2023 26th International Conference on Computer and Information Technology (ICCIT)
13-15 December 2023, Cox's Bazar, Bangladesh

An improved Spectral Analysis based Subspace Detection for Hyperspectral Image Classification

Kazi Mehrab Rashid[1], Md. Ali Hossain[2], Tasmia Jannat[3] and Md. Anisur Rahman[4]
[1,2,3]*Deptartment of Computer Science & Engineering, Rajshahi University of Engineering & Technology, Bangladesh*
[4]*Department of Business Analytics, La Trobe University, Australia*
Email- mahimchy2012@gmail.com, ali.hossain@cse.ruet.ac.bd, jannat22tasmia@gmail.com, anisur.rahman@latrobe.edu.au

Abstract— **Hyperspectral imagery, with its rich spectral information, is indispensable in various remote sensing applications. However, working with hyperspectral data introduces two formidable challenges: the curse of dimensionality and the predicament of class imbalance. The inherent high dimensionality necessitates substantial computational resources, while class imbalance undermines overall classification accuracy due to the scarcity of samples in training for certain classes. To mitigate the curse of dimensionality, this research analyses a few diverse feature extraction techniques, including Principal Component Analysis (PCA), Sparse PCA, segmented PCA, and Linear Discriminant Analysis (LDA). Although these methods can reduce the features from the original dataset, however sometimes depend on certain global statistics such as variance. Also, these methods could not handle the small classes as they could not represent the overall characteristics of the samples. Therefore, the proposed method incorporates the solution to the class imbalance problem as well as reduces the features based on the given classes. The ensuing classification phase employs a kernel Support Vector Machine (KSVM), achieving an accuracy of 88.81% only when the class imbalance exists. Later the class imbalances are resolved through the incorporation of Synthetic Minority Over-sampling Technique (SMOTE). The proposed approach reduces the input features and generates synthetic data for each class based on the proportions original simultaneously rectifying class imbalance while preserving dataset characteristics. Remarkably, this strategy results in an improvement of accuracy to 98.69%. This study underscores the significance of hyperspectral imagery, delineates its inherent challenges, and presents novel solutions that substantially enhance classification performance, making hyperspectral data more accessible and valuable in remote sensing endeavors.**

Keywords— Remote Sensing, Hyperspectral Image, Principal Component Analysis (PCA), SPCA, Linear Discriminant Analysis (LDA), Support Vector Machine (SVM), SMOTE.

I. Introduction

Imaging spectroscopy, also known as hyperspectral imaging, is concerned with the measurement, analysis, and interpretation of spectra acquired from a given scene (or specific object) [1]. Hyperspectral imaging is a specialized and powerful technique employed in remote sensing, earth observation, and numerous scientific disciplines. With the advent of hyperspectral remote sensors, hundreds of narrow contiguous spectral bands can now be captured to provide greater details on the spectral variation of targets than with conventional multispectral systems [2]. This fine spectral resolution allows for the precise identification of materials based on their unique spectral signatures. As a result, hyperspectral imaging has become indispensable in fields such as remote sensing, geology, agriculture, forestry, and environmental science, where it aids in the identification and classification of materials, including soils, vegetation types, and minerals.

However, harnessing the potential of hyperspectral data comes with inherent challenges. One of the significant challenges that come with hyperspectral image is curse of dimensionality. In hyperspectral data, each spectral band represents a dimension, and the data points reside in this high-dimensional space. As the number of spectral bands increases, so does the dimensionality of the data. While this provides an opportunity to discern subtle differences in spectral signatures and identify materials with precision, it introduces several formidable challenges. The curse of dimensionality leads to an exponential growth in the volume of data. With a high number of dimensions, the available data points become sparse, making it challenging to gather enough samples to adequately represent the entire space. In a nutshell, high dimensionality leads to the curse of the dimensionality problem, which causes classification performance to deteriorate for testing data, especially when the number of available labeled training samples is limited [3].

