IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 7 October 2022, accepted 31 October 2022, date of publication 14 November 2022, date of current version 18 November 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3221936

RESEARCH ARTICLE

Extended a Priori Probability (EAPP): A Data-Driven Approach for Machine Learning Binary Classification Tasks

VICENT ORTIZ CASTELLÓ[1], FRANCISCO JAVIER PÉREZ-BENITO[1], OMAR DEL TEJO CATALÁ[1], ISMAEL SALVADOR IGUAL[1], RAFAEL LLOBET[1,2], AND JUAN-CARLOS PEREZ-CORTES[1,3]

[1]Instituto Tecnológico de Informática (ITI), Universitat Politècnica de València, 46022 Valencia, Spain
[2]Department of Computer Systems and Computation (DSIC), Universitat Politècnica de València, 46022 Valencia, Spain
[3]Department of Computing Engineering (DISCA), Universitat Politècnica de València, 46022 Valencia, Spain

Corresponding author: Ismael Salvador Igual (issalig@iti.upv.es)

This work was supported in part by the Generalitat Valenciana through the Valencian Institute of Business Competitiveness (IVACE) Distributed Nominatively to Valencian Technological Innovation Centers under Project IMAMCN/2021/1, in part by the Cervera Network of Excellence Project in Data-Based Enabling Technologies (AI4ES) Co-Funded by the Centre for Industrial and Technological Development—E. P. E. (CDTI), and in part by the European Union through the Next Generation EU Fund within the Cervera Aids Program for Technological Centers under Project CER-20211030.

ABSTRACT The *a priori* probability of a dataset is usually used as a baseline for comparing a particular algorithm's accuracy in a given binary classification task. ZeroR is the simplest algorithm for this, predicting the majority class for all examples. However, this is an extremely simple approach that has no predictive power and does not describe other dataset features that could lead to a more demanding baseline. In this paper, we present the Extended *A Priori* Probability (EAPP), a novel semi-supervised baseline metric for binary classification tasks that considers not only the *a priori* probability but also some possible bias present in the dataset as well as other features that could provide a relatively trivial separability of the target classes. The approach is based on the area under the ROC curve (AUC ROC), known to be quite insensitive to class imbalance. The procedure involves multiobjective feature extraction and a clustering stage in the input space with autoencoders and a subsequent combinatory weighted assignation from clusters to classes depending on the distance to nearest clusters for each class. Class labels are then assigned to establish the combination that maximizes AUC ROC for each number of clusters considered. To avoid overfit in the combined feature extraction and clustering method, a cross-validation scheme is performed in each case. EAPP is defined for different numbers of clusters, starting from the inverse of the minority class proportion, which is useful for a fair comparison among diversely imbalanced datasets. A high EAPP usually relates to an easy binary classification task, but it also may be due to a significant coarse-grained bias in the dataset, when the task is previously known to be difficult. This metric represents a baseline beyond the *a priori* probability to assess the actual capabilities of binary classification models.

INDEX TERMS A priori probability, EAPP, clustering, autoencoder, semisupervised, combinatory, bias.

I. INTRODUCTION

From text analysis [1], [2] or pedestrian detection [3], [4] to healthcare [5], [6], it is unquestionable that Artificial Intelligence (AI) has become more and more useful in almost any technological challenge. The paradigms of Machine Learning (ML) and Deep Learning (DL) in particular, have become state of the art in several scientific fields. However, is such complex technology needed to solve any challenge? Can data affect the knowledge extracted using DL? Is a given trained network as good as it seems or is it just because of the dataset? With all of these questions, it is clear

The associate editor coordinating the review of this manuscript and approving it for publication was Kok-Lim Alvin Yau.

that it is necessary to assess how the algorithms are really performing.

The *a priori* probability is used in classification tasks as the lower bound or baseline that any predictor should achieve. The ZeroR classifier, which simply predicts the majority class, is used for establishing this baseline, which is useful as a benchmark for comparison with other classifiers. Any classifier performing poorer than the ZeroR classifier is considered to have no predictive value.

