The University of Tennessee at Chattanooga
Computer Science and Engineering

Model Analysis and Simulation

CPSC5410

Instructor : Dr. Tomojit Ghosh

Final Project

Group 1

Team members:

CBJ988 J.M. Imtinan Uddin

May 6, 2024

1 Introduction of VAE for augmentation on High Dimensional Biological Data

In the field of bioinformatics, analyzing high-dimensional biological data can be notably challenging. These challenges are often due to the small sample sizes available and the complex nature of the data, which can limit the performance of traditional machine learning models. To enhance the accuracy and effectiveness of these models, our study explores the application of a Variational Autoencoder (VAE), a type of generative model that can create new, synthetic data samples that closely resemble the original data.

Data augmentation, the process of generating new data samples to expand the training dataset, is well-known for improving model performance in various fields. By applying a VAE, we aim to produce synthetic data that maintains the intricate patterns and structures inherent in high-dimensional biological datasets. This augmented data could potentially provide a more robust training environment for machine learning models, thereby improving their ability to generalize to new, unseen data.

In this study, we employed several configurations of data to train and test our models:

1. **Original Data Only:** Models were trained and tested using only the data originally available.

2. **Combined Data (2n):** Models were trained on a dataset that included both the original and an equal amount of synthetic data.

3. **Combined Data (4n):** Models were trained on a dataset where synthetic data doubled the amount of the original data.

4. **Augmented Data (n):** Models were also trained on datasets composed entirely of synthetic data, in amounts equal to the original dataset size, to evaluate the efficacy of the VAE-generated data in isolation.

5. **Augmented Data (2n):** Models were also trained on datasets composed entirely of synthetic data, in amounts double to the original dataset size, to evaluate the efficacy of the VAE-generated data in isolation.

Before training our classifiers, we applied Principal Component Analysis (PCA) to reduce the dimensionality of our datasets. This step was crucial for simplifying the data and enabling more efficient and insightful visual and statistical analysis. PCA was performed in both two-dimensional and three-dimensional spaces to facilitate a clear visualization of the data distribution and clustering.
For the classification tasks, we selected two widely used algorithms: Random Forest and K-Nearest Neighbors (KNN). These models were chosen for their robustness and widespread applicability in handling complex datasets. By applying these classifiers to the different data configurations, we aimed to rigorously assess the impact of VAE-generated synthetic data on model performance across various training scenarios.The primary goal of our investigation is to determine whether VAE-based data augmentation can significantly improve the predictive accuracy of machine learning models dealing with high-dimensional biological data. Through this study, we aim to shed light on the practical benefits and potential limitations of using synthetic data augmentation in the realm of bioinformatics, providing valuable insights into its effectiveness and applicability.

1.1 VAE Augmentation using GLIOMA and GLIOMA_SLCE Dataset

1.1.1 GLIOMA

Code Link

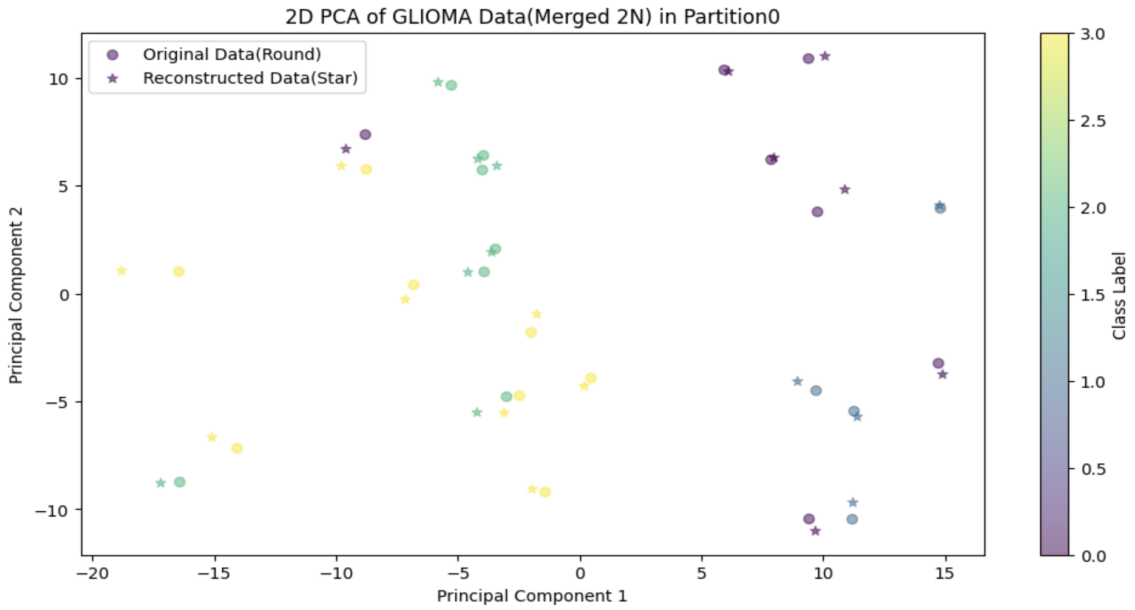

Figure 1: 2D GLIOMA COMBINED 2N(N ORIGINAL + N AUGMENTED)

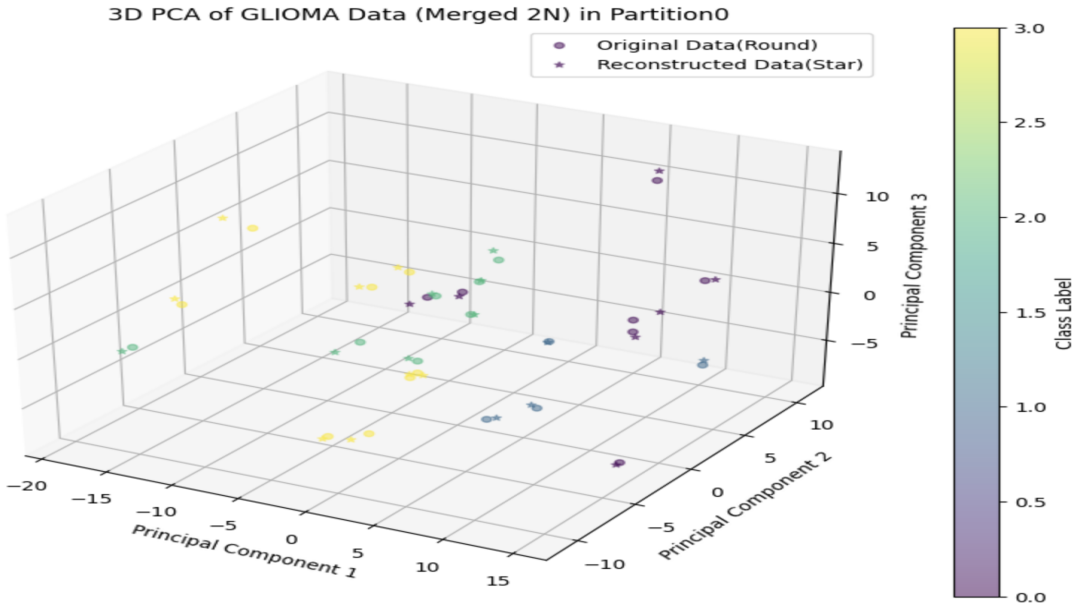

Figure 2: 3D GLIOMA COMBINED 2N (N ORIGINAL + N AUGMENTED)

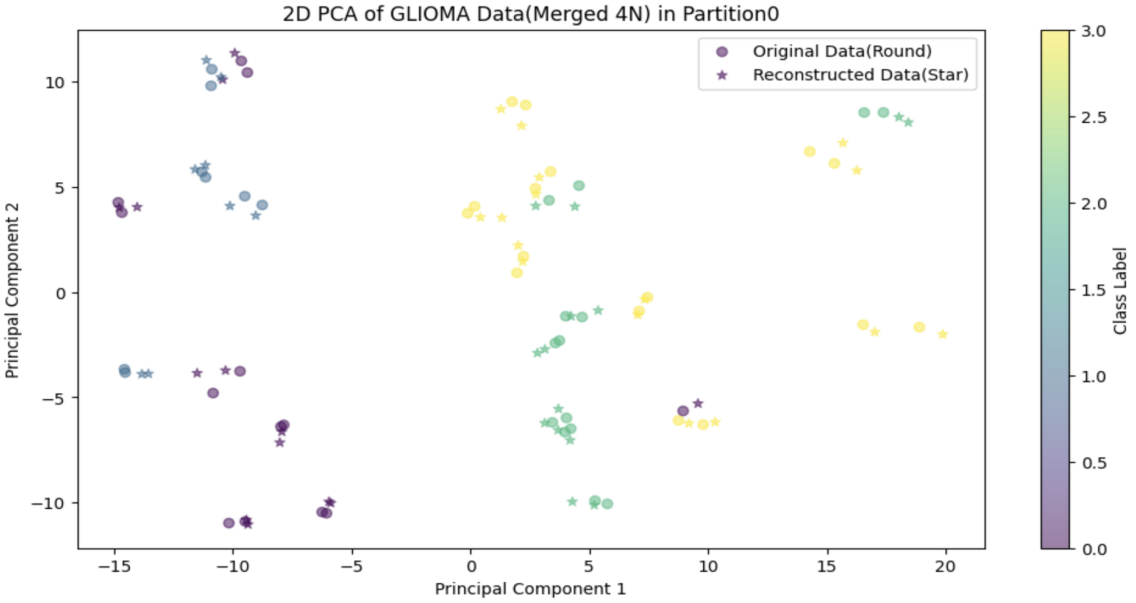

Figure 3: 2D GLIOMA COMBINED 4N SAMPLES(2N ORIGINAL + 2N AUGMENTED)

