Validation of Data Augmentation for Highly Similar Images Using Dimensionality Reduction Techniques and Affinity-Diversity Metrics

1st Toledo-Rios, Juan Salvador
Facultad de Ingeniería
Universidad Autónoma de Querétaro
Querétaro, México

2nd Aceves-Fernandez, Marco Antonio
Facultad de Ingeniería
Universidad Autónoma de Querétaro
Querétaro, México
Corresponding author: marco.aceves@uaq.mx

3rd Cabrera-Hernández, Carmen
Facultad de Ingeniería
Universidad Autónoma de Querétaro
Querétaro, México

4th Tovar-Arriaga, Saúl
Facultad de Ingeniería
Universidad Autónoma de Querétaro
Querétaro, México

***Abstract*—The main objective of this study is to present a methodology to validate data augmentation using specific techniques when the original data are very similar to each other; to assess the diversity of the data, these two different approaches to data validation had to be combined, using quantitative metrics such as affinity and diversity used to estimate the similarity and variability of the data; in addition to dimensionality reduction techniques to visualize the distribution of the augmented data over the original data. In particular, it was given a focus when there is a class imbalance that adds another widespread problem when trying to work with complex models; this combined methodology provides a robust validation of the quality of the augmented data, contributing to a better generalization and performance of the data. For the particular case of this research, the intention is to validate the diversity of a highly similar data set obtained using eye-tracking techniques, in which 55 people are evaluated using three psychometric tests.**

***Index Terms*—Artificial Intelligence, Data Augmentation, t-SNE, Machine Learning, Data Validation, Deep Learning, Data Quality**

I. Introduction

One of the main challenges when working with complex machine learning models and, especially, deep learning models is the fact that a large amount of data is needed for the effectiveness of the different tasks (classification, detection, segmentation, etc.), where quantity and diversity are of utmost importance to avoid generalization of the models.), where the quantity and diversity are of utmost importance to prevent the generalization of the models. The problem is that large and labeled data sets are scarce or expensive to obtain, which is why, to overcome this problem, there is a technique called ***data augmentation***, which consists of creating artificial data controlled from accurate data [1], [2].

The problem of limited data sets affects the ability of models to learn representative features, especially in complex models such as Convolutional Neural Networks (CNN) when used in classification tasks [3] since they usually require a large number of examples to learn patterns to classify labeled instances correctly, the most common failure when having limited data is that the model usually fits the training data very well but fails to classify the validation data [4], [5].

Data augmentation is widely used in model training, particularly when data availability is limited [6]–[8]. By artificially expanding the training set, this method generates new instances from the original data through various transformations. These transformations include operations such as rotation, translation, and adjustments in brightness and contrast. In addition to increasing the dataset size, data augmentation enhances diversity, essential for improving the model's generalization ability. This process helps the deep learning model better generalize to unseen data and reduces the risk of overfitting [1], [9].

Another advantage of working with data augmentation is the ability to balance classes when you have a very noticeable imbalance, generating more examples of minority classes, which would allow the model to be able to learn the classes equally, helping it to improve performance in terms of accuracy and sensitivity [10], [11].

In this context, class imbalance is a problem when using Convolutional Neural Networks (CNN), as the lack of sufficient examples can lead to the model only learning the features of majority classes, making it unable to recognize the patterns of underrepresented classes [4], [12]. This issue is particularly detrimental in applications such as medical diagnostics, where class imbalance is a common and critical concern [13].

Although data augmentation is used and recognized for its effectiveness in increasing the amount of data, it is essential to validate the quality and usefulness of the augmented data. A problem that occurs when images are very similar is that when trying to validate with standard techniques such as cosine similarity or color analysis techniques is the fact that they usually yield results that indicate that there is minimal variation between the original image and the augmented image even though visually they look different; this is where dimensionality reduction techniques come into play, such as t-SNE (T-distributed Stochastic Neighbor Embedding) and PCA (Principal Component Analysis). These techniques make it possible to visualize the distribution of the original data and the augmented data in a lower dimensional space, facilitating the visual evaluation of the representativeness and diversity of the data generated [14]–[16].

This work's main contribution is validating the class balancing process through data augmentation and visualization techniques such as t-SNE when the augmentation and real data are very similar due to the characteristics of the image by itself. These dimensionality reduction techniques evaluate whether the augmented data behaves similarly to the original data.

In addition to t-SNE and PCA, it is necessary to use more ways to be able to validate the generation of the images, so two concepts have to be taken into account: affinity and diversity, which indicate the similarity between the generated images and the original images, the aim is to have a medium diversity and a low affinity, showing that the new data is diverse enough but not so different from the original ones, ensuring that the images still make sense when training the model [17].