However, a more demanding and realistic baseline could be established considering not only the *a priori* probability, but also possible biases in the dataset and/or trivial class separation. For example, a binary classifier reaching an accuracy of 0.8 on a dataset with *a priori* probability of 0.5 but with a relatively obvious bias in the dataset that allows 80% of its samples to be predicted trivially, should be considered as good as a ZeroR classifier.

In this paper, we propose a method to compute a more challenging metric than *a priori* probability, which can be used as a baseline to determine the prediction capabilities of a given classifier. This metric, which we call Extended *A Priori* Probability (EAPP), takes into account not only the *a priori* probability, but also other underlying characteristics of the sample, such as the presence of biases in the data or an obvious class separation. Thus, *bias* in this paper refers to the features of the dataset that are not related to the pure data but external conditions, i.e., different brightness in images taken with different equipment, different acquisition procedures depending on the human knowledge for this sample, or any artifact not naturally present in the data.

There are several types of bias as reported on [7]. First, selection bias is present when a dataset prefers a particular type of image (e.g. indoor or outdoor scenes). Second, capture bias can affect the dataset, i.e., different hospitals apply different settings to the RX equipment and category. Third, label bias appears when different labelers assign different labels to the same type of object. Finally, negative set bias defines what the dataset considers to be the rest of the world.

In recent years, there has been a renewed interest in the study of bias [8], [9], [10], [11], [12]. According to [13], while the *known unknowns* are wrong predictions with low confidence that can cast doubt on accuracy, the *unknown unknowns* are wrong predictions with high confidence of truth that can mask dataset-intrinsic representation problems. Since the classical AI definition proclaims imitation of humans, in terms of ML, the *unknown unknowns* are extremely harmful, raising controversy around ML usage. Attenberg et al. designed an experiment to prove humans can detect bias, *unknown unknowns*, where a machine can not [13]. Bansal et al. proposed a way to automatically discover *unknown unknowns* [14].

A wide range of methods are focused on the removal of bias in the dataset in many different ways. Alvi et al. presented an algorithm to remove spurious features for the task at hand [15]. Others like Khosla et al. proposed a method to undo the bias by learning two sets of weights: bias and visual world weights [16], Tommasi et al. use DeCAF features in order to mitigate bias [17], and Clark combines biased and unbiased ensembles to remove dataset biases [18].

Hoffman and Tzeng proposed a discriminative domain adaptation to minimize the impact of bias on the task [19], [20] which was outperformed by the method given by [21], where stacked autoencoders based on domain adaptation are used to extract domain invariant features. Finally, Zhao et al. presented a way to constrain the training corpus to reduce bias [9]. Also, several feature-adaptive methods have been proposed for the removal of selection bias (or covariate shift) [22], [23] and CNN descriptors demonstrated to be robust against this bias [24].

Some of the works presented before use supervised methods to reduce the bias in a dataset mainly by building a classifier that is not affected by bias or using cross-dataset performance drop [7], [23] as a generalization measure. However, our main objective is to obtain a simple semi-supervised metric that allows evaluation of the ease or complexity of the task beyond the well-established baseline for any binary classification (namely, the *a priori* probability, that is, the proportion of examples belonging to the majority class). If the task is previously known to be difficult but the EAPP is high, then the dataset is likely to have a heavy bias, as a simple algorithm may be able to tell the difference between examples from two classes without supervision, using only the locality of the observations in the representation space.

II. MATERIALS AND METHODS

A. DATASETS

We tested the performance of the EAPP method covering a wide range of scenarios.

- Since the EAPP deals with binary classification tasks, a subset of the handwritten digits dataset MNIST [25] was extracted. The images of "1" and "7", which are relatively similar, were compared to those of "8".
- The ImageNet dataset [26] consists of 3.2 million images covering up to 20000 categories. Under the assumption that the categories *mushroom* and *wedding* may have different environmental elements, we selected the images of these two classes looking at whether the background that do not contain the object could introduce bias to the dataset.
- Our previous work [12] showed the presence of bias in the BIMCV-PADCHEST [27] chest x-ray image dataset. In this paper, we assess the EAPP metric in that real, biased, task.
- The EAPP definition is independent of the input data type. In this sense, we evaluated the metric in a subset of the nCOV2019 dataset [28]. This dataset was shown to have potential bias sources [11] and after replicating the data processing proposed in the previously mentioned work, we evaluated EAPP.