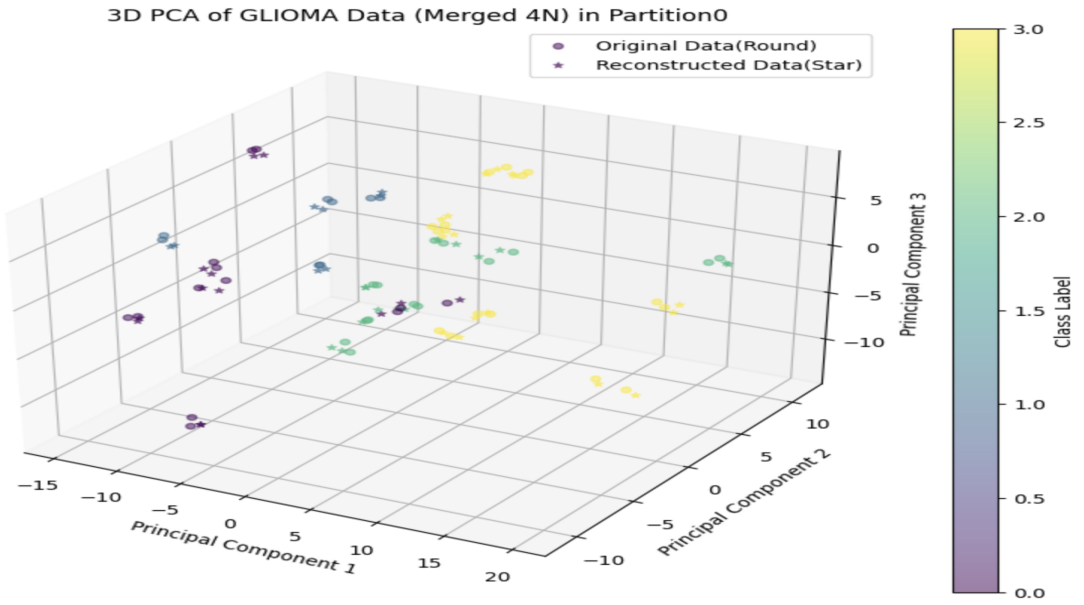

Figure 4: 3D GLIOMA COMBINED 4N SAMPLES(2N ORIGINAL + 2N AUGMENTED)

1.1.2 GLIOMA SLCE

Code Link

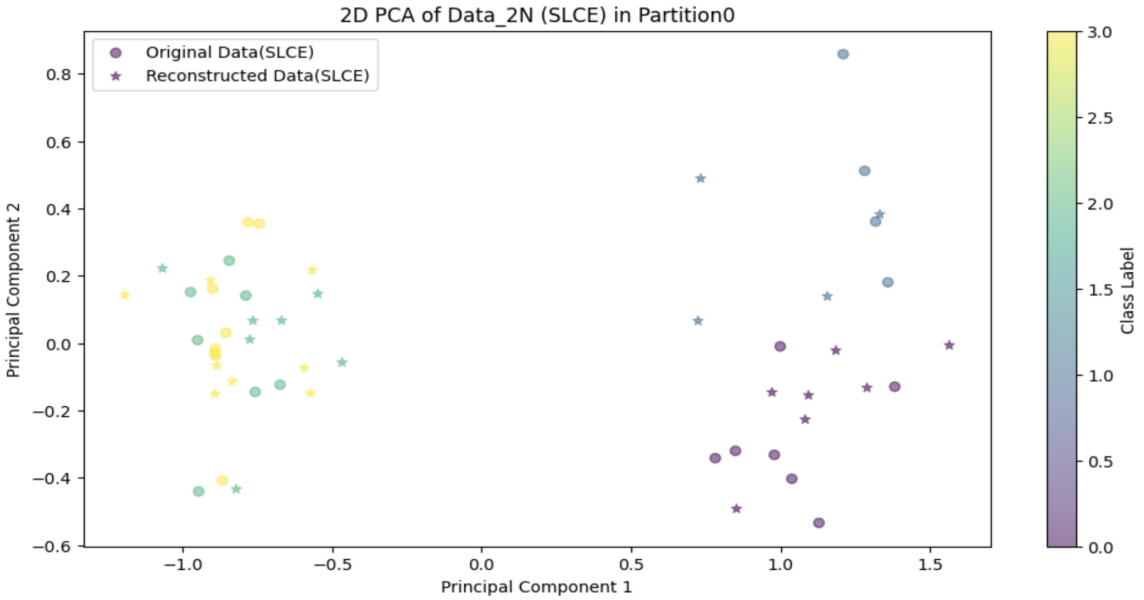

Figure 5: 2D GLIOMA SLCE 2N COMBINED(N ORIGINAL + N AUGMENTED)

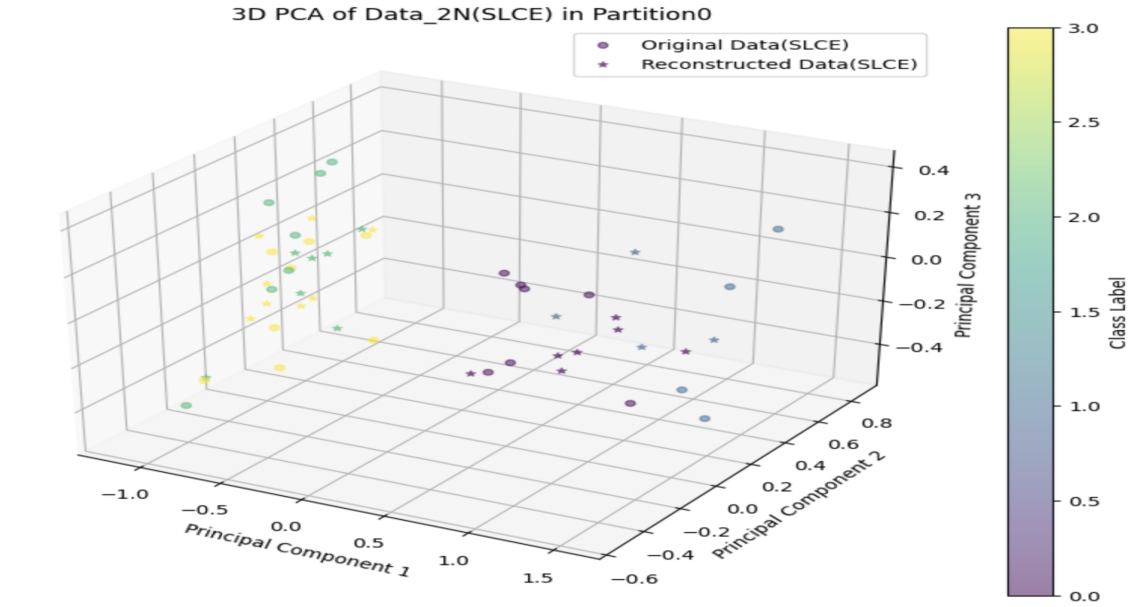

Figure 6: 3D GLIOMA SLCE 2N COMBINED(N ORIGINAL + N AUGMENTED)