Combining both validating methods provides a more robust methodology for assessing the diversity and effectiveness of class balancing through data augmentation. Assessing the representativeness and diversity of the augmented data obtains a more complete picture of the impact of data augmentation on model performance.

Finally, this study aims to provide a straightforward methodology to validate data expansion effectively in terms of diversity and class balancing. This not only impacts improving fairness in AI models but also offers a replicable method for other researchers experiencing similar problems. Ensuring that the augmented data is representative and diverse improves the model's ability to learn valuable features for all classes, leading to a fairer and more efficient system.

II. Related Works

Data augmentation is an essential technique when using robust models to solve data sparsity [18]–[20], but the validation of these synthetically created data is still a little explored area; a common approach is the reduction of dimensionality using statistical techniques such as t-SNE, as mentioned in the work of Van der Maaten and Hinton [14] who explored the distribution of the original data and the augmented data, validating whether the data are sufficiently diverse without being so original to the diverse ones. Another work that uses this technique is the one developed by Kwok and Wong [16], which seeks to analyze the quality of the generated data.

Another technique using dimensionality reduction to validate data augmentation is PCA, where research conducted by de Bro and Rasmus [15] and Abayomi-Alli [21] use PCA as a way to project the augmented data and the original data into a lower dimensional space, to assess the distribution and similarity of the data. Other research relies on the use of PCA to determine whether the augmented data represent the same general structure as the general data by visually validating the effectiveness of these data [22]–[24].

In addition to validations through visual inspections, some techniques propose quantitative metrics to validate data augmentation, such as affinity and diversity, which are essential metrics to measure the quality of the generated data, in the research proposed by Gao [12] makes use of Euclidean distance and correlation coefficient to quantify the similarity between the augmented data and the original data, proving helpful to determine whether the data are representative and diverse.

When working with deep learning, techniques based on the generated embeddings are proposed; Ozgode and Saygili [25] developed a confidence estimation algorithm for the embeddings generated by t-SNE, which allows validating the quality of the data augmented through these generated embeddings. This approach helps identify whether the generated data contributes positively to the training set or introduces unwanted noise.

Another approach that has been emerging for years is the use of generative adversarial networks (GANs) introduced by Goodfellow [26], which, through other research, are used to assess the quality of augmented data, as presented by Radford [27] and Zhang [28], which evaluate whether the data are realistic and practical for training classification models. The importance of GANs lies in their ability to generate synthetic data that is almost indistinguishable from accurate data.

All these works related to the research addressed using data sets with very well-defined classes, so individual approaches are usually insufficient as validation techniques when encountering similar original data. Combining two or more methods is necessary when wanting to estimate the quality of the augmented data.

III. Methodology

Figure 1 represents a flow chart of the methodology proposed for validating the images generated from the original data.

The methodology used to evaluate the quality of the generated images combines visualization techniques with quantitative metrics, as previously mentioned. This approach assesses the representativeness and diversity of the generated images, aiming to ensure that the applied transformations preserve or enhance the essential characteristics. The goal is to enable accurate representation of minority classes and improve the model's capacity to learn and generalize effectively.

Cabrera developed the database used in this research [29]; this database is used to be able to determine the level of attention of the individuals, which consists of three categories by obtaining eye trajectory data from 55 people through an eye tracker, evaluated through a series of 3 psychometric tests; table I represents the distribution of the original data of each category for each test.

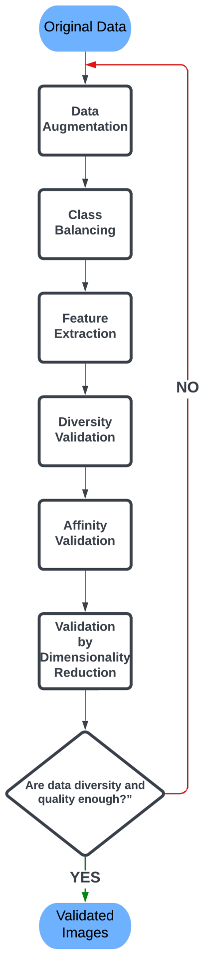
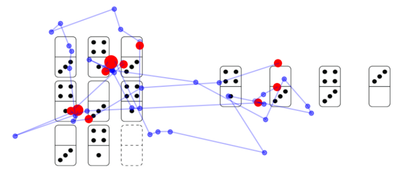


Fig. 1: Flow chart proposed

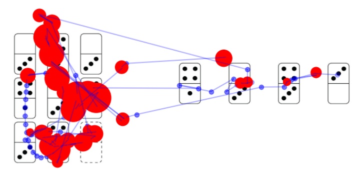
Test	Low Level	Medium Level	High Level	Total
Domino	21	4	18	43
Unfolded Cubes	4	17	21	42
Figure Series	27	8	12	47