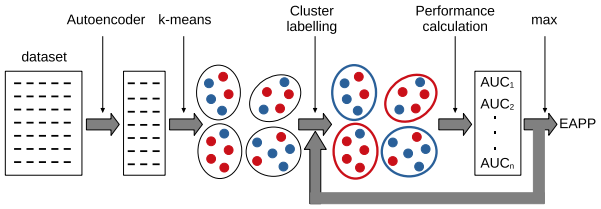


FIGURE 1. EAPP process.

B. DATA PREPROCESSING

The image datasets analyzed contained images of multiple shapes. As Neural Networks are used and scale invariance is not our focus, the task is simplified resizing all images to the same shape to perform feature extraction. Therefore, they were cropped as a square, keeping the same image center, and resized to 128 × 128 pixels. Regarding numerical datasets, all raw features were separately normalized (mean 0 and standard deviation 1).

C. EAPP

The goal of EAPP is to assess how well a non-supervised feature extraction method automatically splits classes into different clusters. The method is based on assigning the same class to all the examples that fall into the same cluster. This is done iteratively for different numbers of clusters and combinations of class assignments. A probability of belonging to a class is assigned to each observation depending on its distance to the centroids of the nearest positive and negative classes' clusters.

The process is divided into 3 stages: feature extraction, clustering, and combinatory analysis. The stages of the complete process, shown in Figure 1, are described below.

1) FEATURE EXTRACTION

To compute the EAPP metric, labels must not be used during the training phase. Therefore, feature extraction is performed using unsupervised learning methods only. Algorithms such as Convolutional AutoEncoders (CAEs) are valid candidates. An autoencoder forces its inputs to fit into a reduced latent space and then tries to rebuild the original input from that smaller representation. The structure of the network can be seen in Figure 2.

2) CLUSTERING

Additionally, this algorithm should group together observations that have similar features, as they are likely to be from the same class. Therefore, clustering algorithms such as K-means are useful to find the inner clusters that group samples of the dataset. Even if only two classes are present within the data, we cannot assume that the latent representations of both classes are linearly separable. Therefore, using a number of clusters equal to the number of classes might not represent an adequate EAPP value. For instance, classifying the XOR problem may be an easy task but cannot be performed using two clusters. The optimal number of clusters is unknown,

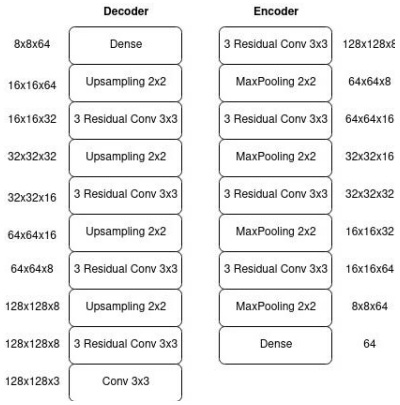


FIGURE 2. Structure of the CAE.

so the algorithm should explore multiple values up to a certain limit. In this sense, the limitation of cluster cardinality k controls the likelihood of overestimating the fit by chance. Arguably, if the number of maximum clusters is small enough compared to the number of samples in the dataset, the overfitting probability is bounded.

Moreover, it is important to highlight that the *a priori* probability for both classes gives a clear insight into the minimum number of clusters to consider to fit the data well. This is the case if the clustering method tends to group a similar number of examples in each cluster, as is the case of K-means, assuming that instances of the same class are close in the input space. Given a binary classification task, in this scenario, the k values should start from $1/p_0$ upwards, being p_0 the *a priori* probability for the minority class. For example, for $p_0 = 0.5$ (a perfectly balanced dataset), the minimum cluster number is $k = 2$, while for $p_0 = 0.25$, the range starts at $k = 4$. Interestingly, this matter allows us to establish a fair comparison among different datasets regardless of their diverse *a priori* probability values. If $1/p_0$ is the exact k value for the number of clusters needed to establish a comparison in a dataset, then, this value can be inferred by linear interpolation between both $\lfloor(1/p_0)\rfloor$ and $ceil(1/p_0)$ EAPP values. For instance, if $p_0 = 0.4$, then the exact k is 2.5, so $\text{EAPP}(2.5) \approx (\text{EAPP}(2) + \text{EAPP}(3))/2$.

