (SVM). However, PCA consistently achieved superior performance across all evaluation metrics, including Accuracy, Precision, Recall, and F1-Score. The PCA + SVM model achieved an F1-Score of 84.14%, slightly outperforming the Chi-Square + SVM model, which obtained an F1-Score of 83.65%. One possible explanation for the superior performance of PCA lies in its ability to transform the original feature space into a set of orthogonal principal components. Unlike feature selection methods that retain only a subset of the original variables, PCA captures information from all features and combines them into new components that maximize variance. As a result, important information contained in multiple correlated attributes can be preserved more effectively. This characteristic enables the SVM classifier to identify decision boundaries with improved discrimination capability.

In contrast, the Chi-Square method evaluates each feature independently based on its statistical association with the target variable. Although this approach successfully identifies the most relevant attributes, it may overlook interactions among features. Consequently, some useful information embedded in lower-ranked variables may be discarded during the feature selection process. This limitation may explain the slightly lower performance achieved by the Chi-Square-based model. The feature importance analysis further revealed that Have You Ever Had Suicidal Thoughts?, Financial Stress, and Academic Pressure were the most influential variables associated with depression. Among these features, suicidal thoughts obtained the highest Chi-Square score by a substantial margin, indicating a strong relationship with the target class. This finding is consistent with previous mental health studies reporting that suicidal ideation is closely associated with depressive symptoms and psychological distress among university students. Financial stress and academic pressure also emerged as significant predictors. University students frequently encounter challenges related to tuition fees, living expenses, academic workload, examinations, and future career uncertainty.

These stressors may contribute to emotional exhaustion and psychological vulnerability, increasing the likelihood of depression. Therefore, the identification of these factors provides valuable insights for educational institutions seeking to develop early intervention and mental health support programs. Although PCA outperformed Chi-Square, the performance difference between the two approaches was relatively small, with an accuracy improvement of approximately 0.48%. This finding suggests that the original dataset already contains a relatively compact and informative feature set. Since the dataset consists of only a limited number of predictor variables, the benefits of dimensionality reduction are naturally constrained. Nevertheless, PCA was still able to preserve feature interactions and reduce redundancy more effectively than Chi-Square, resulting in slightly better classification performance.

From a practical perspective, the findings indicate that feature engineering plays an important role in enhancing machine learning performance for student mental health prediction. The results suggest that PCA-based feature extraction is a suitable preprocessing strategy when the objective is to maximize predictive performance, whereas Chi-Square feature selection may be preferable when model interpretability is prioritized. Educational institutions and researchers can utilize these approaches to develop intelligent screening systems that assist in the early identification of students at risk of depression. Overall, this study demonstrates that PCA-based feature extraction combined with Support Vector Machine provides the most effective framework among the evaluated approaches for student depression classification. The findings contribute to the growing body of research in educational data mining and mental health analytics by highlighting the impact of feature engineering techniques on classification performance.

4. CONCLUSION

This study evaluated the effectiveness of two feature engineering techniques, namely PCA-based feature extraction and Chi-Square feature selection, for student depression classification using the Support Vector Machine (SVM) algorithm. The experiments were conducted using the Student Depression Dataset, which contains demographic, academic, lifestyle, and psychological factors associated with students' mental health conditions. The results demonstrated that both feature engineering approaches were capable of supporting depression classification with satisfactory performance. However, PCA-based feature extraction consistently outperformed Chi-Square feature selection across all evaluation metrics. The PCA + SVM model achieved the highest performance, with an Accuracy of 84.09%, Precision of 84.24%, Recall of 84.09%, and F1-Score of 84.14%. In comparison, the Chi-Square + SVM model obtained an Accuracy of 83.61%, Precision of 83.75%, Recall of 83.61%, and F1-Score of 83.65%. These findings indicate that PCA was more effective in preserving informative patterns and reducing feature redundancy, enabling the classifier to achieve better predictive performance. Furthermore, the Chi-Square analysis revealed that Have You Ever Had Suicidal Thoughts?, Financial Stress, and Academic Pressure were the most influential features associated with student depression. This finding highlights the significant role of psychological, financial, and academic factors in predicting mental health conditions among university students. Overall, the study concludes that PCA-based feature extraction is a more suitable feature engineering approach than Chi-Square feature selection for student depression classification using SVM. Future research may explore advanced hyperparameter optimization techniques, alternative classification algorithms, and larger or more diverse datasets to further improve predictive performance and model generalizability.

ACKNOWLEDGMENTS

The authors would like to thank Universitas Sains dan Teknologi Indonesia (USTI) for its support and facilities provided throughout this research. The authors also appreciate all parties who contributed directly or indirectly to the completion of this study.

REFERENCES

[1] Li, Q., Li, J., & Fan, Y. (2025). Addressing mental health in university students: a call for action. *Frontiers in Public Health*, **13**, 1614999.