TABLE I: Labels distribution

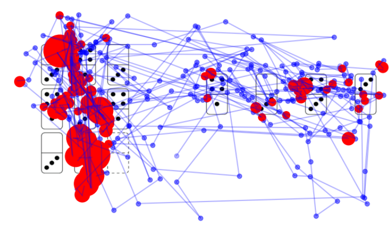
It is possible to determine the imbalance of the categories, in which some classes barely contribute to 10% of the total data, so when training an artificial intelligence model with this information, it would be tough to determine that the model learns enough to classify that minority category compared to the other two majority categories, another problem is that when generating three categories of the same image depending only on specific unique characteristics per category may be unclear how they differ from each other, figure 2 represents the three categories of a test, in which the similarity between each category can be observed. If individual validation techniques analyze the images, little distribution of images will be obtained. To solve these two problems, we propose this combined validation by balancing the classes through data augmentation, allowing to guarantee that the images are unique and, at the same time, are representative of the original images.



(a) High Level



(b) Medium Level



(c) Low Level

Fig. 2: Example of the data classification used in the database to measure attention levels [29]

A. *Data Augmentation*

Once the number of subjects per class is determined, the first step is to balance these classes by data augmentation. Class balancing is paramount as it helps the model have the same data per category. In the project, a function was developed to apply stochastic transformations such as rotation, translation, zoom, and color change to the images in the dataset. These transformations add visual variation to the images, improving the model's generalization and reducing overfitting to limited data.

The ranges of each transformation are as follows:

- **Translation:** This function shifts the image horizontally and vertically. Parameters: $x` : X - axis$ (horizontal) offset in pixels. Range: $[-10, 10]$. $y` : Y - axis$ (vertical) offset in pixels. Range: $[-10, 10]$. Description: A translation matrix moves the image in the 2D plane. The displacement range is randomly defined between -10 and 10 pixels, which allows the image to move both left/right and up/down.

- **Zoom:** This function zooms in on the image. Parameters: factor: Zoom factor applied to the image. Range: $[1.0, 1.5]$. Description: The image is resized using a random zoom factor. A factor greater than 1.0 indicates a zoom-in, while a factor less than 1.0 (though not used here) would indicate a zoom-out. The resulting image is centered to maintain the same resolution.

- **Rotation:** This function rotates the image around its center. Parameters: angle: Rotation angle in degrees. Range: $[-30, 30]$. Description: Uses a rotation matrix to rotate the image around its center. The rotation angle is randomly selected between -30 and 30 degrees, allowing clockwise and counterclockwise rotations.

- **Brightness:** This function adjusts the image's brightness. Parameters: Factor: Brightness adjustment factor. Range: $[0.8, 1.2]$. Description: The brightness of the image is modified by a multiplicative factor. A factor greater than 1.0 increases the brightness, while a factor less than 1.0 reduces it. This allows you to create variations of the image with different brightness levels.

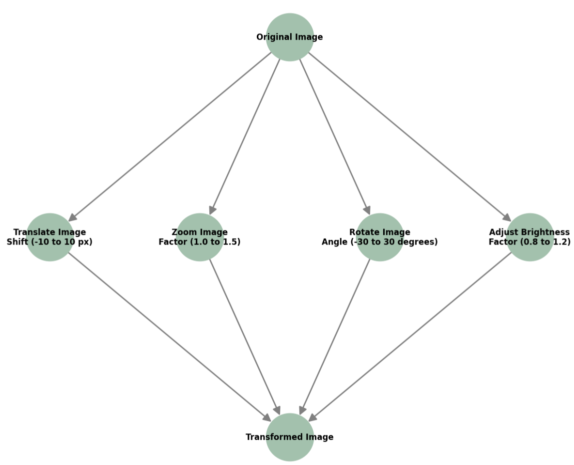


Fig. 3: Data Augmentation diagram process

Figure 3 shows how the data augmentation process of each image is randomized to avoid bias towards a specific transformation that benefits the model, ensuring that the images are generated uniformly. Likewise, to avoid developing more images than possible per test, it was estimated how many unique images per image can be generated; as each transformation is randomly selected, it is guaranteed that only one transformation per image is applied; the calculation of these images is presented in table II.

Transformation Type	**Combinations**
Translation (X, Y from -10 to 10 px)	441
Zoom (Factor 1.0 to 1.5, step 0.01)	50
Rotation (Angle -30 to 30 degrees, step 1)	61
Color Adjustment (Factor 0.8 to 1.2, step 0.01)	41
Total Possible Combinations	593

TABLE II: Summary of Discretized Transformation Combinations

As can be seen in table II, it is possible to generate 593 unique images without repeating the process, so for this project, a limit of 700 images per class was established, having a total of 2100 images, thus finalizing the balancing of the classes through data augmentation.