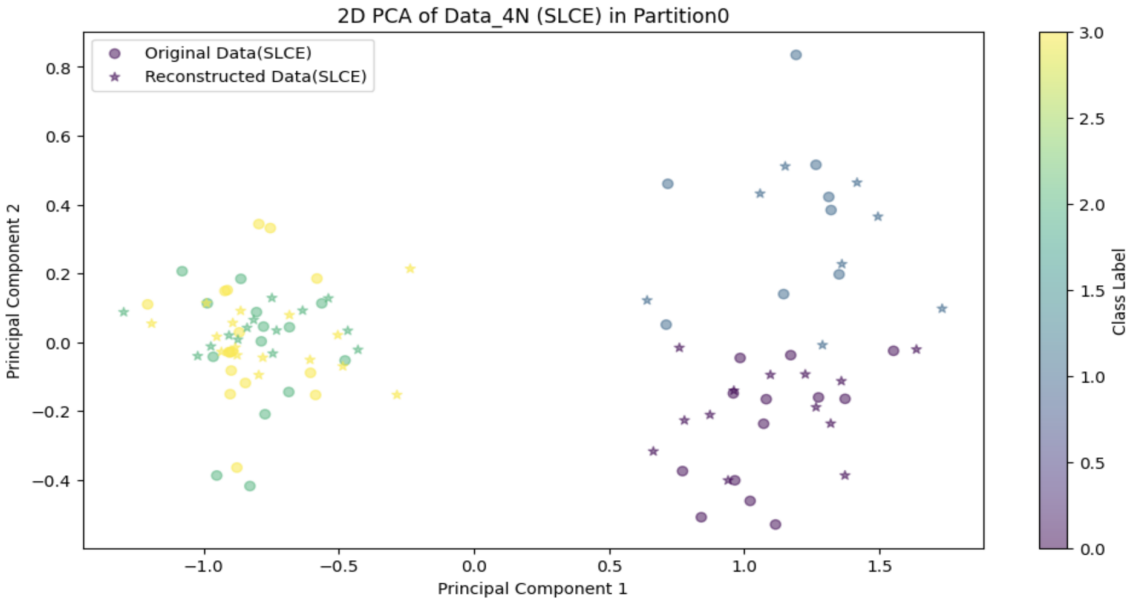

Figure 7: 2D GLIOMA SLCE 4N COMBINED (2N ORIGINAL + 2N AUGMENTED)

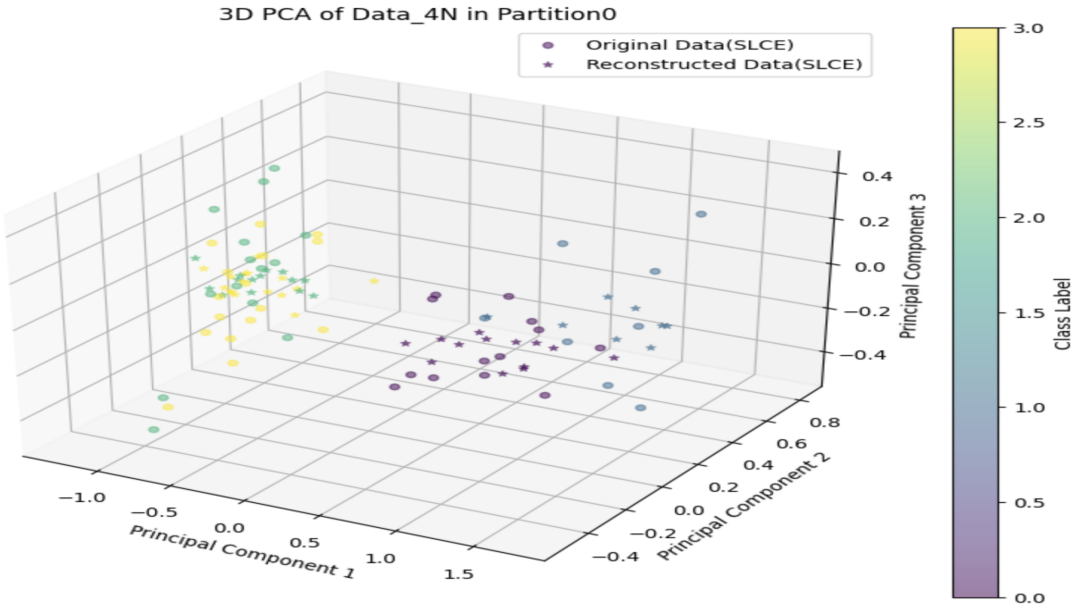

Figure 8: 3D GLIOMA SLCE 4N COMBINED (2N ORIGINAL + 2N AUGMENTED)

Table 1: Mean Accuracy of RF and KNN Models Across Different Augmentation Datasets of GLIOMA

Dataset Name	Mean RF Accuracy (%)	Mean KNN Accuracy (%)
Combined 2n(n synthetic + n original)	74.17	70.83
Combined 4n (2n synthetic + 2n original)	74.79	70.00
Original Data n (only original)	72.92	69.79
Augmented 2n (only synthetic)	73.33	71.88
Augmented n (only synthetic)	72.50	68.96

Table 2: Mean Accuracy of RF and KNN Models Across Different Augmentation Datasets of GLIOMA_SLCE

Dataset Name	Mean RF Accuracy (%)	Mean KNN Accuracy(%)
Combined 2n SLCE (n synthetic + n original)	69.37	66.87
Combined 4n SLCE (2n synthetic + 2n original)	68.12	67.29
Original Data(n) SLCE (n original)	66.25	69.37
Augmented 2n SLCE (2n synthetic)	61.88	63.75
Augmented n SLCE (n synthetic)	62.71	62.92

1.1.3 Effects of VAE Augmentation on GLIOMA and GLIOMA_SLCE Dataset

The scatter plots generated from the Principal Component Analysis (PCA) of both GLIOMA and GLIOMA_SLCE data show clear distinctions between the original and reconstructed datasets. In the 2D PCA visualizations, the spread of data points indicates that while there is some overlap, the original and reconstructed data maintain distinct groupings. This suggests that the reconstruction process maintains some degree of the original data's variability, which is crucial for preserving the dataset's integrity.

When observing the 3D PCA plots, we can see that the additional dimension provides a deeper insight into the data structure, revealing clusters that were not as apparent in the 2D representations. The complexity of the data seems to be captured well in these 3D plots, which might help in better understanding the underlying patterns.

From the tables provided, we notice a consistent pattern in the mean accuracy percentages across different models and datasets. The Random Forest (RF) model generally shows a higher mean accuracy than the K-Nearest Neighbors (KNN) model across the datasets. The combined datasets with both synthetic and original data (Combined 2n and 4n) demonstrate higher accuracy in comparison to the datasets with only original or only synthetic data. This could imply that a mix of original and synthetic data helps improve model performance by providing a richer set of samples for training.

Specifically for the GLIOMA_SLCE data, the highest accuracy is achieved with the original dataset when using the KNN model, which indicates that for this specific scenario, the KNN model might be capturing the nuances of the original data better than the RF model.

In conclusion, the integration of synthetic data with original data seems to provide an enhancement in model performance, with the RF model slightly outperforming the KNN model in most cases. These findings can serve as a valuable insight into the use of synthetic data augmentation in improving machine learning models, especially in complex domains such as medical diagnosis where data is often limited.

1.1.4 Summary of VAE Augmentation on GLIOMA and GLIOMA_SLCE Dataset

- The PCA scatter plots for GLIOMA and GLIOMA_SLCE data reveal a distinction between original and reconstructed datasets, with each maintaining unique groupings.
- In the 3D PCA visualizations, additional dimensionality aids in uncovering more complex data structures, suggesting a good capture of the dataset's variability.
- The Random Forest (RF) model consistently shows higher mean accuracy than the K-Nearest Neighbors (KNN) model across various datasets.
- Datasets combined with synthetic and original data present improved model accuracies, indicating that a mixture of data types enriches the training sample pool.
- For GLIOMA_SLCE data, the KNN model achieves the highest accuracy with the original dataset, suggesting its potential in capturing the original data's

characteristics.

- Overall, synthetic data augmentation appears beneficial in enhancing machine learning model performance, especially in the RF model, and holds particular promise in fields with limited data like medical diagnostics.

