Skill Gap Detection and Curriculum Optimization using PCA

1st Esha Kumari
Computer Science and Engineering
Lovely Professional University
Punjab, India
esha.12322827@lpu.in

2nd Premanand Sahu
Computer Science and Engineering
Lovely Professional University
Punjab, India
premananda.29813@lpu.co.in

***Abstract*—The increasing gap between academic skill development and industry expectations has become a major concern in engineering education. Students often graduate with theoretical knowledge but lack the practical skills required for modern technical roles, leading to reduced employability and performance in real-world environments. Detecting this skill gap early can significantly improve the effectiveness of training programs and curriculum design. This research proposes the use of Principal Component Analysis (PCA) as a statistical and machine-learning-based approach to identify hidden patterns in student skill datasets and highlight areas where students fall short of industry standards. PCA is applied to a multidimensional dataset consisting of technical and soft skill scores, allowing high-dimensional data to be transformed into a smaller set of principal components that capture the maximum variance. These components reveal relationships between different skills, classify skill clusters, and allow for the comparison between student performance and industry skill benchmarks. By examining component loadings and visualizing the distribution of student scores in the transformed PCA space, this study identifies significant gaps in programming, algorithmic thinking, and problem-solving abilities among students. The results demonstrate that PCA not only simplifies complex assessments but also provides an interpretable framework for skill gap analysis. The findings highlight the potential for integrating PCA-based analysis into academic evaluation systems, teacher dashboards, and personalized learning recommendations. This research therefore contributes to the field of educational analytics by offering a data-driven, scalable, and efficient method for detecting skill deficiencies, which can guide institutions in designing more industry-aligned training modules and improving student readiness for employment.**

***Index Terms*—Skill Gap Detection, Principal Component Analysis, Educational Analytics, Curriculum Optimization, Dimensionality Reduction**

I. Introduction

In recent years, the demand for industry ready graduates has increased significantly as technology continues to evolve at a rapid pace. Companies expect students to possess not only strong theoretical knowledge but also practical skills such as programming efficiency, analytical thinking, teamwork, and problem-solving. However, multiple studies and placement reports show a noticeable mismatch between what students learn in academic environments and what industries actually require. This mismatch is widely referred to as the skill gap.

Skill gaps can occur for several reasons. Curricula often become outdated compared to the speed at which technology changes, students may not get enough exposure to hands-on learning, and certain essential soft skills are not evaluated consistently. Traditional methods of identifying these gaps such as manual observation, classroom assessments, or feedback from instructors tend to be slow, subjective, and limited in accuracy. Hence, educational institutions are shifting towards data-driven and machine learning-based evaluation systems that offer more reliable and scalable insights.

Among various analytical methods, Principal Component Analysis (PCA) stands out because of its ability to simplify complex datasets while still preserving essential information. In a typical skill assessment scenario, students are evaluated on many different skills, creating a high-dimensional dataset. PCA helps transform this dataset into a smaller number of principal components that still explain most of the variability. This allows patterns to become clearer for example, how students cluster based on strengths, which skills tend to correlate, and which components represent the major differences between student performance and industry expectations.

In this research, PCA is used as the core method for skill gap detection. By comparing the principal component patterns of student skills with industry benchmarks, the study identifies which areas show the highest level of deviation. The goal is to help institutions understand where students are lagging, so they can update teaching methods, redesign curriculum modules, or provide targeted training programs.

This paper aims to demonstrate that PCA is not only a mathematical dimensionality reduction tool but also a valuable educational analytics technique. Ultimately, this work contributes to creating a more effective, objective, and data-supported approach to preparing students for future careers.

II. Literature Review

Principal Component Analysis (PCA) has been widely recognized as an effective approach for managing high-

dimensional datasets. Early theoretical work established PCA as a mathematical framework that transforms correlated variables into a reduced set of orthogonal components while retaining the majority of data variance. This transformation enables simplified analysis of complex datasets, making PCA particularly suitable for educational data that involve multiple interdependent skill attributes [1], [2].