Because of this, all graphs are plotted similarly: the X axis starts from $k = \lfloor(1/p_0)\rfloor$ upwards, but the zone below $k = 1/p_0$ is shaded because it falls below the minimum theoretical k value to establish a fair comparison for clusters of similar size, as explained above. Furthermore, to assess the level of improvement obtained by chance by reaching higher k values, a baseline curve for randomly shuffled classes (that is, respecting the original p_0 values for the task) is presented. For any k, a number of different random shuffles are performed and the maximum EAPP values reached are saved as a population, from which the mean and the 95% confidence intervals are plotted in blue.

3) HYBRID TECHNIQUE

To achieve better clustering for EAPP computation, one option is to perform clustering and feature extraction

separately. In this case, these processes are independent, with no restrictions on their training metric (i.e. loss). However, both stages may also be considered together, as a part of a particular algorithm, with goal interdependence (for example, with a loss function composed of two terms, one for each process).

Assuming that an incremental approach to the clustering phase would benefit the latent space configuration, the initial number of clusters is set at $k = \lfloor(1/p_0)\rfloor$, and this value is increased after the current configuration converges. After each convergence, the feature extraction network and previous cluster centers are kept, and an additional cluster is added. This is performed by splitting the largest cluster along its axis of largest variance. Then, the training phase of the feature extraction network and clustering is performed until convergence. In each iteration of the incremental approach, the closest cluster center index for each sample is stored.

4) COMBINATORY ANALYSIS

Once a particular clustering scheme has been computed for a given feature representation, this proposal aims to assess the ease or complexity of the classification task by searching a higher baseline above the *a priori* probability, dependent only on the number of instances of each class. The underlying idea is, once the clustering process converges for a range of different cluster cardinalities k in the training stage, to save the model information (namely, the centroids) and apply this clustering to new test data but, for each cluster assignation (for each k considered), assigning all the possible different combinations of binary labels to the clusters, leaving out the trivial ones (the extreme all-negative and all-positive correspondences, since they would lead to a strongly unbalanced, useless classification). For example, for $k = 2$, the only chances are cluster 1 assigned to the positive class and cluster 2 to the negative class, and conversely (we discard the all-negative and all-positive assignations, as said before). This correspondence can be represented as a binary number, where each digit represents a particular cluster and the particular value it takes (0 or 1) represents its correspondence to the negative or positive class, respectively. Similarly, for $k = 3$, the only chances are 001, 010, 011, 100, 101, and 110, since the extreme 000 and 111 are discarded. In the general case, for each k, $(2^k - 2)$ cases are computed, and the combination of better results is taken as a representative of this k, and the particular metrics used are to be discussed.

Moreover, we sort the set of observations based on the distance to the inferred clusters. The assignation to a binary class for each example in each combination is then performed depending on the distance between that instance and the centroids of the nearest positive and negative classes clusters, so that a continuous score is available.

Hence, to sum up, for each k, we compute all the possible binary assignations leaving out both extreme configurations and we obtain a similarity indicator depending on the distance to the nearest positive and negative clusters. Then, using this sorting procedure, we compute a performance index, in this case the Area Under the ROC, and select the assignation which leads to the best score. Finally, we plot the maximum AUC ROC values for k. Note that, if the particular divisive clustering method used is not hierarchical, there is no guarantee for the curve to be increasing in a monotonic way, but this behavior is predominant since with more clusters there are more chances to fit the data accurately. It is also worth to note that as it is mentioned before, EAPP evaluates the complexity of a task and this could also be seen as a randomness test [29] where non-random data would get higher EAPP than random data.