1.2 VAE Augmentation using CLL_SUB and CLL_SUB_SLCE Dataset

1.2.1 CLL_SUB

Code Link

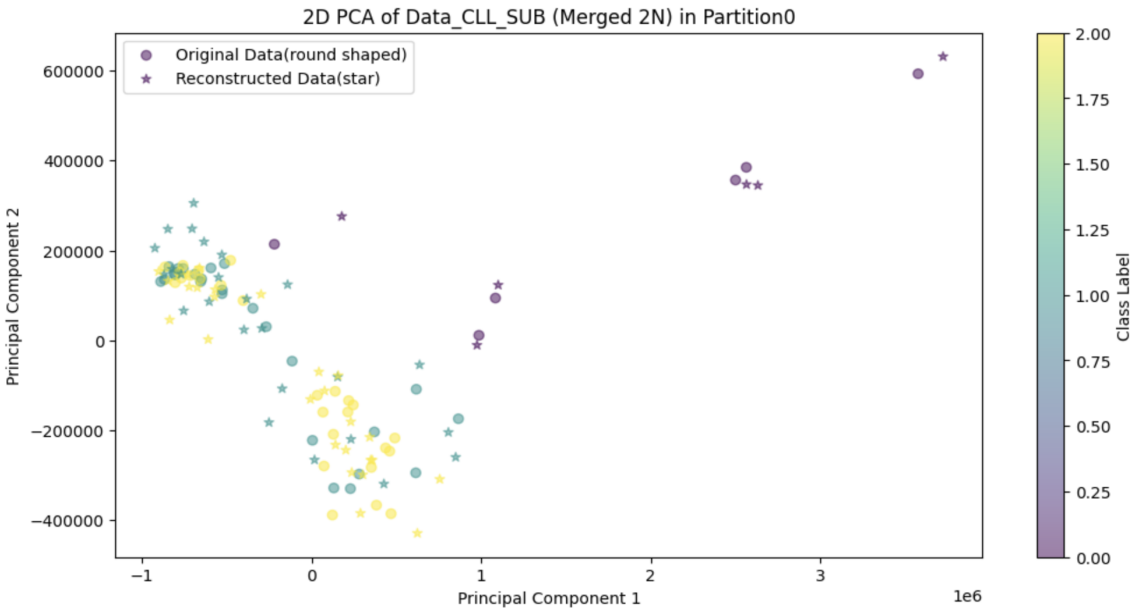

Figure 9: 2D CLL_SUB COMBINED 2N(N ORIGINAL + N AUGMENTED)

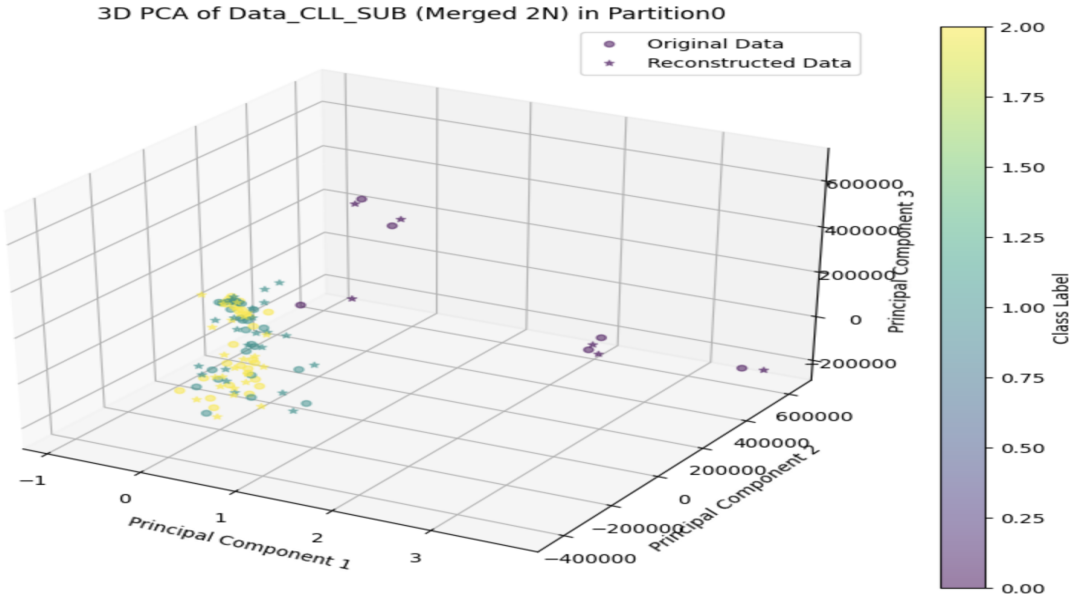

Figure 10: 3D CLL_SUB COMBINED 2N (N ORIGINAL + N AUGMENTED)

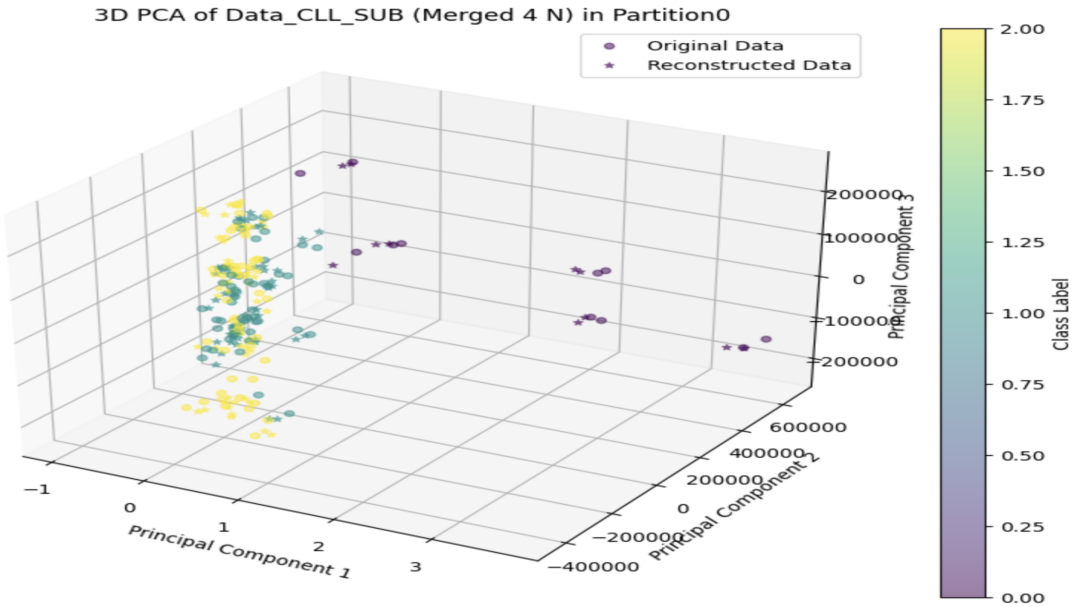

Figure 11: 2D CLL_SUB COMBINED 4N SAMPLES(2N ORIGINAL + 2N AUGMENTED)

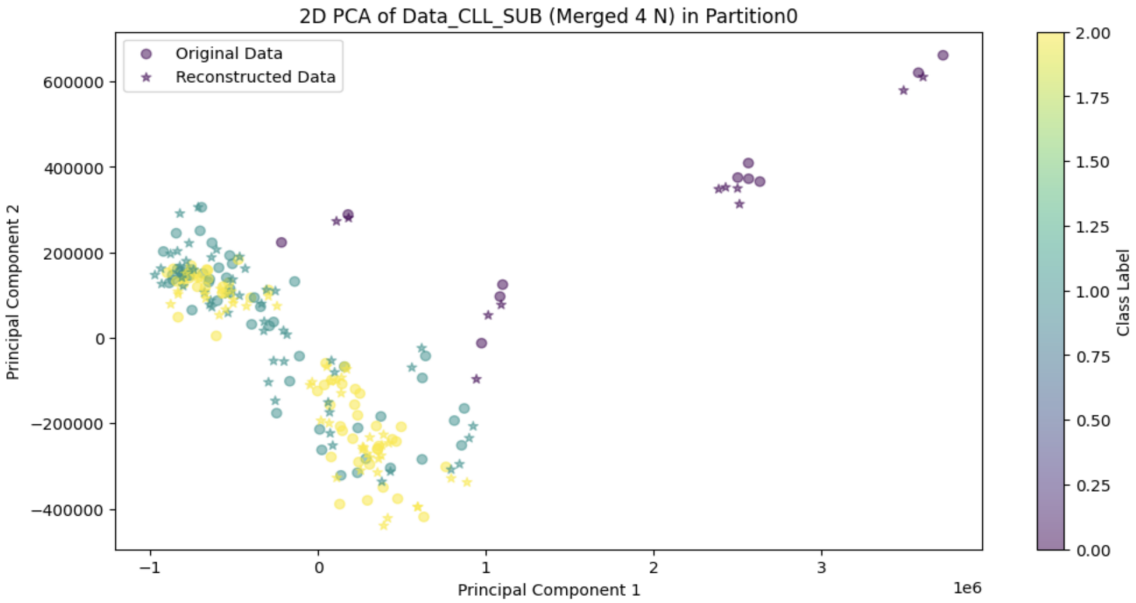

Figure 12: 3D CLL_SUB COMBINED 4N SAMPLES(2N ORIGINAL + 2N AUGMENTED)