Furthermore, in hyperspectral image analysis, a prevalent and often daunting challenge is the class imbalance issue. This phenomenon arises from the uneven distribution of objects or materials across different categories within hyperspectral data. It means that certain classes possess a significantly smaller number of instances or pixels compared to others. The class imbalance issue introduces several complications in the analysis and classification of hyperspectral data, as it can lead to biased models that favor the majority class and reduced sensitivity to minority classes, which may represent crucial or rare phenomena. Moreover, traditional accuracy metrics can be misleading when dealing with imbalanced data, making it imperative to employ specialized techniques such as oversampling, under-sampling, ensemble methods, and cost-sensitive learning to address this challenge effectively. Indeed, this problem is relevant in classification tasks, as sample labeling is quite difficult and expensive in practice, while it is absolutely necessary to model the minority classes in order to conduct a complete and accurate land cover analysis [4].

In this paper, the crucial challenges posed by high dimensionality and the imbalance of class samples during hyperspectral image classification are addressed. Hence, an integrated approach that combines both the dimensionality reduction approaches with a novel strategy for addressing class imbalance problem is proposed. By synergizing these

979-8-3503-5901-5/23/$31.00 ©2023 IEEE

techniques, we aim to unlock the latent information within hyperspectral data, elevating the accuracy and reliability of classification outcomes.

II. Literature review

In the current literature, few approaches utilize the dimensionality reduction approaches as primary tasks to improve the classification problem. Among these, few of them considers only the subset of classes (i.e., classes with high samples) to avoid the problems of hyperspectral image classification. Although in [5], the class imbalance problem was discussed by combining SMOTE with centroid-based clustering and its consequences are presented however it extensively does not review related works. This paper utilizes the potential of the above idea for advancing hyperspectral image classification. The study's significance lies in its effort to improve the reliability of remote sensing data interpretation in the context of climate change and informed decision-making.

But the main problem lies with [5] is they used Synthetic Minority Oversampling Technique (SMOTE) for the purpose of oversampling only the minority classes. What happens in this process is that the ratio of each class presented in the main dataset is interrupted. Because here they were focusing on increasing the number of samples for only minority class ignoring the rest of the classes. Hence, the minority class in the main dataset will be the dominant class in the updated dataset that doesn't preserve the properties of the main one. In [6] semi supervised learning is presented to address the challenge of imbalanced datasets in hyperspectral remote sensing image classification. The issue of varying training sample numbers across classes is a concern for applications in environmental monitoring, agriculture, and mineralogy. While existing literature highlights the importance of tackling imbalanced datasets, this paper innovatively applies semi-supervised learning to hyperspectral image classification. The proposed "NearPseudo" technique creates pseudo-labels from an initialization classifier, augmenting minority class samples. A feedback mechanism enhances pseudo-label reliability through consistency checks.

While this study has demonstrated the strong performance of NearPseudo, it is essential to acknowledge its contextual dependency. The effectiveness of NearPseudo relies heavily on specific conditions. Optimal results are achieved when a substantial volume of unlabeled data is available for efficient processing with NearPseudo. A significant pool of unlabeled samples is required to facilitate the selection of suitable samples for augmenting imbalanced datasets effectively. And also, this has used oversampling strategy which again made the minority class as dominant class. But the problem happens here is that the properties of the dataset is hampered. Like the previous ,paper other classes were totally ignored and only the minority class has been given the attention.

III. Methodology

The studied and proposed methodologies center on tackling the challenges of hyperspectral image analysis through a synergistic integration of techniques. The experiment commences by applying dimensionality reduction methods, specifically Principal Component Analysis (PCA), Sparse and Segmented PCA, and Linear Discriminant Analysis (LDA), to effectively extract pertinent features while mitigating the complexities of high dimensionality. To address class imbalance, we introduce a novel approach to the Synthetic Minority Over-sampling Technique (SMOTE) tailored to maintain the inherent distribution of the original dataset. This balanced dataset is then fed into kernel Support Vector Machine (KSVM) classifiers to enhance classification accuracy. By fusing dimensionality reduction, modified SMOTE, and SVM classification, the methodology provides a comprehensive framework to unlock the latent insights within hyperspectral data while overcoming challenges posed by dimensionality and class imbalance. Here Figure 2 is our proposed methodology.