B. Feature extraction

When balancing the data, to perform the analysis of the augmented data, it is necessary to extract the characteristics of the images; this is a previous stage to the analysis since it helps us to obtain the characteristics of the images numerically, in the case of the project a CNN was used with the frozen dense cells for the extraction of the characteristics. Specifically, the model used is a VGG16 [30]; this allows us to represent the images as vectors of information of characteristics of high dimension, facilitating the analysis of similarity and diversity.

C. Diversity Evaluation

The diversity of the dataset refers to the measure of variability between the original and synthetic elements. This metric is critical in data augmentation validation applications, as it allows one to evaluate the different points within the feature space.

To measure and evaluate the diversity of the augmented data, it is necessary to calculate the Euclidean distance between the generated images and the real images; this helps to identify whether the set of augmented images covers an adequate range of variations. The Euclidean distance between images is calculated as follows,

$$\text{Diversity} = \frac{1}{\binom{N}{2}} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \|x_i - x_j\| \tag{1}$$

Where $\binom{N}{2} = \frac{N(N-1)}{2}$ is the total number of possible combinations of two distinct elements of the data set, thus ensuring that each pair (i, j) is counted only once [31].

This version of the metric is frequently used in analysis to assess the dispersion of points within a group. Using unordered combinations avoids counting duplicates, which provides an accurate and consistent representation of the internal variability of the data [32]. In the context of machine learning, measuring the diversity of a data set is critical to assess whether the points generated using data augmentation techniques are sufficiently varied, which can improve the generalization of the model [33].

D. Affinity Evaluation

The affinity measures the similarity between each class's synthetic and original images. This allows us to determine if the variations are significant or too similar. The following parameters must be calculated to determine and evaluate the affinity per class:

- **Calculate the Average of Characteristics per Class:** For each class c, it calculates the average of the features of all images belonging to that class:

$$\bar{x}_c = \frac{1}{N_c} \sum_{i=1}^{N_c} x_i \tag{2}$$

  where x_i is the feature vector of the image i in the class c, and N_c is the number of images in the class c.

- **Calculate the Average Distance per Class:** The affinity is calculated as the average of the Euclidean distances between all pairs of images in the same class:

$$\text{Affinity}_c = \frac{1}{N_c(N_c - 1)} \sum_{i=1}^{N_c} \sum_{j \neq i} \|x_i - x_j\| \tag{3}$$

  where $\| \cdot \|$ represents the Euclidean norm.

- **Average Affinity:** The overall affinity is obtained by averaging the affinity of all classes:

$$\text{Affinity}_{\text{avg}} = \frac{1}{C} \sum_{c=1}^{C} \text{Affinity}_c \tag{4}$$

  Where C is the total number of classes.

Once the affinity has been calculated, the following analysis should be performed, where a value Close to 0 indicates that the images within a class are very similar. This may be desirable because the images are consistent and homogeneous, which is helpful for model training. However, if the affinity is too low, it may be a sign of overfitting if the model is seeing too little range of variation. A value near 1.0 indicates that the images within a class are very different. This can be a problem because images within the same class do not share enough common characteristics, making it difficult for the model to learn about the factors that define that class [17].

E. Validation with Dimensionality Reduction Techniques

Once diversity and affinity metrics have been calculated, it is necessary to visually validate the distribution of the generated images, for which dimensionality reduction techniques are needed to visualize them in a two-dimensional space since the vectors generated through feature extraction are of high dimensionality. The methods used are as follows:

1) PCA: It is a dimensionality reduction technique used to transform a high dimensional data set into a lower dimensional data set [15], [34] while retaining as much of the variability of the data as possible. This is achieved by identifying principal directions (also called principal components). These components are linear combinations of the original variables and are ordered such that the first principal component captures most of the variability; the second captures most of the variability that was not captured by the first, and so on [31], [33], [35].

Visualizing the original and augmented data in the same space allows visualizing if the augmented data are well distributed concerning the original data. Suppose the augmented data are distributed around or over the original data. In that case, it indicates that the data augmentation is adequate since data similar to the original data are being generated but with variations that do not make them identical; if there is no variation or they are too concentrated among themselves, it means that the augmented data do not represent sufficient diversity.

To calculate the principal components, the following steps are performed:

1) **Standardize the Data:** First, the data set is standardized so each feature has a mean of zero and a variance of one. This can be written as:

$$X' = \frac{X - \mu}{\sigma} \tag{5}$$

   Where:
   - X is the original data set,
   - μ is the mean vector of the data set, and
   - σ is the vector of standard deviations.