1.2.2 CLL_SUB SLCE

Code Link

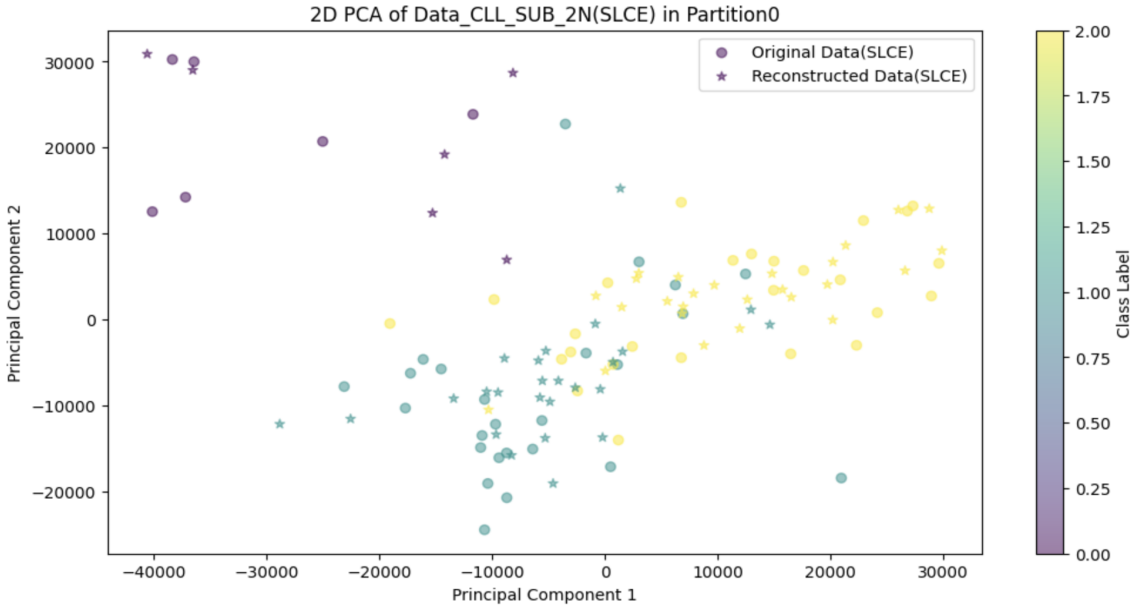

Figure 13: 2D CLL_SUB SLCE COMBINED 2N(N ORIGINAL + N AUGMENTED)

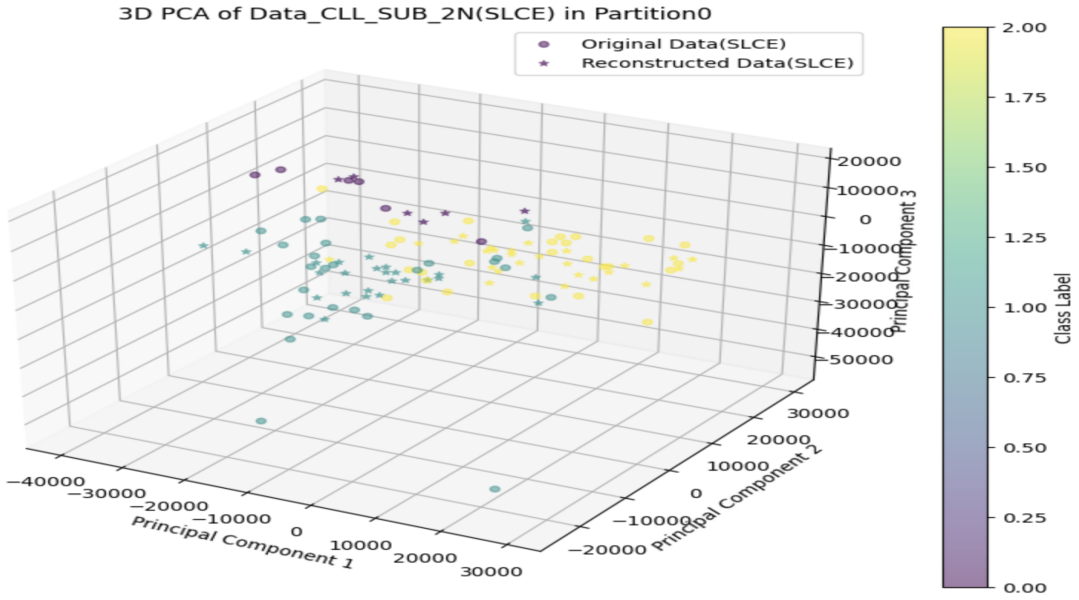

Figure 14: 3D CLL_SUB SLCE COMBINED 2N (N ORIGINAL + N AUGMENTED)

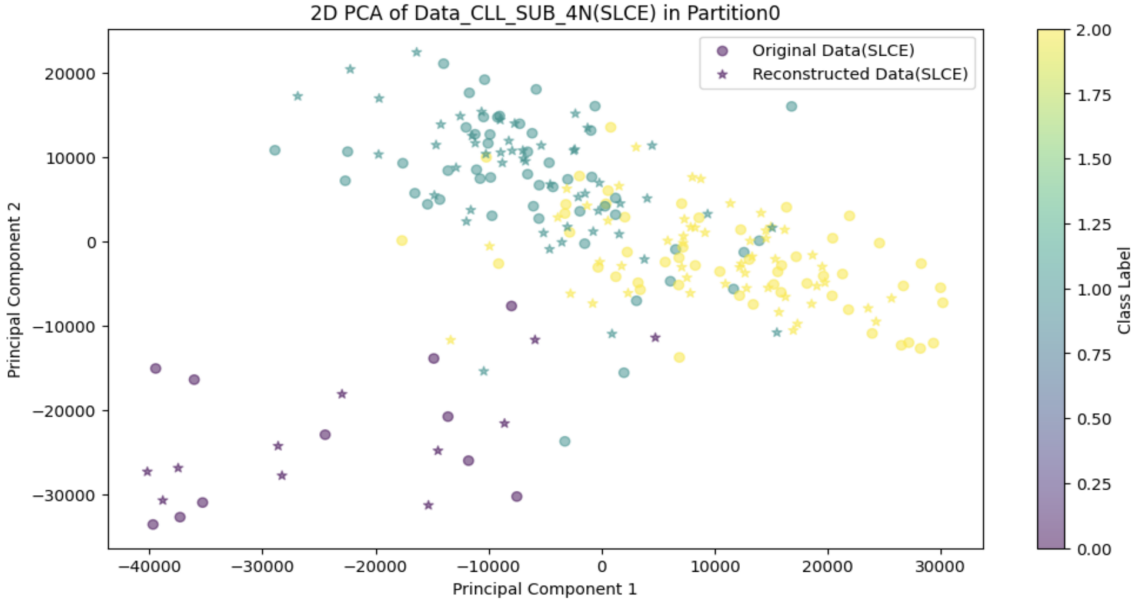

Figure 15: 2D CLL_SUB SLCE COMBINED 4N SAMPLES(2N ORIGINAL + 2N AUGMENTED)

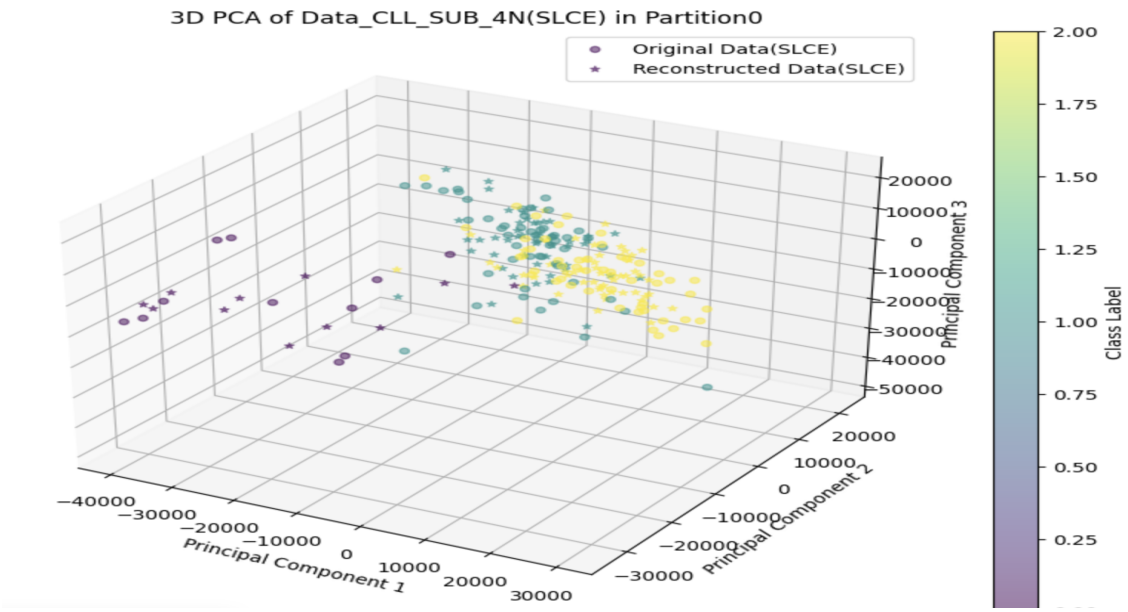

Figure 16: 3D CLL_SUB SLCE COMBINED 4N SAMPLES(2N ORIGINAL + 2N AUGMENTED)