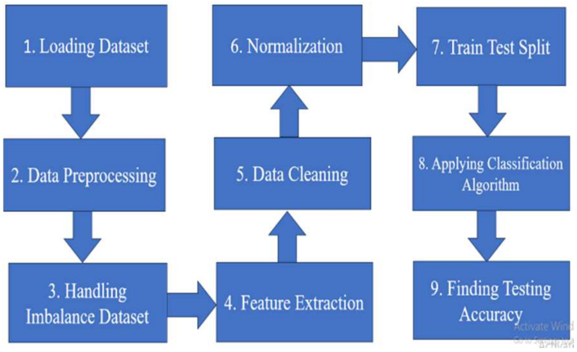

Figure 1: Proposed Methodology

A. Dataset Description

Before diving deep into all the steps of methodology, let's first discuss the dataset and this is Indian Pines hyperspectral dataset which is shown in Figure 1. The Indian Pines dataset is collected by Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) sensor over the agricultural area of Indian Pine test site, which is in North-western Indiana [7]. It comprises a grid of dimensions 145×145 pixels and encompasses 220 spectral bands, each capturing specific wavelength information across the electromagnetic spectrum. So the dataset we are working with is of size (145×145×220). The scene was collected over a mixed agricultural area with a spatial resolution of 20 m and a spectral resolution of 10 nm in the wavelength range of 0.4–2.4 µm [8]. Focusing on agricultural elements, the dataset features 16 distinct land cover classes, including Alfalfa, Corn-notill, Corn-mintill, Corn, Grass-pasture, Grass-trees, Grass-pasture-mowed, Hay-windrowed, Oats, Soybean-notill, Soybean-mintill, Soybean-clean, Wheat, Woods, Buildings-Grass-Trees-Drives, and Stone-Steel-Towers. .

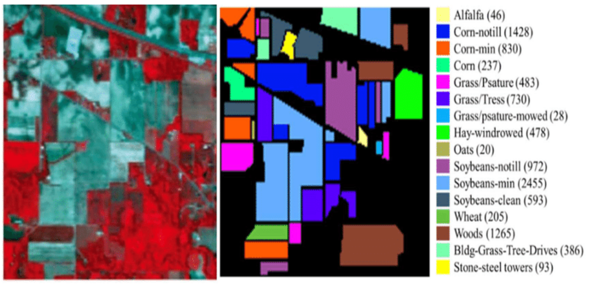

Figure 2: Indian Pine Dataset caputured by an AVIRIS sensor [9]

B. Loading Dataset

Now, let's dive into all the steps of methodology one by one. This is the first step that is loading the dataset. The AVIRIS Hyperspectral Indian Pine Dataset which has 220 image bands and each band contain 145 × 145 pixels. So the dimension of this dataset is (145 × 145 × 220). The Indian Pine dataset and its ground truth image are presented in Figure 2. This dataset contains mixed agriculture area having 16 different land cover classes which are listed in Figure 2.

C. Data Preprocessing

Since dealing with 3D or more dimension data will be intimidating as the higher the number of dimensions is the higher the complexity. So instead of working with 3D data, it is converted 2D data for the processing. Since the data has 220 number of bands captured on different wavelength, each image band termed as input feature. Therefore, the number of features is 220. So, the shape of updated dataset will be 2 dimensional (21025 × 220). Likewise, the ground truth data of 2D shape (145 × 145) will also be made 1D by multiplying 145 × 145 = 21025.