2) **Compute the Covariance Matrix:** Next, the covariance matrix C of the standardized data is calculated:

$$C = \frac{1}{n-1} X'^T X' \tag{6}$$

   Where:
   - X' is the standardized data set,
   - n is the number of observations.

3) **Compute Eigenvectors and Eigenvalues:** The next step is to compute the eigenvectors (v) and eigenvalues (λ) of the covariance matrix:

$$Cv = \lambda v \tag{7}$$

   The eigenvectors represent the principal components, and the eigenvalues represent the variance explained by each principal component.

4) **Order and Select Principal Components:** The eigenvectors are ordered by the eigenvalues in decreasing order to determine the principal components that capture the most variance. Typically, the first k eigenvectors are selected, where k is the desired number of dimensions.

5) **Project the Data:** Finally, the standardized data set is projected onto the selected eigenvectors to obtain the transformed data:

$$Y = X'V_k \tag{8}$$

Where:

- Y is the transformed data set,
- V_k is the matrix containing the first k eigenvectors.

This process reduces the dimensionality of the data while retaining as much information as possible in terms of variance.

2) t-SNE: Like PCA, it is a dimensionality reduction technique, mainly used for the visualization of data in low dimensional spaces (2 or 3 dimensions) [36]; it is beneficial because the vector is high dimensional and serves to find inherent patterns since it seeks to understand the relationship between groups of data. In this case, it is applied to compare the augmented data with the original data.

If the augmented data and the original data form overlapping clusters, they share very similar characteristics, which indicates that the augmented data are of adequate quality and reflect the properties of the original data. On the other hand, if the clusters are widely separated and dispersed, it indicates that the augmented data does not represent the characteristics of the original distribution.

t-SNE converts the high-dimensional Euclidean distances between data points into conditional probabilities representing similarities. The technique tries to minimize the divergence between two distributions: one that measures pairwise similarities in the high-dimensional space and one that measures pairwise similarities in the low-dimensional space. The objective function is defined as:

$$C = \sum_i \sum_{j \neq i} P_{ij} \log \frac{P_{ij}}{Q_{ij}} \tag{9}$$

Where:

- P_{ij} represents the similarity between high-dimensional points x_i and x_j.
- Q_{ij} represents the similarity between low-dimensional points y_i and y_j.

The similarity P_{ij} is calculated using a Gaussian distribution centered at point x_i:

$$P_{j|i} = \frac{\exp(-\|x_i - x_j\|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|x_i - x_k\|^2 / 2\sigma_i^2)} \tag{10}$$

where σ_i represents the variance of the Gaussian centered on x_i.

The low-dimensional similarity Q_{ij} is defined using a Student t-distribution with one degree of freedom:

$$Q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq l} (1 + \|y_k - y_l\|^2)^{-1}} \tag{11}$$

By minimizing the cost function C, t-SNE aims to ensure that similar points in the high-dimensional space remain close in the low-dimensional embedding. In contrast, dissimilar points are modeled far apart. This makes t-SNE a powerful tool for visualizing the data structure and assessing the quality of the augmented dataset.

IV. Results

This section presents the results obtained in each methodology step from the techniques described and implemented in the study. The results will focus on how data augmentation improves the quality and representativeness of the original dataset.

In addition, comparative analyses will be presented between the characteristics of the augmented images and the original ones through affinity and diversity. Finally, the results will validate whether the data augmentation has contributed to a good distribution of the data, and visual validations will be made through dimensionality reduction techniques.

After balancing the classes presented in the table I, the classes present the results in the table III. Estimates were made heuristically to determine that in comparison of metrics and computational expense, 700 images per test were sufficient to estimate all metrics; as can be seen, the classes have the same amount of images with a total of 2100 images per test.

Test	Low Level	Medium Level	High Level	Total
Domino	700	700	700	2100
Unfolded Cubes	700	700	700	2100
Figure Series	700	700	700	2100

TABLE III: Labels distribution after balancing

In addition, the augmented images can be seen randomly in figure 4, which generates the transformations of each image; for the visual representation, only the results of the increased data from one test will be shown, even though all three tests have been balanced, the other results will show all three tests.