5) CROSS-VALIDATION TRAINING SCHEME

Our complete algorithm is semi-supervised, so it faces the problem of overfitting as do others of this kind. In fact, even some basic algorithms such as the naive K-means clustering, which is totally unsupervised, could suffer from it [30]. For this reason, a 10 cross-validation was established as a standard for all our experiments. To be precise, for each dataset, the whole process is split into two parts: the training process, in which 90% of the data is used to learn all clustering schemes for any k (centroids), and a particular internal representation, and the testing process, in which the remaining 10% is processed with all these parameters obtained by training. This procedure is repeated 10 times for each experiment to allow a fairer comparison throughout all datasets. However, for nCov2019, the cross-validation process is 50-fold because of its reduced size, as it was not appropriate to suppress 10% of the data for training in each fold. Thus, we trained with 98% of data in each iteration.

III. RESULTS

A. IMPLEMENTATION DETAILS

The experiments were carried out on diverse datasets, most of them based on images, but also on non-image, structured data. This approach is valid regardless of the nature of data as long as a binary classification problem underlies the task. For reasons of clarity, the experiments on image datasets are presented first in order of increasing difficulty and finally, an additional experiment on a numerical dataset is shown. Note that curves are plotted with k starting at $\lfloor(1/p_0)\rfloor$, being p_0 the *a priori* probability for the minority class in each task, but with the zone between $\lfloor(1/p_0)\rfloor$ and $1/p_0$ shadowed because it falls outside the range of fair comparison. Again for reasons of clarity, the vertical axis starts at 0.5, as EAPP is defined in terms of AUC ROC, whose mathematical expectation is 0.5 for a random assignation. Finally, for assessing the likelihood of adapting the data by chance as k increases, a baseline curve with confidence intervals is presented, representing the best data fit for each k for random shuffle of the original data.

B. IMAGE DATASETS

In this subsection, we aim to study the EAPP behavior for diverse binary classification tasks within different

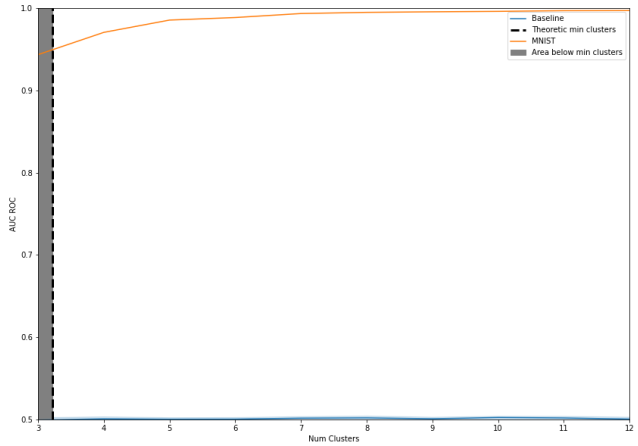


FIGURE 3. EAPP trend for multiple cluster number values for reduced MNIST dataset (digit '8' vs '1' and '7' combined).

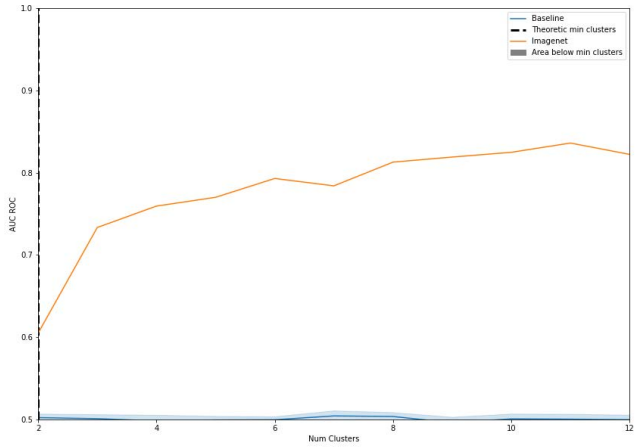


FIGURE 4. EAPP trend for multiple cluster number values for reduced ImageNet dataset (wedding vs mushroom).