D. Handling Imbalance Dataset

Since it has already been told before that Indian pine dataset has some classes that have insufficient number of samples which leads to poor accuracy. So, the way the class imbalance issue has been solved is by introducing a modified approach to the Synthetic Minority Over-sampling Technique (SMOTE). This proposed approach involves calculating required sample ratios for each class before generating synthetic data. The methodology was applied across all 16 classes of the Indian Pines dataset. So, what we have performed here is that, first we calculate the total number of number of samples for class no 1 to class no 16. After that we measure the ratio of each class in the original dataset. Later, we pick a number of additional data that will be added on the original dataset. After that, the number of synthetic data generation for each class will be decided according to the ratio of each class which was measured before. For example, we have two classes A and B and their sample numbers are 6 and 4 respectively. So, the total number of samples here is 10 and class A has 60% of total samples and class B has 40% of total samples. Now if we decide to add 100 more data, then 60% of the additional data will be added for class A and 40% of the additional data will be added for class B. The reason we have approached this method is that even if we add any amount of synthetic data, the properties and the ratio of each class of the dataset will be maintained and kept intact properly. Overall, this ensures that the synthesized data retained the essential properties of the original dataset.

E. Feature Extraction

Feature extraction is a fundamental process in the domain of machine learning and pattern recognition, where intricate and often high-dimensional data are transformed into a more compact and informative representation, known as features. Feature extraction can be achieved by mapping the input images into a new feature space where the first few components carry most of the information for classification tasks [10]. The primary objective of feature extraction is to distill the relevant and discriminative aspects of the original data, enabling enhanced model performance, reduced computational complexity, and improved generalization capabilities. Feature extraction techniques that we have applied in this research paper are:

- Principal Component Analysis (PCA)
- Segmented PCA
- Sparse PCA
- Linear Discriminant Analysis

Principal component analysis (PCA) is a mathematical algorithm that reduces the dimensionality of the data while retaining most of the variation in the data set [11]. Sparse PCA, a variation of traditional Principal Component Analysis (PCA), promotes sparsity in the principal components by favoring mostly zero values. Segmented PCA is a modified version of Principal Component Analysis (PCA) that applies PCA separately to different segments or subsets of data. Segmented PCA can extract the overall characteristics of each segmented dataset [12]. Linear discriminant analysis (LDA) is one of the dimensionality reduction methods that maximizes the ratio of between-class variance to the within-class variance in any particular data set thereby guaranteeing maximal separability [13].

F. Data Cleaning

If we observe the ground truth of Indian pine dataset closely , we will notice that the actual number of classes in the .GIS file is 17 where it should have been 16. But the reason of extra one class being present in the dataset is the white background. And even the number of this unwanted or noisy class is 10659 which is the highest among all classes. Since this class is not what we want , so this class will be deleted from the dataset.

G. Normalization

Normalization is a data preprocessing technique used to standardize the range and distribution of features in a dataset. It involves adjusting the values of variables to bring them within a common scale, typically between 0 and 1 or with a mean of 0 and a standard deviation of 1. Normalization is employed to ensure that all features contribute equally to the analysis and to prevent certain features from dominating others due to their larger magnitude. There are many normalization techniques but the one we have used for this research is Z-score normalization.

H. Train Test Split

For the purpose of this research, 70% of the data has been trained and the rest 30% of the data will be used for testing. The value for random state has been considered to be 42.

I. Applying Classification Algorithm

There are many classification algorithms, for example – logistic regression, k nearest neighbor, support vector machine, decision tree and random forest. But among all these, we have chosen support vector machine (SVM). The reason we chose SVM are:

- It is high effective in high dimensional spaces.
- It has versatile kernel functions.
- It has the ability to handle nonlinear data.
- It is robust to outliers.
- It is originally designed for binary classification but can be extended to handle multiclass problem.

- It finds the optimal hyperplane by solving a convex optimization problem.