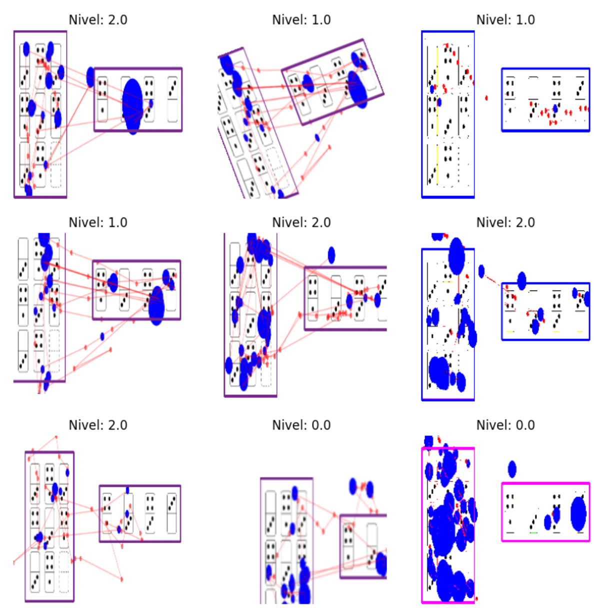


Fig. 4: Data augmentation representation

After balancing the classes, the similarity between the augmented and original images can be observed, which prompted further exploration of further validation techniques. As described in the methodology, the diversity and affinity metrics are calculated, and the results of each test are presented in the following tables:

Diversity	Domino	Cubes	Figures
Original data	16.9536	15.8990	17.6986
Data augmentation	24.0319	26.4942	27.1704

TABLE IV: Diversity of Classes

Table IV shows how, after implementing data augmentation, the diversity of each test increased without losing the main characteristics of each original image, indicating that there is a large number of examples per class that can be passed to the model to work with, demonstrating a high range of variations of each image, allowing to work with complex models.

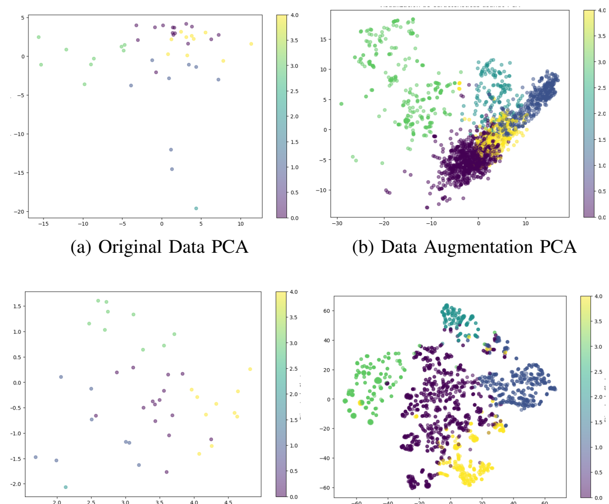
Affinity by Class	Domino	Cubes	Figures
Class 1	18.4720	15.7035	17.36037
	24.1867	26.0103	26.7926
Class 2	16.2699	17.9823	19.8720
	22.4138	26.5377	26.2795
Class 3	16.1854	14.5449	18.3725
	24.6234	26.0252	27.5354

TABLE V: Affinity by Class

As table V shows, a steady increase in the affinity between the two established metrics can be observed, indicating that data augmentation has been successful in generating new images that are more similar to the original images of the same class, which is critical to improving the performance of the model in classification tasks.

The next step involves visually inspecting the distribution of the augmented data compared to the original data using dimensionality reduction techniques. This will include comparing both PCA and t-SNE, where a uniform distribution is expected. Achieving such a distribution would validate the findings proposed in this research.

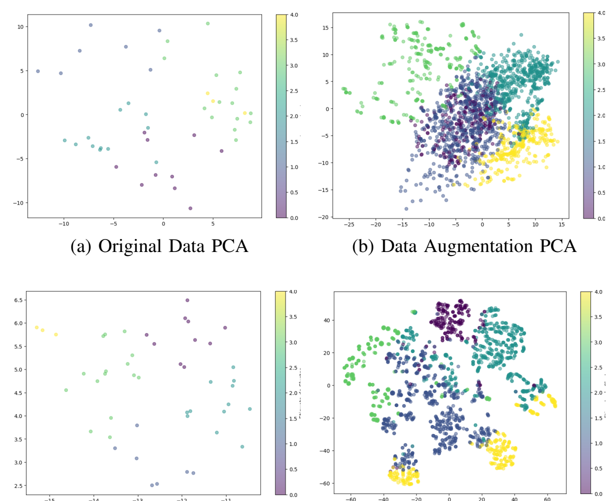
It can be observed that in figure 5, the distribution of the classes of the domino test has a high variability, complementing with the quantitative metrics, it can be determined that the data increase has a high quality, ensuring that there is a high range of examples per class without these being necessarily the same as the original image but also not so different that it does not make sense with the original image.



(a) Original Data PCA (b) Data Augmentation PCA

(c) Original Data t-SNE (d) Data Augmentation t-SNE

Fig. 5: Analysis using PCA and t-SNE for Domino test data with and without data augmentation.