Table 3: Mean Accuracy of RF and KNN Models Across Different Augmentation Datasets of CLL_SUB

Dataset	Mean RF Accuracy (%)	Mean KNN Accuracy (%)
Combined 2n (n original + n synthetic)	64.91	54.09
Combined 4n (2n original + 2n synthetic)	65.79	55.36
Original Data (n) (n original)	67.45	53.31
Augmented 2n (2n synthetic)	59.75	56.14
Augmented n (n synthetic)	61.70	53.22

Table 4: Mean Accuracy of RF and KNN Models Across Different Augmentation Datasets of CLL_SUB_SLCE

Dataset	Mean RF Accuracy (%)	Mean KNN Accuracy (%)
Combined 2n SLCE (n original + n synthetic)	75.44	72.12
Combined 4n SLCE (2n original + 2n synthetic)	74.37	72.61
Original Data (n) (n original)	76.41	71.54
Augmented 2n SLCE (2n synthetic)	73.59	70.18
Augmented n SLCE (n synthetic)	70.76	69.30

1.2.3 Effects of VAE Augmentation on CLL_SUB and CLL_SUB_SLCE Dataset

From the examination of the Principal Component Analysis (PCA) visualizations and the corresponding accuracy data for both the Random Forest (RF) and K-Nearest Neighbors (KNN) models, several conclusions can be drawn. The PCA plots for the CLL_SUB and CLL_SUB_SLCE datasets clearly differentiate between the original and reconstructed data, indicating that the reconstructed data retains a distinct pattern, which is crucial for modeling purposes. The inclusion of a third dimension in the PCA visualizations enhances the visibility of the data's structure, suggesting that higher-dimensional analysis may be beneficial for capturing complex relationships within the data.

When analyzing model performances, it appears that the RF model consistently outperforms the KNN model, with the combination of original and synthetic data generally resulting in better accuracy than using either data type alone. Particularly for the CLL_SUB_SLCE dataset, combining original and synthetic data in equal proportions (Combined 2n) yields the most favorable RF accuracy. This highlights the importance of incorporating original data into the training set to maximize model performance. Conversely, datasets augmented exclusively with synthetic data

tend to have lower accuracy rates, emphasizing the irreplaceable value of original data. These insights affirm the advantageous role of synthetic data augmentation in enhancing machine learning models, especially when such models are used in complex domains where data may be scarce or challenging to collect.

1.2.4 Summary

- PCA plots for CLL_SUB and CLL_SUB_SLCE datasets reveal that original and reconstructed datasets can be differentiated, suggesting a retained uniqueness in the reconstructed sets.
- The third dimension in the 3D PCA plots provides further insights, highlighting complex structures and groupings within the data.
- Comparisons of model accuracies indicate that the Random Forest (RF) model outperforms the K-Nearest Neighbors (KNN) model in the majority of cases.
- Combined datasets, containing both original and synthetic data, yield better performance in models compared to datasets containing solely original or synthetic data.
- Specifically, for CLL_SUB_SLCE datasets, the combined 2n dataset (with equal parts original and synthetic data) achieves higher RF accuracy compared to other data compositions.
- Augmented datasets, which consist solely of synthetic data, show lower accuracies in comparison, which underscores the value of having original data in training models.
- The findings support the utility of synthetic data in improving machine learning model performance, particularly when it is used to augment rather than replace original datasets.

1.3 SMK_CAN & SMK_CAN SLCE

1.3.1 SMK_CAN

Code Link

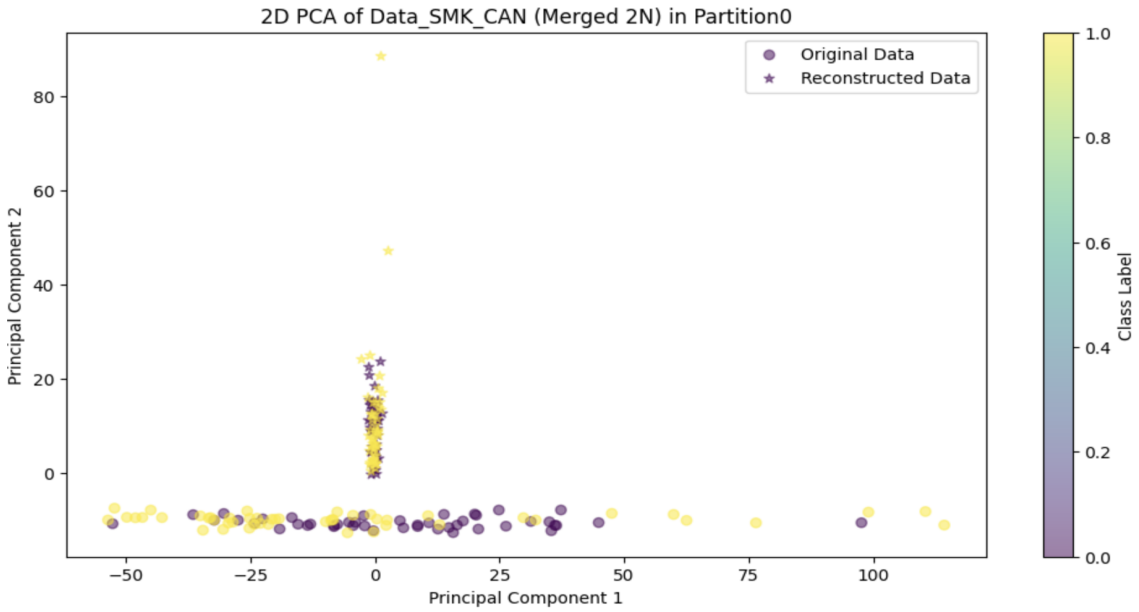

Figure 17: 2D SMK_CAN COMBINED 2N(N ORIGINAL + N AUGMENTED)

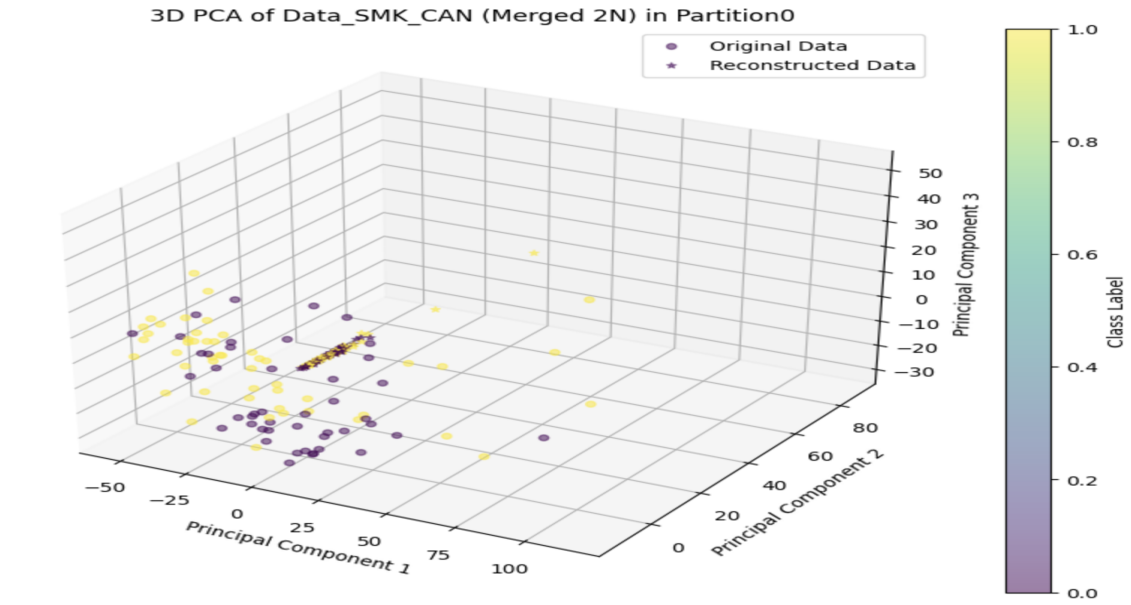

Figure 18: 3D SMK_CAN COMBINED 2N (N ORIGINAL + N AUGMENTED)