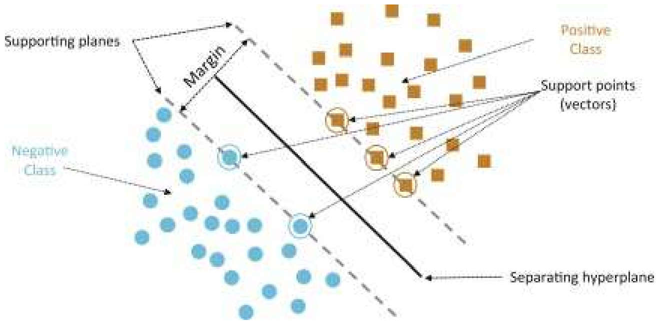

Figure 3: Support Vector Machine working Principle [14]

This is how Support Vector Machine works:

- Firstly, it selects a random hyperplane.
- Then two planes are drawn which are parallel to the hyperplane and go through support vectors. Here, support vectors are those data points that lie nearest to the hyperplane.
- Then the distance between two parallel planes is calculated. This distance is called margin.
- The above steps continue until best hyperplane is found out. Here, the best hyperplane refers to that decision boundary for which the distance between the closest samples to boundary is maximum.

SVMs are originally designed for linearly separable data. However, many real-world datasets are not linearly separable, and this is where the kernel trick comes in. In essence, the kernel trick is a mathematical method that enables SVMs to handle non-linear data by projecting it into a higher-dimensional space, where a linear decision boundary can be found. It's a key feature that makes SVMs versatile for various machine learning applications.

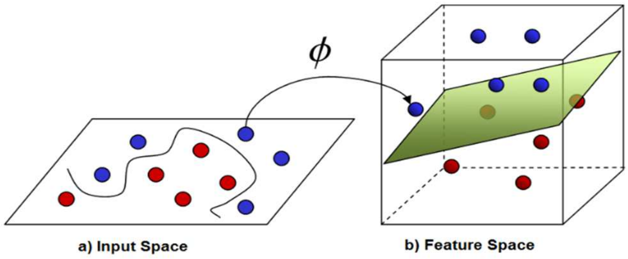

Figure 4: Kernel Trick in SVM [15]

Common kernel functions include the linear kernel, polynomial kernel, radial basis function (RBF) kernel and sigmoid kernel. For this research we have used RBF kernel. Because the RBF kernel excels in capturing non-linear patterns in data, making it versatile and it doesn't require explicit high-dimensional mapping.

J. *Finding Testing Accuracy*

After training the model with cross validation and grid search, the best hyperparameter value is obtained to finally test the model.

IV. Results

In the experimentation, using the Indian Pines hyperspectral dataset, SVM classification without feature extraction yielded an accuracy of 88.55%. Incorporating dimensionality reduction, PCA + SVM achieved 88.68%, SPCA + SVM achieved 88.81%, LDA + SVM achieved 86.94%, and Segmented PCA + SVM achieved 86.97%. This has been shown in Figure 5.

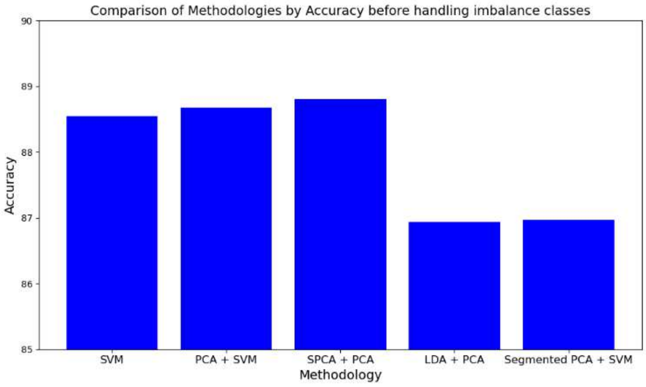

Figure 5: Accuracy before SMOTE

After performing rigorous analysis, this average level accuracy creates the scope of further processing the data for an improved classification. Hence the class imbalance problem appears as a vital concern and we plan to address it. Addressing class imbalance with the proposed Modified SMOTE technique significantly boosted SVM accuracy to 98.3%. Further, integrating Modified SMOTE with PCA + SVM achieved 98.53%, Modified SMOTE + SPCA + SVM achieved 98.69%, Modified SMOTE + LDA + SVM achieved 96.73%, and Modified SMOTE + Segmented PCA + SVM achieved 98.43%. This has been visualized in Figure 6.