The applicability of PCA in practical multivariate analysis has been demonstrated across various domains. Prior studies have shown that PCA can uncover hidden relationships and structural patterns that are difficult to observe in the original feature space. By reducing redundancy and noise, PCA improves both data interpretability and analytical robustness, especially when dealing with large-scale datasets [3].

In the context of engineering and computing education, several studies have reported a growing mismatch between student competencies and industry expectations. Research examining learning styles and curriculum effectiveness has highlighted persistent gaps in technical knowledge, analytical reasoning, and practical problem-solving abilities. These findings suggest that traditional assessment methods are often insufficient for capturing multidimensional skill deficiencies [4], [5].

To address these limitations, researchers have increasingly adopted machine learning techniques for skill gap analysis. Systematic reviews in this area indicate that dimensionality reduction methods play a critical role in improving model efficiency, scalability, and transparency. PCA, in particular, has been shown to enhance the performance of predictive and evaluative models by eliminating irrelevant or highly correlated features [6], [7].

Recent studies have applied PCA-based skill profiling to assess student performance in technical courses. By projecting student data into a lower-dimensional space, these approaches enable clearer differentiation of learner capabilities and learning patterns. Such methods have proven effective in preserving meaningful performance information while reducing computational complexity [8].

Data mining research further supports the integration of PCA as a preprocessing step before applying clustering or classification algorithms. Feature reduction has been shown to improve cluster cohesion and prediction accuracy, particularly in educational datasets characterized by high dimensionality and overlapping attributes [9], [10].

Several empirical studies combining PCA with clustering techniques have reported enhanced skill grouping and more consistent assessment outcomes. The removal of redundant variables contributes to improved cluster quality and facilitates clearer identification of student strengths and weaknesses, supporting data-driven curriculum planning [11], [12].

From an implementation perspective, widely used machine learning libraries provide efficient and reliable tools for applying PCA within educational analytics workflows. Additionally, decision-making frameworks and global workforce reports emphasize the growing importance of analytical skill assessment models for aligning academic training with evolving labor market demands [13]-[15].

III. Methodology

A. Flowchart

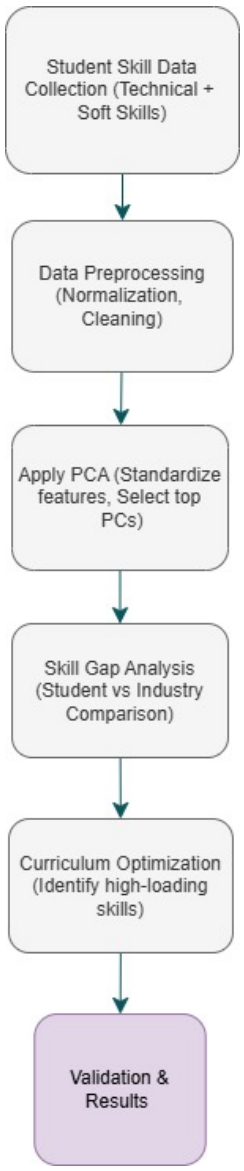

Fig. 1. Proposed methodological flowchart for skill gap detection and curriculum optimization using Principal Component Analysis (PCA)

B. Data Collection

i. Student skill data is collected from academic records, assessments, and self-evaluation surveys.
ii. The dataset includes both technical skills (such as programming, algorithms, databases, and analytical ability) and soft skills (such as communication, teamwork, and problem-solving).
iii. Each skill is represented numerically on a common scale to allow mathematical analysis.
iv. Alongside student data, industry skill benchmark values are defined to represent expected proficiency levels required for employability.

C. Data Preprocessing

i. The collected dataset is first examined for missing or inconsistent values.
ii. Missing values are handled using mean imputation to maintain dataset balance.
iii. All skill attributes are standardized using normalization techniques to ensure equal contribution of each skill.
iv. Outliers are identified and treated to prevent distortion during PCA computation.