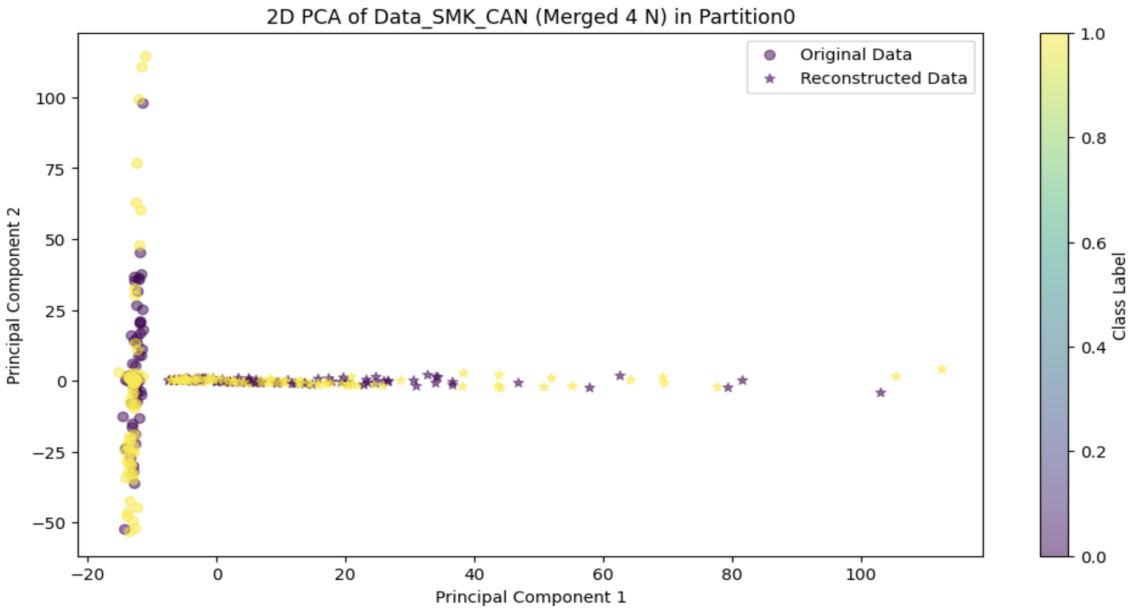

Figure 19: 2D SMK_CAN COMBINED 4N SAMPLES(2N ORIGINAL + 2N AUGMENTED)

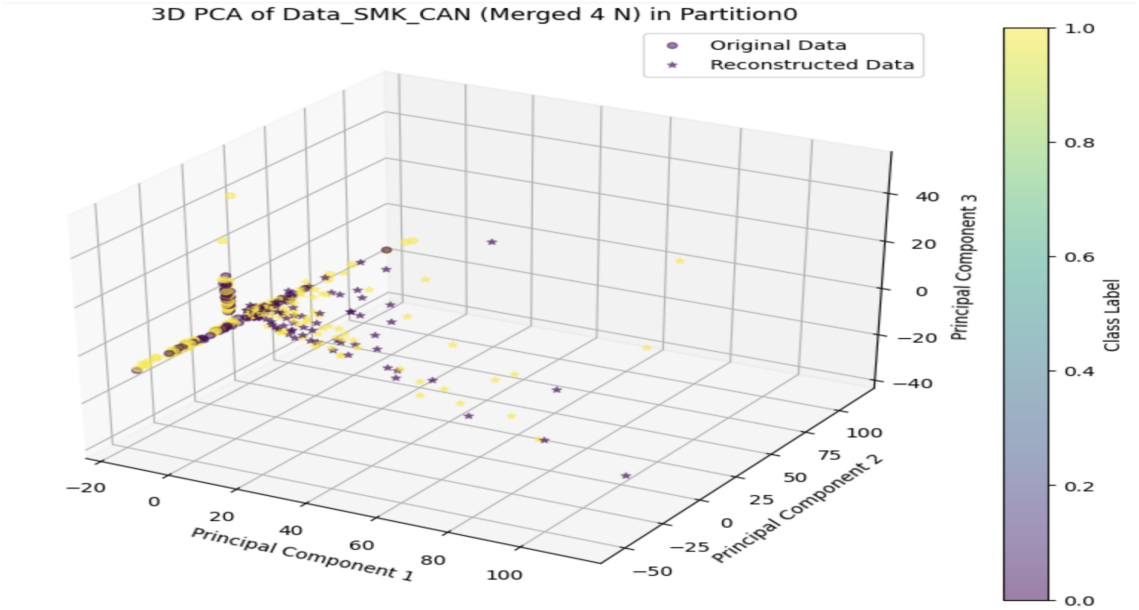

Figure 20: 3D SMK_CAN COMBINED 4N SAMPLES(2N ORIGINAL + 2N AUGMENTED)

1.3.2 SMK_CAN SLCE

Code Link

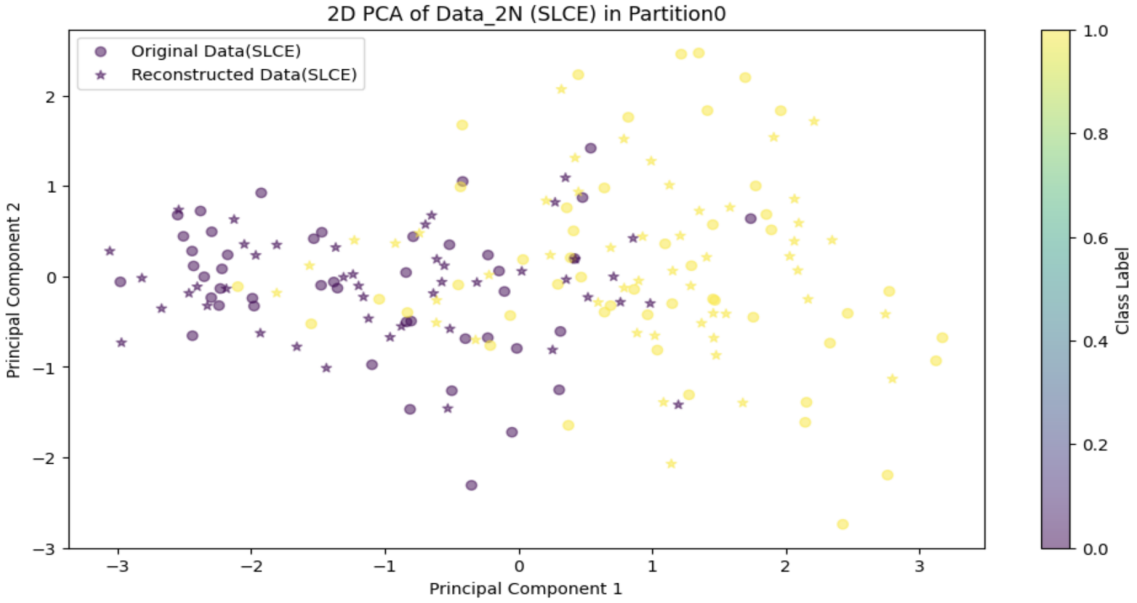

Figure 21: 2D SMK_CAN SLCE MERGED 2N(N ORIGINAL + N AUGMENTED)