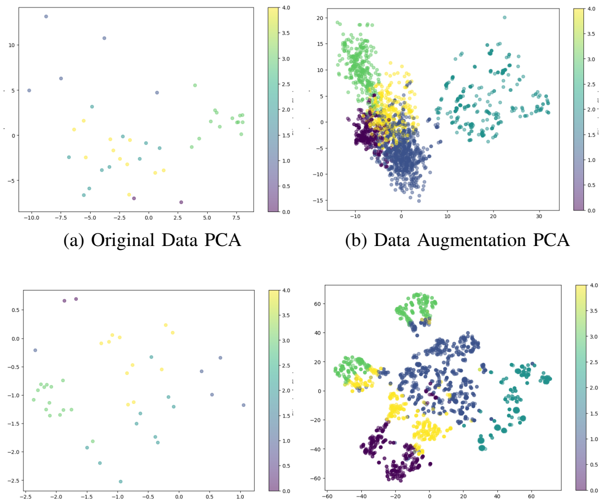
(a) Original Data PCA (b) Data Augmentation PCA

(c) Original Data t-SNE (d) Data Augmentation t-SNE

Fig. 6: Analysis using PCA and t-SNE for Figure series test data with and without data augmentation.

As with the domino test, it can be estimated after seeing the figure 6 that the diversity of the augmented data for the figure series test has a high but representative dispersion, preserving the main characteristics of the original data.

The behavior of the cube test data is very similar to the domino test; after looking at the figure 7, that shows all tests have a high variability of images from the original images, even though the original images are very similar to each other on visual inspection.



(a) Original Data PCA (b) Data Augmentation PCA

(c) Original Data t-SNE (d) Data Augmentation t-SNE

Fig. 7: Analysis using PCA and t-SNE for Cubes test data with and without data augmentation.

V. Conclusions

The validation of data augmentation on sets of very similar images is still a task that has been little explored; besides being a prevalent task along with class imbalance, it is expected that with this research, a robust methodology to determine validation by combining quantitative techniques such as component analysis will be proposed, such as diversity and affinity in the quantitative metrics part, as well as t-SNE and PCA in the component analysis techniques, since they allow us, in addition to visual inspection, to determine how varied our set is before and after data augmentation.

One of the findings is that it is possible to validate the variability of the augmented images even though the original images are very similar between classes, allowing us to work with more data on these characteristics in the future. As observed in the results section, the distributions were uniform, avoiding excessive conglomerations on the original images and increasing affinity and diversity after balancing the classes.

This may impact the future by allowing more robust models to be used when data have similar characteristics, helping to avoid overfitting and making them more sensitive to detecting unique attributes of each class.

Future work remains to validate these new datasets through artificial intelligence models, such as machine learning and deep learning models, and to examine what other metrics could help validate the information as more data becomes available.

In conclusion, the paper highlights the importance of validating data augmentation techniques using a comprehensive approach that combines visualization tools (t-SNE and PCA) and quantitative metrics (Affinity and Diversity). These methods provide an in-depth analysis of the quality and effectiveness of the data generated, ensuring that the data generated not only increases the number of samples per class but also preserves the essential characteristics of the class to which they belong.

References

[1] C. Shorten and T. M. Khoshgoftaar, “A survey on image data augmentation for deep learning,” *Journal of big data*, vol. 6, no. 1, pp. 1–48, 2019.

[2] L. Perez and J. Wang, “The effectiveness of data augmentation in image classification using deep learning,” *arXiv preprint arXiv:1712.04621*, 2017.

[3] X. Zhao, L. Wang, Y. Zhang, X. Han, M. Deveci, and M. Parmar, “A review of convolutional neural networks in computer vision,” *Artificial Intelligence Review*, vol. 57, no. 4, p. 99, 2024.

[4] J. L. Leevy, T. M. Khoshgoftaar, R. A. Bauder, and N. Seliya, “A survey on addressing high-class imbalance in big data,” *Journal of Big Data*, vol. 5, no. 1, pp. 1–30, 2018.

[5] V. López, A. Fernández, S. García, V. Palade, and F. Herrera, “An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics,” *Information sciences*, vol. 250, pp. 113–141, 2013.

[6] T. Talaei Khoei, H. Ould Slimane, and N. Kaabouch, “Deep learning: Systematic review, models, challenges, and research directions,” *Neural Computing and Applications*, vol. 35, no. 31, pp. 23 103–23 124, 2023.

[7] S. E. Whang, Y. Roh, H. Song, and J.-G. Lee, “Data collection and quality challenges in deep learning: A data-centric ai perspective,” *The VLDB Journal*, vol. 32, no. 4, pp. 791–813, 2023.