D. Feature Representation

i. Preprocessed data is transformed into a structured skill matrix, where rows represent students and columns represent individual skills.
ii. This matrix serves as the input for PCA.
iii. Converting raw skill data into matrix form allows efficient mathematical processing and comparison.

E. Mathematical Formulation of PCA

Let $X \in \mathbb{R}^{n \times p}$ denote the student–skill matrix, where n represents the number of students and p denotes the number of skill attributes. Each row of X corresponds to a student profile, while each column represents a specific skill.

Prior to applying PCA, the data is standardized using z-score normalization to ensure that all skills contribute equally to the analysis. The standardized value z_{ij} of skill j for student i is computed as:

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j} \tag{1}$$

where:

- x_{ij} denotes the original score of student i for skill j, and
- μ_j represent the mean
- σ_j standard deviation of skill j, respectively.

After standardization, the covariance matrix Σ of the normalized data matrix Z is calculated as:

$$\Sigma = \frac{1}{n-1} Z^T Z \tag{2}$$

Principal components are obtained by computing the eigenvalues and eigenvectors of the covariance matrix. The standardized data is then projected onto the principal component space using:

$$Y = ZW \tag{3}$$

where W is the matrix of eigenvectors corresponding to the largest eigenvalues, and Y represents the transformed data in the reduced-dimensional PCA space.

F. Dimensionality Reduction using PCA

i. Principal Component Analysis is applied to reduce the dimensionality of the skill matrix while retaining maximum variance.
ii. The covariance matrix of the standardized data is computed.
iii. Eigenvalues and eigenvectors are extracted to identify principal components.
iv. The top principal components explaining the majority of variance are selected.
v. Student skill data is projected into the reduced PCA space, simplifying complex multi-skill relationships.

G. Skill Gap Detection

i. Skill gap analysis is performed by comparing student skill representations, informed by PCA-derived components, against predefined industry benchmark proficiency levels.
ii. The difference between these component values indicates the magnitude of skill gaps.
iii. Each principal component is analyzed to understand which original skills contribute most to the observed gaps.
iv. Larger deviations represent critical areas requiring academic or training intervention.

H. Curriculum Optimization and Recommendation

i. Based on the detected skill gaps, targeted curriculum improvements are suggested.
ii. Students with similar gap patterns are grouped to enable focused training modules.
iii. Recommendations may include additional coursework, practical labs, industry-aligned projects, or skill-specific workshops.
iv. This step ensures that academic learning outcomes are better aligned with industry requirements.

I. Evaluation and Interpretation

i. The effectiveness of the PCA-based approach is evaluated by observing the consistency of principal components across different student batches.
ii. Visualizations such as scatter plots and gap charts are used to interpret results.
iii. Faculty feedback and academic observations are used to validate the practical relevance of detected gaps.
iv. The results confirm that the proposed methodology provides a structured and scalable approach to skill gap detection.

IV. Results and Discussion

This section presents the outcomes of applying Principal Component Analysis (PCA) to the student skill dataset and discusses the implications for skill gap detection and curriculum optimization. The analysis focuses on understanding how multidimensional skill attributes are transformed into meaningful principal components and how these components reflect underlying competency patterns

among students.

By examining explained variance, component loadings, and PCA-based representations, this section highlights key trends in student performance and identifies areas where deviations from industry benchmarks are most prominent. The results not only demonstrate the effectiveness of PCA in reducing data complexity but also provide actionable insights that can support data-driven academic decision-making and curriculum enhancement strategies.

A. PCA Variance Analysis

PCA was applied to the standardized skill matrix to reduce dimensionality while retaining the most significant information. The cumulative explained variance plot indicates that a small number of principal components capture the majority of variance present in the dataset. This confirms the presence of strong correlations among several skill attributes and validates the effectiveness of PCA for dimensionality reduction in this context.

The first principal component accounts for the highest variance and primarily represents foundational technical competencies, while subsequent components capture more specialized and advanced skill dimensions.