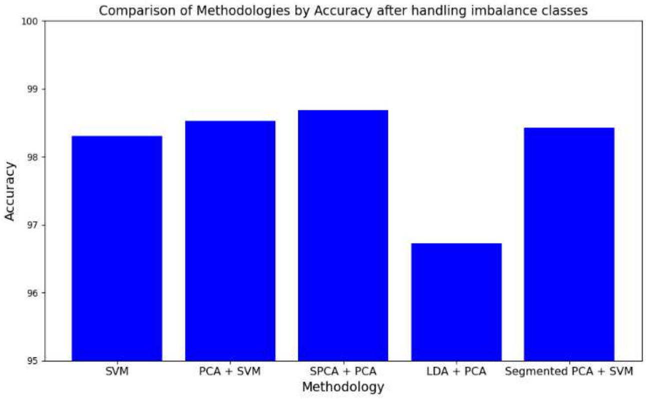

Figure 6: Accuracy after SMOTE

These outcomes emphasize the importance of tackling dimensionality and class imbalance challenges in hyperspectral image analysis, showcasing improved accuracy and enhancing remote sensing applications. Overall accuracy of before and after solving class imbalance issue have been shown in Table 1.

TABLE 1. ACCURACY OF DIFFERENT METHODS BEFORE AND AFTER SOLVING CLASS IMBALANCE ISSUE

Methods	Accuracy (Before solving class imbalance issue)	Accuracy (After solving class imbalance issue)
SVM	88.55%	98.3%
PCA + SVM	88.68%	98.53%
SPCA + SVM	88.81%	98.69%
LDA + SVM	86.94%	96.73%
Segmented PCA + SVM	86.97%	98.43%

Table 1 depicts the overall accuracy of before and after solving class imbalance issue. Now let’s see what the accuracy of each class was before and after applying modified SMOTE. First of all, let’s look at Table 2 that shows the class names, the accuracy of each class using SVM before handling imbalance class and original number of test samples for each class.

TABLE 2. CLASS-WISE ACCURACY OF SVM BEFORE APPLYING MODIFIED SMOTE AND NUMBER OF TEST SAMPLES

Class Names	Accuracy (Before solving class imbalance issue)	Original Number of Test Samples
Alfalfa	69%	16
Corn-notill	89%	449
Corn-mintill	82%	267
Corn	70%	69
Grass-pasture	95%	149
Grass-trees	97%	244
Grass-pasture-mowed	80%	5
Hay-windrowed	99%	144
Oats	80%	5
Soybean-notill	76%	287
Soybean-mintill	90%	725
Soybean-clean	88%	187
Wheat	100%	54
Woods	98%	354
Building-Grass-Trees-Drives	61%	124
Stone-Steel-Towers	100%	31

Now let’s look at Table 3. Here, the number of samples has been increased through data augmentation to solve class imbalance issue because few classes for instances - alfalfa, corn, oats etc. have only a few samples that may not describe the overall statistics of the classes.

TABLE 3: CLASS-WISE ACCURACY OF SVM AFTER APPLYING MODIFIED SMOTE AND NUMBER OF TEST SAMPLES

Class Names	Accuracy (After solving class imbalance issue)	Increased Number of Test Samples
Alfalfa	100%	44
Corn-notill	96%	1305
Corn-mintill	93%	734
Corn	97%	211
Grass-pasture	99%	444
Grass-trees	100%	637
Grass-pasture-mowed	100%	19
Hay-windrowed	100%	432
Oats	100%	21
Soybean-notill	93%	836
Soybean-mintill	96%	2149
Soybean-clean	97%	547
Wheat	100%	195
Woods	100%	1109
Building-Grass-Trees-Drives	95%	345
Stone-Steel-Towers	100%	80

Now let’s see the accuracy comparison of each class using PCA + SVM and Modified SMOTE + PCA + SVM through line graph has been shown in Figure 7.