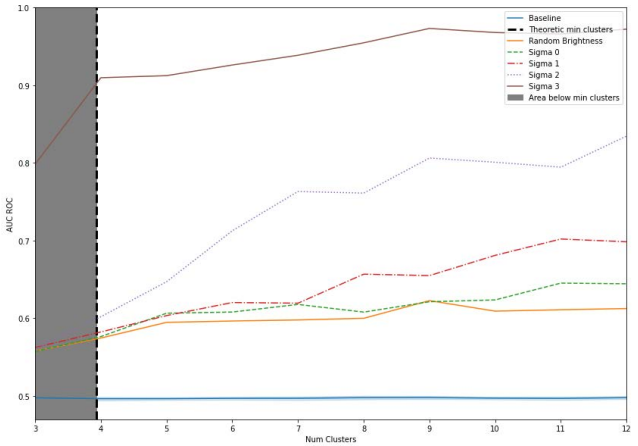


FIGURE 5. EAPP trend for multiple cluster number values for BIMCV dataset (cases vs controls).

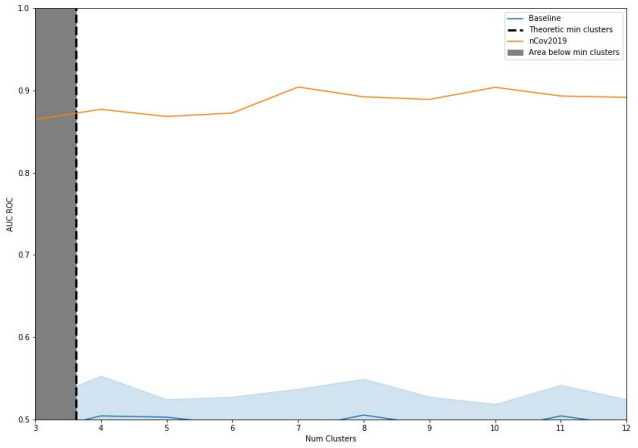


FIGURE 6. Latent space and cluster assignment for several cluster numbers.

well-known image datasets. For simplicity, we present our experiments sorted by increasing perceived difficulty of the detection task.

1) MNIST

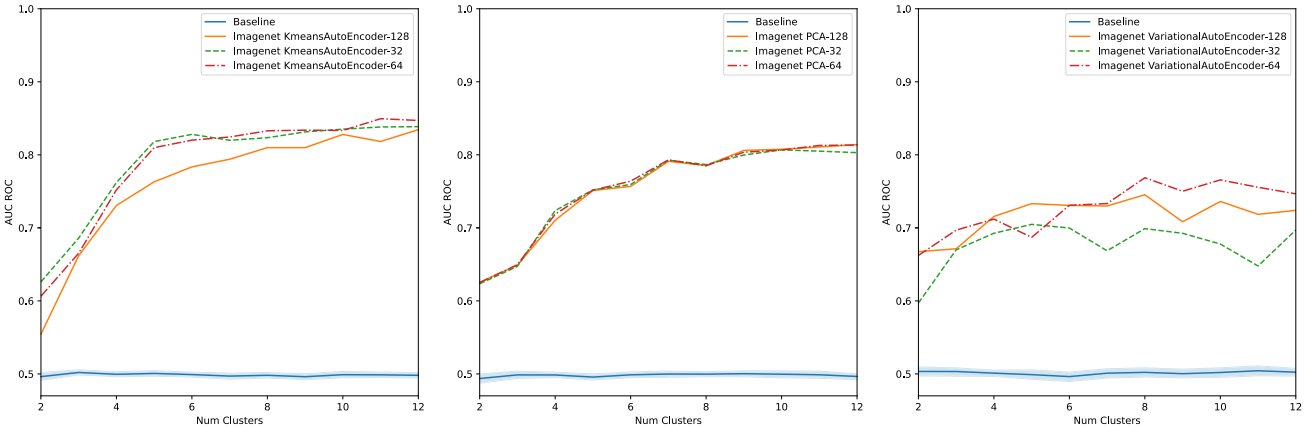
MNIST is a standard database of handwritten digits commonly used for training classifiers. It consists of the 10 different arabic numerals, but our approach is defined for binary classification. Therefore, we selected the set {'1', '7'} for class w_0, and {'8'} for w_1. Since all digits are represented evenly in the dataset, p_0 is around 0.33 (minority class). Therefore, the results for this MNIST task, from $k = \lfloor (1/p_0) \rfloor = 3$, are presented in Figure 3.