[8] A. Shrestha and A. Mahmood, “Review of deep learning algorithms and architectures,” *IEEE access*, vol. 7, pp. 53 040–53 065, 2019.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[10] D. A. Dablain, C. Bellinger, B. Krawczyk, and N. V. Chawla, “Efficient augmentation for imbalanced deep learning,” in *2023 IEEE 39th International Conference on Data Engineering (ICDE)*. IEEE, 2023, pp. 1433–1446.

[11] S. C. Wong, A. Gatt, V. Stamatescu, and M. D. McDonnell, “Understanding data augmentation for classification: when to warp?” in *2016 international conference on digital image computing: techniques and applications (DICTA)*. IEEE, 2016, pp. 1–6.

[12] G. Haixiang, L. Yijing, L. Shuai, Y. Jing, K. Hongrich, and W. Jin, “Learning from class-imbalanced data: Review of methods and applications,” *Expert Systems with Applications*, vol. 73, pp. 220–239, 2017.

[13] L. A. Jeni, J. F. Cohn, and F. De La Torre, “Facing imbalanced data–recommendations for the use of performance metrics,” *Proceedings of the IEEE*, vol. 104, no. 12, pp. 2185–2197, 2017.

[14] L. Van der Maaten and G. Hinton, “Visualizing data using t-sne,” *Journal of Machine Learning Research*, vol. 9, no. 11, pp. 2579–2605, 2008.

[15] R. Bro and A. K. Smilde, “Principal component analysis,” *Analytical methods*, vol. 6, no. 9, pp. 2812–2831, 2014.

[16] T. Kwok and A. Wong, “Data augmentation techniques for improved performance in deep learning models,” *Artificial Intelligence Review*, vol. 53, no. 2, pp. 1021–1040, 2020.

[17] R. Gontijo-Lopes, S. J. Smullin, E. D. Cubuk, and E. Dyer, “Affinity and diversity: Quantifying mechanisms of data augmentation,” *arXiv preprint arXiv:2002.08973*, 2020.

[18] M. Wattenberg, F. Viégas, and I. Johnson, “How to use t-sne effectively,” *Distill*, 2016.

[19] J. A. Lee, D. H. Peluffo-Ordóñez, and M. Verleysen, “Multi-scale similarities in stochastic neighbour embedding: Reducing dimensionality while preserving both local and global structure,” *Neurocomputing*, vol. 169, pp. 246–261, 2015.

[20] Y. Saeys, S. Van Gassen, and B. N. Lambrecht, “Computational flow cytometry: helping to make sense of high-dimensional immunology data,” *Nature Reviews Immunology*, vol. 16, pp. 449–462, 2016.

[21] O. O. Abayomi-Alli, R. Damaševičius, M. Wieczorek, and M. Woźniak, “Data augmentation using principal component resampling for image recognition by deep learning,” in *Artificial Intelligence and Soft Computing: 19th International Conference, ICAISC 2020, Zakopane, Poland, October 12-14, 2020, Proceedings, Part II 19*. Springer, 2020, pp. 39–48.

[22] J. Shlens, “A tutorial on principal component analysis,” *arXiv preprint arXiv:1404.1100*, 2014.

[23] D. Cozzolino, W. U. Cynkar, R. G. Dambergs, N. Shah, and P. Smith, "Multivariate methods in grape and wine analysis," *International Journal of Wine Research*, vol. 1, pp. 123–130, 2009.
[24] K. Kjeldahl and R. Bro, "Some common misunderstandings in chemometrics," *Journal of Chemometrics*, vol. 24, no. 8, pp. 558–564, 2010.
[25] B. Ozgode Yigin and G. Saygili, "Confidence estimation for t-sne embeddings using random forest," *International Journal of Machine Learning and Cybernetics*, vol. 13, no. 12, pp. 3981–3992, 2022.
[26] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.
[27] A. Radford, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.
[28] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *International conference on machine learning*. PMLR, 2019, pp. 7354–7363.
[29] M. del Carmen Cabrera-Hernández, C. A. García-Ezquerra, M. A. Aceves-Fernández, J. C. Pedraza-Ortega, and S. Tovar-Arriaga, "A dataset on eye movement tracking during the resolution of neuropsychological tests on a screen," *Data in Brief*, vol. 55, p. 110601, 2024.
[30] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
[31] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.
[32] B. S. Everitt, S. Landau, M. Leese, and D. Stahl, *Cluster Analysis*. Wiley, 2011.
[33] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009.
[34] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley interdisciplinary reviews: computational statistics*, vol. 2, no. 4, pp. 433–459, 2010.
[35] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *Journal of Educational Psychology*, vol. 24, pp. 417–441, 498–520, 1933.
[36] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.