TABLE I
SAMPLE PCA-BASED SKILL REPRESENTATION AND GAP ANALYSIS

Student ID	Skill Profile (Input)	Dominant PCA Components	Gap Level	Method
1	Programming, Databases, Data Structures	PC1 (0.72), PC2 (0.31)	Low	PCA
2	Programming, ML, Cloud Computing	PC1 (0.65), PC3 (0.44)	Medium	PCA
3	ML, DevOps, System Design	PC2 (0.70), PC3 (0.52)	High	PCA
4	Programming, Web Development	PC1 (0.81), PC2 (0.28)	Low	PCA
5	Cloud, DevOps, ML	PC3 (0.76), PC2 (0.41)	High	PCA

B. Interpretation of Principal Components

Analysis of PCA loadings reveals that core skills such as programming fundamentals, data structures, and database systems contribute strongly to the leading principal component. This suggests that these skills collectively form a dominant competency dimension shared across most learners.

In contrast, advanced skills including machine learning, cloud computing, DevOps, and system design exhibit higher loadings in later principal components. This indicates increased variability and specialization among students, highlighting areas where proficiency levels differ significantly across the population.

C. Skill Gap Detection

Skill gap analysis was performed by comparing average student proficiency levels against predefined industry benchmark targets. The results demonstrate notable gaps in advanced technical skills, particularly in machine learning, cloud computing, DevOps, and system design. These gaps suggest that while students generally possess adequate foundational skills, there is a need for focused improvement in emerging and industry-relevant competencies. The PCA-informed representation enables clearer identification of these gaps by reducing redundancy and emphasizing the most influential skill dimensions.

D. Analytical Insights from PCA-Based Results

The PCA-based analysis provides deeper analytical insights beyond simple skill-wise comparisons by capturing latent relationships among multiple competencies. Unlike traditional evaluation approaches that assess individual skills independently, the PCA framework aggregates correlated attributes into principal components, enabling a more holistic understanding of student capability patterns.

The dominant principal components reveal how foundational and advanced skills interact to influence overall student performance. For instance, students exhibiting strong loadings in components associated with core technical skills demonstrate consistent competency across multiple domains, whereas lower component scores indicate fragmented or uneven skill development. Such patterns are difficult to identify using conventional score-based evaluation methods.

Additionally, the PCA-transformed space facilitates clearer differentiation between students with similar aggregate scores but differing skill compositions. This distinction is particularly valuable for identifying hidden gaps in advanced or emerging technologies that may not be apparent from raw performance metrics alone. By examining component contributions, educators can distinguish between students who require foundational reinforcement and those who need targeted exposure to specialized skills.

These analytical insights serve as a critical bridge between numerical PCA outcomes and visual statistical representations. They ensure that the subsequent statistical graphs are interpreted within a meaningful analytical context, strengthening the reliability and interpretability of the results presented in the following subsection.

E. Evaluation Metrics and Validation

The effectiveness of the proposed framework is evaluated using both statistical and structural validation criteria. The explained variance ratio is analyzed to ensure that the reduced PCA representation retains sufficient informational content from the original feature space. Additionally, internal clustering validation metrics, including silhouette score, are employed to assess the cohesion and separation of learner groups in the reduced-dimensional space. These metrics confirm that PCA enhances interpretability while supporting stable and meaningful cluster formation. It is emphasized that the

proposed framework is intended for exploratory skill analysis rather than predictive classification, aligning with best practices in educational data mining.

F. Statistical Data

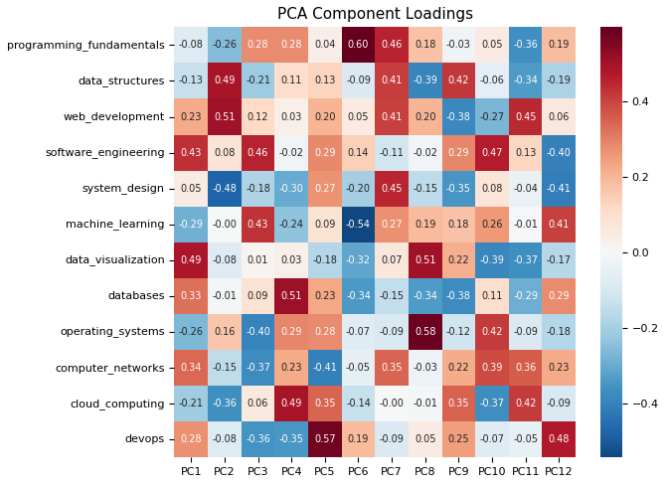

Fig. 2. Heatmap representation of principal component loadings, indicating which skills play a stronger role in each component.