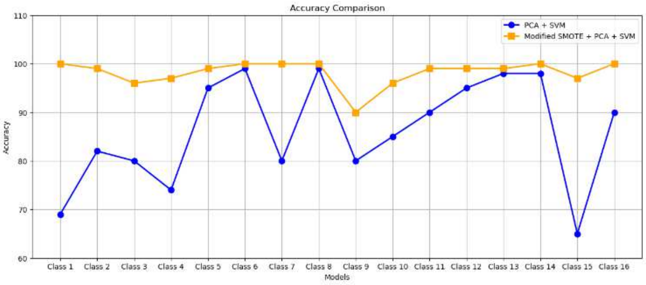

Figure 7: Accuracy comparison of each class using PCA + SVM and Modified SMOTE + PCA + SVM

The accuracy comparison of each class using SPCA + SVM and Modified SMOTE + SPCA + SVM through line graph has been shown in Figure 8.

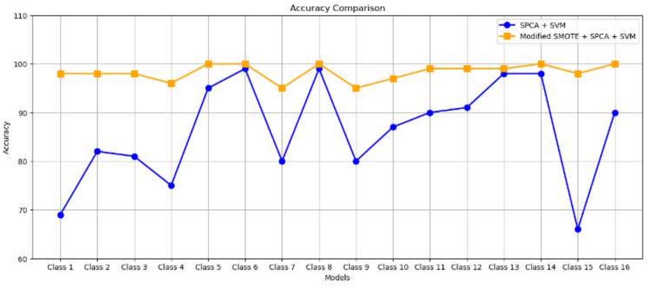

Figure 8: Accuracy comparison of each class using SPCA + SVM and Modified SMOTE + SPCA + SVM

The accuracy comparison of each class using LDA + SVM and Modified SMOTE + LDA + SVM through line graph has been shown in Figure 9.

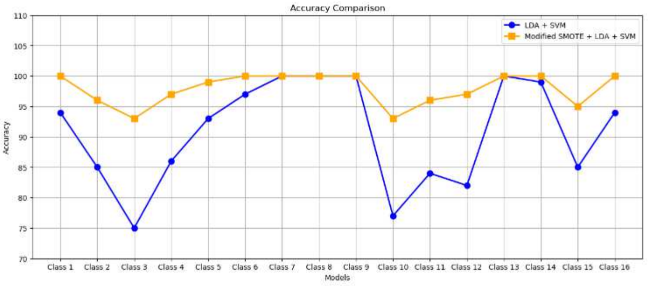

Figure 9: Accuracy comparison of each class using LDA + SVM and Modified SMOTE + LDA + SVM

The accuracy comparison of each class using Segmented PCA + SVM and Modified SMOTE + Segmented PCA + SVM through line graph has been shown in Figure 10.

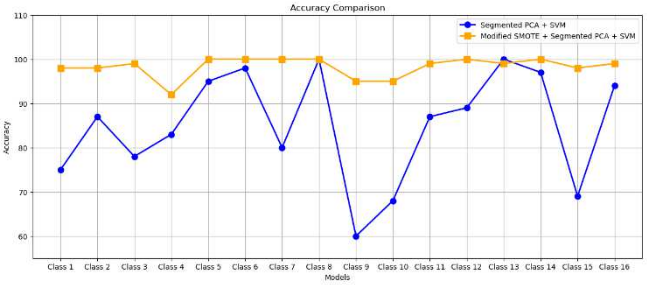

Figure 10: accuracy comparison of each class using Segmented PCA + SVM and Modified SMOTE + Segmented PCA + SVM

V. CONCLUSION

This study addresses the complexities of hyperspectral image analysis through dimensionality reduction and class imbalance handling. We demonstrate the significance of these techniques in enhancing classification accuracy and unlocking the potential of hyperspectral data. By employing dimensionality reduction methods, including PCA and Segmented PCA, along with LDA, we achieve substantial accuracy improvements over the baseline SVM classification. Notably, PCA + SVM reaches 88.68% and SPCA + SVM achieves 88.81%, highlighting their efficacy in capturing crucial features within high-dimensional data. While applying SVM on whole dataset, the accuracy was 88.55%. This means that feature extraction techniques have been able to reduce number of dimension while keeping the accuracy as high possible that helps us working with only few of bands rather working with all the bands.