It comes as no surprise that this binary classification task yields very high EAPP values already from $k = 2$. It is the expected behavior for such an easy task. It is noticeable that the '8' vs other task is easier than other combinations because the digits '1' and '7' are visually similar, so the clustering process is more likely to group them together than for other combinations (for example, '1' and '8' digits). EAPP is able to correctly differentiate among groups of digits. This simple task serves as a starting point from which we will reach more difficult binary tasks, such as the following more difficult classification tasks.

2) ImageNet

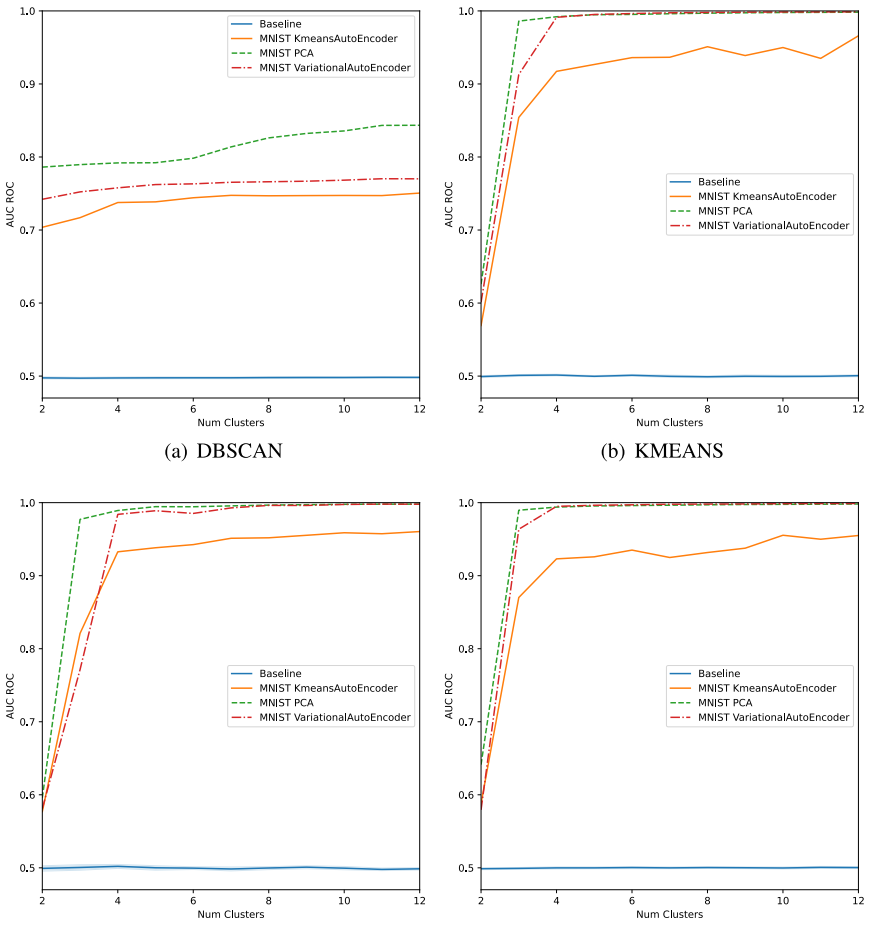
With MNIST experiments, a particular spatial distribution for bright pixels is easily noticeable. To overcome this, a slightly more complex dataset which more image richness is used: ImageNet. It is another well-established image dataset containing more than 20,000 categories, but again, we focused on two visually different categories (wedding and mushroom) so as to work with an appropriate subset for binary classification. The results for this ImageNet task, from $k = \lfloor (1/p_0) \rfloor = 2$, are shown in Figure 4.

This experiment also leads to significant EAPP values. However, they are not as high as in MNIST (Figure 3) since the task is now more complex, as the variability in images increased noticeably. In this regard, EAPP for



(a) KMAE (b) PCA (c) VAE

FIGURE 7. Comparison between different latent space sizes and feature extraction algorithms for a specific clustering method in the ImageNet dataset.