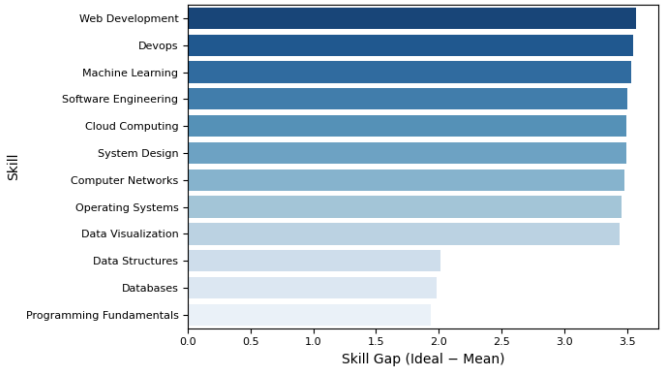

Fig. 3. Skill gap analysis illustrating the difference between ideal proficiency levels and average student performance.

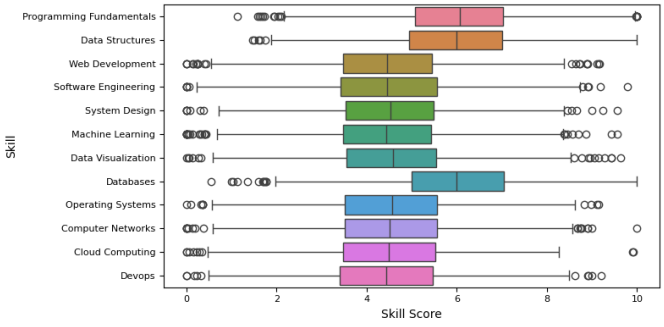

Fig. 4. Boxplot showing the distribution of student skill scores across different technical domains.

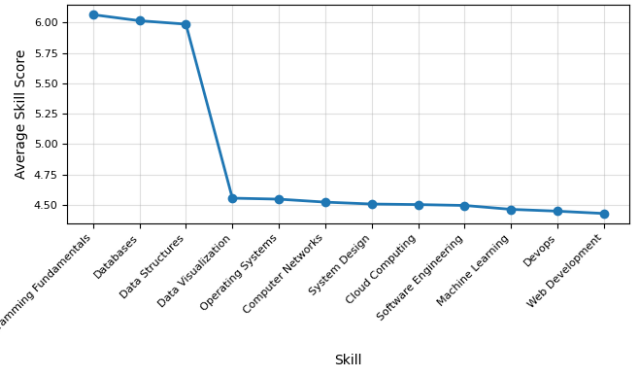

Fig. 5. Line graph depicting average student performance trends across technical skills, highlighting variations in proficiency levels.

V. Conclusion

This research presented a systematic framework for skill gap detection and curriculum optimization using Principal Component Analysis (PCA). The primary objective was to analyze high-dimensional student skill data and extract meaningful latent patterns that reflect underlying competency structures. By applying PCA, the dimensionality of the dataset was effectively reduced while retaining the majority of informative variance, thereby simplifying complex multi-skill relationships without loss of critical information.

The experimental analysis demonstrated that student performance is influenced by multiple skill dimensions rather than a single dominant factor. The explained variance analysis confirmed the suitability of PCA for this domain, while the component loading visualization provided interpretable insights into the contribution of individual skills to each principal component. Furthermore, the skill gap analysis revealed noticeable disparities between average student proficiency levels and predefined benchmark values, highlighting specific technical domains that require focused academic and training interventions.