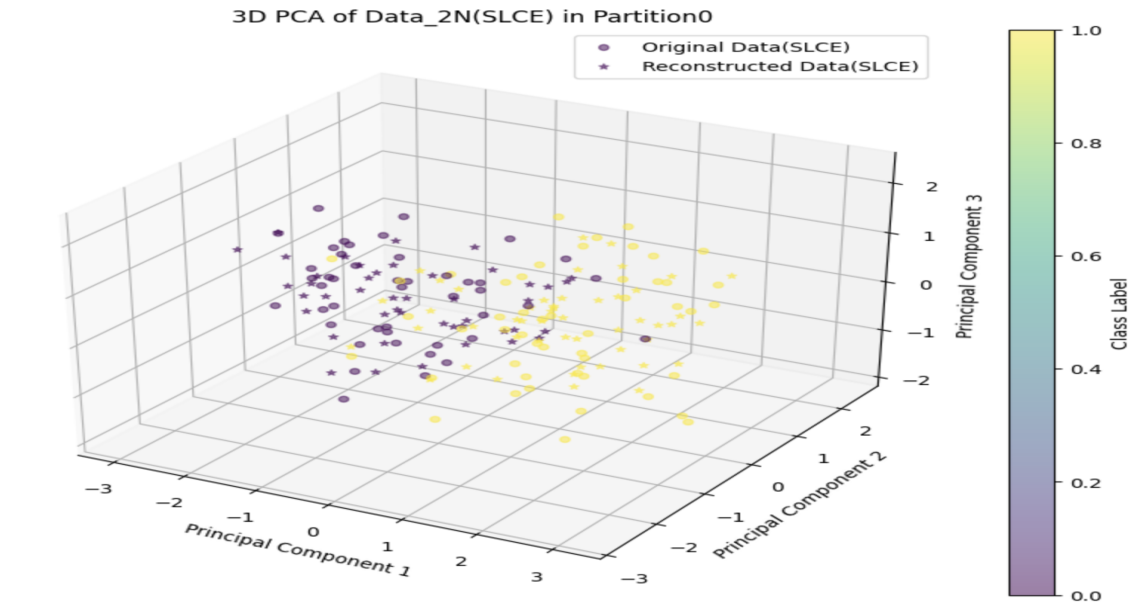

Figure 22: 3D SMK_CAN SLCE MERGED 2N (N ORIGINAL + N AUGMENTED)

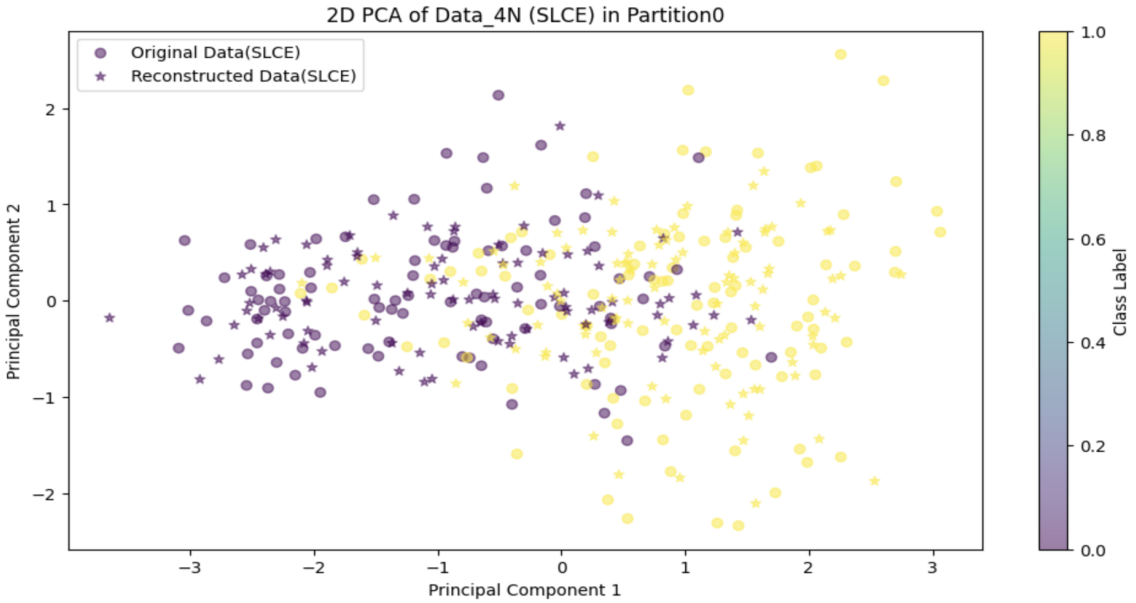

Figure 23: 2D SMK_CAN SLCE MERGED 4N SAMPLES(2N ORIGINAL + 2N AUGMENTED)

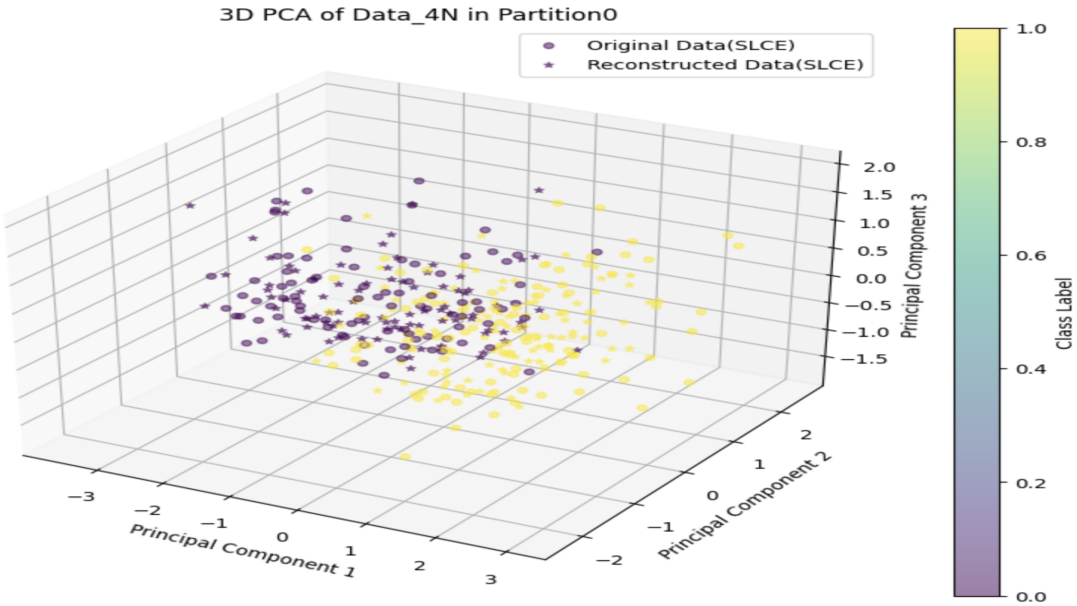

Figure 24: 3D SMK_CAN SLCE 4N SAMPLES(2N ORIGINAL + 2N AUGMENTED)

Table 5: Mean Accuracy of RF and KNN Models Across Different Datasets in SMK_CAN

Dataset Name	Mean RF Accuracy (%)	Mean KNN Accuracy (%)
Combined 2n (n original + n synthetic)	67.69	59.41
Combined 4n (2n original + 2n synthetic)	68.10	62.72
Original Train Data (n original)	67.42	65.86
Augmented n (n synthetic)	47.58	51.40
Augmented 2n (2n synthetic)	51.253	50.53

Table 6: Mean Accuracy of RF and KNN Models Across Different Configurations and Partitions in SMK_CAN SLCE

Dataset Name	Mean RF Accuracy (%)	Mean KNN Accuracy (%)
Combined 2n SLCE (n original + n synthetic)	69.30	69.95
Combined 4n SLCE (2n original + 2n synthetic)	69.03	67.80
Original Train Data (n original)	68.76	67.31
Augmented 2n SLCE (2n synthetic)	67.85	68.82
Augmented n SLCE (n synthetic)	67.96	69.73

1.3.3 Effects of VAE Augmentation on SMK_CAN and SMK_CAN_SLCE Dataset

The 2D and 3D PCA plots for the datasets—especially the ones processed with SLCE—demonstrate a decent correlation between the original and reconstructed data, although some variations are noticeable. This variation could be indicative of the noise or the reduction in dimensionality affecting the representation of the data points post-augmentation or reconstruction.

From the accuracy tables provided, it's clear that feature selection through SLCE has generally led to an improvement in model accuracy across different configurations for the SMK_CAN datasets. This suggests that the top 50 features selected by SLCE hold significant predictive power and are robust across augmented data as well.

Looking at the tables for CLL_SUB datasets, a similar trend can be seen. With SLCE, there's a notable increase in accuracy for both RF and KNN models. This could imply that despite the augmentation, which typically aims to expand the training data and improve model generalization, SLCE's concise and relevant feature set can result in better performance.

2 Overall conclusion of VAE approaches

For Glioma Dataset, combining original data and synthetic data really improved the random forest and knn model's performance.

For both SMK_CAN and CLL_SUB datasets, the use of SLCE for feature selection has, in several instances, resulted in higher model accuracies than using augmented data alone. This finding stresses the importance of feature selection in enhancing model prediction and could be vital for future modeling endeavors.

It's also remarkable that in some cases, the models trained on combined datasets do not outperform those trained on the original datasets, which could suggest that the augmentation process may not always align well with the underlying distribution of the data or that it introduces redundancy or irrelevant variations.

In summary, these findings highlight the significance of selecting the right features in data modeling. While augmentation is a common strategy to address limited data, feature selection can be equally—if not more—impactful. This points to a strategic approach of combining SLCE feature selection with careful data augmentation to optimize model performance.