Group

- Dubai CW PG Thursday Group 6

Group Members

- Mohamed Aman | Pratibha Yadubanshi | Faizan Watare | Fardeen Khan

F21DL Coursework Part 5 – Research Question

Question

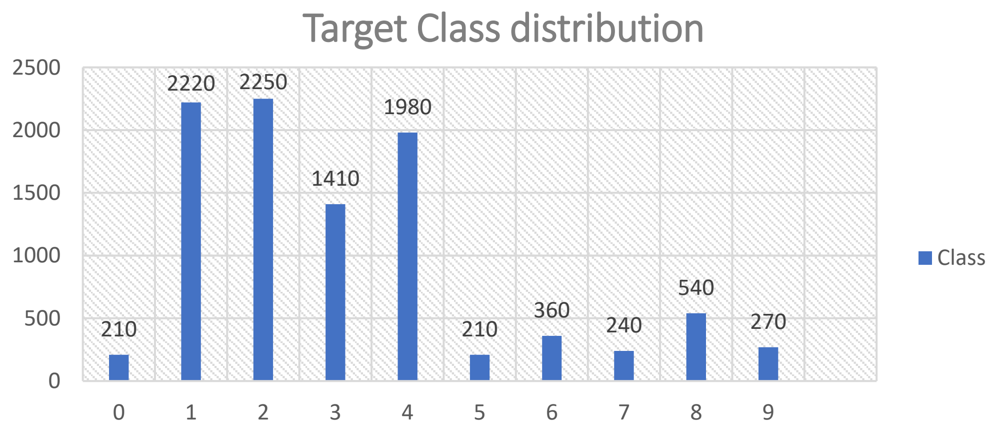
How can the performance of a machine learning model get influenced by data balancing through augmentation and synthetic data? Additionally, what consequences does this balancing approach have on the overall accuracy and reliability of an image dataset?

Approach:

- **Exploratory Data Analysis:** Analyse and identify the specific classes that are underrepresented in the dataset.
- **Balancing Techniques**: Experiment with various data balancing techniques, such as oversampling, under sampling, or using synthetic and augmentation data generation methods.
- **Model Training**: Train machine learning models, such as decision trees or neural networks, on both the imbalanced and balanced datasets.
- **Performance Evaluation**: Evaluate the impact of balancing techniques compare the accuracy, precision, recall, and F1-score of both models.
- **Generalization**: Assess how well the models generalize to unseen data after applying the balancing strategies.

Exploratory Data Analysis (EDA) and Data Visualization

During the data analysis and visualization phase, we found that the dataset consists of 9690 images, each represented by 2304 features (48x48 pixels). The dataset consists of 10 distinct classes. To standardize the data, we normalized it by dividing each value by 255, resulting in scaled values between 0 and 1. Then we visualized the dataset and found that it is indeed an imbalanced dataset. We identified the underrepresented classes from this table.



Initial baseline scores

To establish a baseline score, we implemented various classifiers on the imbalanced dataset and documented the performance metrics. These scores provide an initial assessment of classifier performance on the imbalanced dataset, forming the basis for future development strategies.

	Naïve Bayes	Decision Tree	Random Forest	Logistic Regression
Accuracy	17.2%	68.5%	77.5%	87.9%
Precision (weighted Average)	36.7%	68.6%	78.1%	88%
Recall (weighted Average)	17.2%	68.5%	77.5%	87.9%
F1-Score (weighted Average)	22.5%	68.4%	76.7%	87.4%

Balancing Techniques

We explored various data balancing techniques, including the utilization of augmented data and synthetic data. The following table outlines the performance metrics of different models across various balancing approaches.

- Random over sampling
- Random under sampling
- SMOTE
- Mix of Over sampling and under sampling
- Augmentation using Keras ImageDataGenerator

The following table presents the model's performance metrics across these different approaches:

	Performance	Naïve Bayes	SVM	Decision Tree	Random Forest	Logistic Regression
Random Over Sampling	**Accuracy**	17.2%	71.6%	67.1%	77%	88.1%
	Precision	36.8%	72.7%	67%	76.9%	88.3%
	Recall	17.2%	71.6%	67.1%	77%	88.1%
	F1-Score	22.5%	71.40%	67%	76.1%	87.9%
	Accuracy	15.7%	44.1%	53.5%	65.5%	82%