Additionally, we tackle class imbalance, which significantly impacts outcomes. The baseline SVM accuracy is 88.55%. Our innovative Modified Synthetic Minority Over-sampling Technique (SMOTE) lifts accuracy to 98.3%, a significant enhancement. Furthermore, combining dimensionality reduction with Modified SMOTE yields even higher accuracy. Notably, Modified SMOTE + PCA + SVM achieves 98.53%, while Modified SMOTE + SPCA + SVM reaches 98.69%. These findings underscore the synergy between dimensionality reduction and class imbalance handling. The proposed work offers a holistic framework that aligns these critical aspects, paving the way for accurate classification in hyperspectral image analysis. This contribution advances the field and encourages further exploration for enhanced insights.

VI. REFERENCES

[1] Plaza, A., Benediktsson, J. A., Boardman, J. W., Brazile, J., Bruzzone, L., Camps-Valls, G., ... & Trianni, G. (2009). Recent advances in techniques for hyperspectral image processing. Remote sensing of environment, 113, S110-S122.

[2] Hossain, M. A., Jia, X., & Pickering, M. (2013). Subspace detection using a mutual information measure for hyperspectral image classification. IEEE Geoscience and Remote Sensing Letters, 11(2), 424-428.

[3] Feng, F., Li, W., Du, Q., & Zhang, B. (2017). Dimensionality reduction of hyperspectral image with graph-based discriminant analysis considering spectral similarity. Remote sensing, 9(4), 323.

[4] Roy, S. K., Haut, J. M., Paoletti, M. E., Dubey, S. R., & Plaza, A. (2021). Generative adversarial minority oversampling for spectral–spatial hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing, 60, 1-15.

[5] Singh, P. S., Singh, V. P., Pandey, M. K., & Karthikeyan, S. (2022). Enhanced classification of hyperspectral images using improvised oversampling and undersampling techniques. International Journal of Information Technology, 14(1), 389-396.

[6] Zheng, X., Jia, J., Chen, J., Guo, S., Sun, L., Zhou, C., & Wang, Y. (2022). Hyperspectral image classification with imbalanced data based on semi-supervised learning. Applied Sciences, 12(8), 3943.

[7] Mishu, S. Z., Ahmed, B., Hossain, M. A., & Uddin, M. P. (2020). Effective subspace detection based on the measurement of both the spectral and spatial information for hyperspectral image classification. International Journal of Remote Sensing, 41(19), 7541-7564.

[8] Hossain, M. A., Jia, X., & Benediktsson, J. A. (2016). One-class oriented feature selection and classification of heterogeneous remote sensing images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 9(4), 1606-1612.

[9] Liu, J., Zhang, L., Huang, H., Zheng, C., & Yang, S. (2020). Multiple Characteristics Similarity Metric Method for Hyperspectral Image Classification. IEEE Access, 8, 9501-9512.

[10] Islam, R., Ahmed, B., & Hossain, A. (2022). Feature reduction of hyperspectral image for classification. Journal of Spatial Science, 67(2), 331-351.

[11] Jolliffe, I. T. (2002). Principal component analysis for special types of data (pp. 338-372). Springer New York.

[12] Uddin, M. P., Mamun, M. A., & Hossain, M. A. (2021). PCA-based feature reduction for hyperspectral remote sensing image classification. IETE Technical Review, 38(4), 377-396.

[13] Balakrishnama, S., & Ganapathiraju, A. (1998). Linear discriminant analysis-a brief tutorial. Institute for Signal and information Processing, 18(1998), 1-8.

[14] Gholami, R., & Fakhari, N. (2017). Support vector machine: principles, parameters, and applications. In Handbook of neural computation (pp. 515-535). Academic Press.

[15] Vocaturo, E., Perna, D., & Zumpano, E. (2019, November). Machine learning techniques for automated melanoma detection. In 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (pp. 2310-2317). IEEE.