Overall, the results validate the effectiveness of the proposed PCA-based framework as an exploratory and data-driven method for identifying skill gaps. Unlike traditional assessment approaches that evaluate skills in isolation, this methodology captures holistic competency patterns and supports scalable analysis across large student populations. The findings demonstrate the potential of PCA to assist educational institutions in aligning curriculum design with evolving industry requirements through objective and interpretable analysis.

Despite the strengths of the proposed approach, several limitations must be acknowledged. The analysis is based on static skill snapshots and does not capture temporal evolution of student competencies. Additionally, benchmark proficiency targets are defined using assumed thresholds rather than direct employer feedback, which may influence the magnitude of detected gaps. The study is further limited by dataset scope, as results may vary across institutions with different curricular structures. Finally, while PCA enhances interpretability, it does not model causal relationships between skills and employa-

bility outcomes. These limitations highlight opportunities for methodological extensions and broader validation.

A. Future Scope

The proposed framework can be extended in several directions to enhance its applicability and practical impact. Future work may integrate supervised learning models with PCA-transformed features to enable employability prediction and quantitative evaluation using standard classification metrics. Incorporating longitudinal student data could further support the analysis of skill progression over time, allowing institutions to assess the effectiveness of curriculum modifications and training programs. In addition, expanding the dataset to include students from multiple institutions and diverse academic backgrounds may improve the robustness and generalizability of the findings. The framework can also be enhanced by incorporating soft skills, industry feedback, and adaptive learning recommendations to enable personalized curriculum pathways. These extensions would strengthen the role of data-driven approaches in curriculum planning and contribute to improved workforce readiness.

REFERENCES

[1] A. Jolliffe, Principal Component Analysis, 2nd ed. New York, NY, USA: Springer, 2002.

[2] I. T. Jolliffe and J. Cadima, "Principal component analysis: A review and recent developments," Philosophical Transactions of the Royal Society A, vol. 374, no. 2065, pp. 1–16, 2016.

[3] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," Chemometrics and Intelligent Laboratory Systems, vol. 2, no. 1–3, pp. 37–52, 1987.

[4] R. M. Felder and R. Brent, "Understanding student learning styles and skill gaps in engineering education," Journal of Engineering Education, vol. 94, no. 1, pp. 57–72, 2005.

[5] N. C. Brown and S. Sentance, "Computing education: Skills gaps, curriculum design, and learner assessment," ACM Transactions on Computing Education, vol. 19, no. 4, pp. 1–28, 2019.

[6] A. S. R. Srinivas and S. P. Kumar, "Skill gap analysis using machine learning techniques: A systematic review," International Journal of Advanced Computer Science and Applications, vol. 11, no. 7, pp. 98–106, 2020.

[7] M. N. Akhtar, H. Hassan, and T. R. Sipra, "Data-driven workforce analytics for skill gap detection in higher education," Education and Information Technologies, vol. 27, pp. 1567–1584, 2022.

[8] P. Guo, "Understanding student performance in programming courses through skill profiling and PCA-based analysis," IEEE Transactions on Education, vol. 64, no. 3, pp. 210–218, 2021.

[9] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. Burlington, MA, USA: Morgan Kaufmann, 2011.

[10] R. R. Reddy and A. S. Rao, "Predicting educational outcomes using machine learning and PCA," Procedia Computer Science, vol. 172, pp. 468–474, 2020.

[11] S. Mishra and R. S. Saini, "Skill-gap identification using clustering and dimensionality reduction techniques," IEEE Access, vol. 9, pp. 98,124–98,135, 2021.

[12] R. Aggarwal and S. Kaur, "A PCA and K-means based skill assessment model for technical students," International Journal of Engineering Research and Technology, vol. 8, no. 6, pp. 122–127, 2019.

[13] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[14] T. L. Saaty, "Decision-making in complex environments: Application to competency measurement," European Journal of Operational Research, vol. 199, no. 3, pp. 867–872, 2009.

[15] World Economic Forum, "The future of jobs report: Skills evolution and global workforce trends," WEF, Geneva, Switzerland, Rep., 2023.