Random Under Sampling	**Precision**	32.8%	52.8%	56.5%	67.8%	83.5%
	Recall	15.7%	44.1%	53.5%	65.5%	82%
	F1-Score	19.6%	46.4%	54.4%	65.8%	82.2%
SMOTE	**Accuracy**	17.8%	71.1%	69.5%	77.1%	87.7%
	Precision	36.4%	71.7%	69.8%	77.5%	88%
	Recall	18.8%	71.1%	69.5%	77.1%	87.7%
	F1-Score	22.5%	70.8%	69.4%	76.4%	87.5%
SMOTE + Under Sampling	**Accuracy**	16.8%	66.3%	63%	75.8%	87.6%
	Precision	34.9%	68%	63.2%	75.9%	88.1%
	Recall	16.8%	66.3%	63%	75.8%	87.6%
	F1-Score	21.4%	66.4%	63.1%	75.3%	87.4%
Keras	**Accuracy**	12.4%	66.5%	55.3%	68.6%	75.9%
	Precision	37.4%	69.7%	60.3%	70.8%	78.5%
	Recall	12.4%	66.5%	55.3%	68.6%	75.9%
	F1-Score	17.3%	67.6%	57.4%	69.4%	76.9%

Initial baseline scores revealed the challenges posed by class imbalances, with Naïve Bayes showing lower accuracy compared to other classifiers. Random Over Sampling improved accuracy across classifiers, with Logistic Regression exhibiting high accuracy and precision. However, Random Under Sampling may lead to a decrease in accuracy. SMOTE shows promising results, contributing to increased accuracy and precision in several cases. The combined SMOTE and Under Sampling approach presents a balanced trade-off, addressing class imbalances while maintaining competitive performance metrics.

Dimensionality reduced dataset

We applied PCA with a 99% of variance retention. Then, the reduced-dimension data was evaluated using various classifiers. This approach aims to capture the essential information in the dataset while reducing its dimensionality, potentially improving the efficiency and generalization of the machine learning models.

Performance	Naïve Bayes	Decision Tree	Random Forest	Logistic Regression
Accuracy	14.4%	13.2%	19.3%	10%
Precision	20.1%	19.5%	19.5%	16.6%
Recall	14.4%	13.2%	19.3%	10%
F1-Score	14.5%	15.2%	18.9%	11.4%

These low scores indicate that the application of PCA with a 0.99 variance retention threshold did not lead to a significant improvement in the classifiers' performance on the transformed data. It suggests that the reduced-dimensional representation may not capture sufficient information.

Data Augmentation

After the application of data augmentation using Keras ImageDataGenerator, which included settings such as rotation, width shift, height shift, zoom, and fill mode, a balanced dataset was created with

2250 images per class. Then, this augmented dataset was tested on various classifiers, yielding the following test metrics:

Performance	Naïve Bayes	Decision Tree	Random Forest	Logistic Regression
Accuracy	12.4%	55.3%	68.6%	75.9%
Precision	37.4%	60.3%	70.8%	78.5%
Recall	12.4%	55.3%	68.6%	75.9%
F1-Score	17.3%	57.4%	69.4%	76.9%

Comparing the scores directly, it's evident that the SMOTE-transformed data generally outperformed the dataset augmented with Keras ImageDataGenerator across various classifiers. Specifically, the SMOTE-transformed data showed higher scores in terms of accuracy, precision, recall, and F1 Score compared to the augmented dataset. While data augmentation is beneficial for introducing diversity in training data and improving generalization, SMOTE is designed to address imbalances in class distribution.

CNN

The implementation of Convolutional Neural Network (CNN) model incorporating both SMOTE synthetic data and augmented samples resulted in impressive accuracies of **93.3%** and **94.2%,** respectively. This dual strategy demonstrated its effectiveness in enhancing the model's performance. The use of CNNs, particularly when trained on a diverse dataset comprising synthetically generated instances from SMOTE and augmented data, contributed to improved generalization and robustness.

Conclusion

In summary, the study into data balancing techniques for an imbalanced image dataset revealed that SMOTE, Random Over Sampling, and a combination of SMOTE with Under Sampling substantially improved classifier performance, addressing the challenges posed by class imbalances. While dimensionality reduction through PCA did not yield significant improvements, data augmentation with Keras ImageDataGenerator showcased lower performance compared to SMOTE. The implementation of Convolutional Neural Networks incorporating SMOTE synthetic data and augmented samples demonstrated remarkable accuracy, emphasizing the effectiveness of a dual strategy for enhanced model generalization and robustness. Overall, the strategic application of data balancing techniques, especially SMOTE, holds promise for optimizing machine learning model performance in imbalanced image datasets.

- Github Link: https://github.com/amannuhman/F21DL-CW

Thank You