[2] Chong, L. Z., Foo, L. K., & Chua, S. L. (2025). Student burnout: A review on factors contributing to burnout across different student populations. *Behavioral Sciences*, **15**(2), 170.
[3] Hainrich, H. & Raclaw, M. (2026). Analysis of Academic Burnout and Its Impact on Students' Mental Health. *Acta Psychologia*, **4**(4), 174–184.
[4] Liu, Z., Xie, Y., Sun, Z., Liu, D., Yin, H., & Shi, L. (2023). Factors associated with academic burnout and its prevalence among university students: a cross-sectional study. *BMC Medical Education*, **23**(1), 317.
[5] Tran, T. X., Vo, T. T. T., & Ho, C. (2023). From academic resilience to academic burnout among international university students during the post-COVID-19 new normal: An empirical study in Taiwan. *Behavioral Sciences*, **13**(3), 206.
[6] Zhang, J., Meng, J., & Wen, X. (2025). The relationship between stress and academic burnout in college students: evidence from longitudinal data on indirect effects. *Frontiers in Psychology*, **16**, 1517920.
[7] Geraci, A., D'Amico, A., Fernández-Berrocal, P., & Cabello, R. (2026). The relationship among burnout, emotional intelligence, and self-efficacy in school teachers. A systematic review. *International Journal of Educational Research*, **137**, 102972.
[8] Van Hoy, A., & Rzeszutek, M. (2022). Burnout and psychological wellbeing among psychotherapists: A systematic review. *Frontiers in Psychology*, **13**, 928191.
[9] Drira, M., Ben Hassine, S., Zhang, M., & Smith, S. (2024). Machine learning methods in student mental health research: An ethics-centered systematic literature review. *Applied Sciences*, **14**(24), 11738.
[10] Madububambachu, U., Ukpebor, A., & Ihezue, U. (2024). Machine learning techniques to predict mental health diagnoses: A systematic literature review. *Clinical Practice and Epidemiology in Mental Health: CP & EMH*, **20**, e17450179315688.
[11] Nath, M. D., Ahamed, M. K. U., Ahmed, O., Ahmed, T., Roy, S., & Uddin, M. N. (2025). Smart web interface for student mental health prediction using machine learning with blockchain technology. *Neuroscience Informatics*, 100236.
[12] Feng, X., Luo, J., Yang, Y., El Baz, D., & Shi, L. (2025). Health Misinformation Detection: Approaches, Challenges and Opportunities. *INQUIRY: The Journal of Health Care Organization, Provision, and Financing*, **62**, 00469580251384784.
[13] Kar, A., Nath, N., Kemprai, U. & Aman, . (2024). Performance Analysis of Support Vector Machine (SVM) on Challenging Datasets for Forest Fire Detection. *International Journal of Communications, Network and System Sciences*, **17**(02), 11–29.
[14] Zhao, X., He, F., & Li, K. (2025). Introducing an efficient method for feature extraction in image retrieval systems. *Scientific Reports*, **15**(1), 43664.
[15] Ramos-Pérez, I., Barbero-Aparicio, J. A., Canepa-Oneto, A., Arnaiz-González, Á., & Maudes-Raedo, J. (2024). An extensive performance comparison between feature reduction and feature selection preprocessing algorithms on imbalanced wide data. *Information*, **15**(4), 223.
[16] Shah, S. N. A., Issar, K., & Parveen, R. (2026). A hybrid feature extraction framework combining PCA and mutual information for gene expression based lung cancer classification. *PloS One*, **21**(2), e0342160.
[17] Grządzielewska, M. (2021). Using machine learning in burnout prediction: A survey. *Child and Adolescent Social Work Journal*, **38**(2), 175–180.
[18] Pereira, M. G., Santos, M., Magalhães, R., Rodrigues, C., Araújo, O., & Durães, D. (2025). Burnout risk profiles in psychology students: An exploratory study with machine learning. *Behavioral Sciences*, **15**(4), 505.
[19] Yeskuatov, E., Foo, L. K., & Chua, S. L. (2025, December). Detecting Burnout Among Undergraduate Computing Students with Supervised Machine Learning. *Healthcare*, **13**(23), 3182.

[20] Ayon, S. S., Al Mamun, A., Hossain, M. E., Alamro, W., Allawi, Y. M., Prova, N. N. I., ... & Abadleh, A. (2026). Explainable AI framework for improved Thalassemia mental health classification and feature selection. *PLoS One*, **21**(1), e0341168.

[21] Azevedo, B. F., Rocha, A. M. A., & Pereira, A. I. (2024). Hybrid approaches to optimization and machine learning methods: a systematic literature review. *Machine Learning*, **113**(7), 4055–4097.