(a) DBSCAN (b) KMEANS

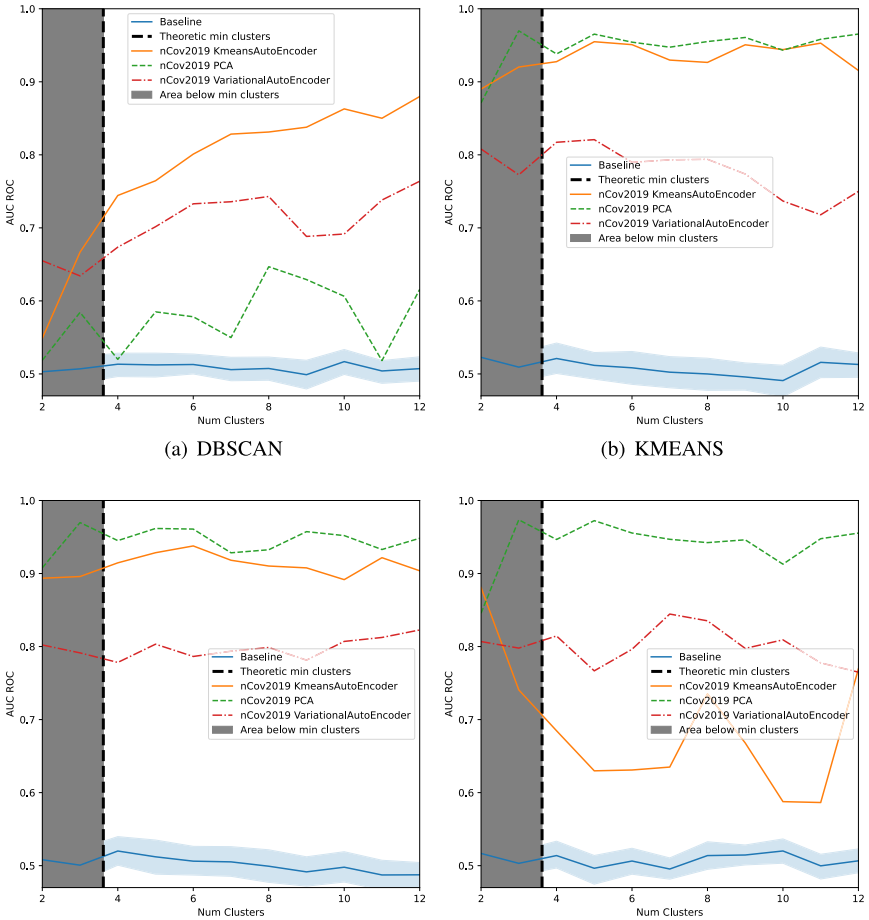
(c) Spectral Clustering (d) Agglomerative Clustering

FIGURE 8. Comparison between different clustering methods and feature extraction algorithms for MNIST dataset.

the ImageNet task seems to plateau around 0.8, whilst for MNIST, it reaches almost 1.

3) BIMCV

In our previous paper [12], we carefully designed a morphological segmentation scheme by which an important bias was detected in some chest X-ray image datasets (mainly BIMCV). That methodology consisted of comparing the classification performance of the whole images with images where areas of the lungs and the background had been removed. From this, we could check that the background was accountable for most of the detection accuracy, despite the



(a) DBSCAN (b) KMEANS

(c) Spectral Clustering (d) Agglomerative Clustering

FIGURE 9. Comparison between different clustering methods and feature extraction algorithms for nCov2019 dataset.

fact that all the information of the disease is expected to be inside the lungs. In particular, this is a case of capture bias where different settings on acquisition are used as task information by the CNN or other algorithms. Therefore, our aim is to check if this fine-grained bias was easily perceptible using our simple EAPP method. The EAPP results for BIMCV dataset, from $k = \lfloor (1/p_0) \rfloor = 3$, are presented in Figure 5 (*sigma 0* curve in green).

The results show moderate EAPP values around 0.6, suggesting that the generic approach is not as powerful as the *ad-hoc* morphological method that excluded the lung. To check if further biases are detected, we introduced different levels (sigma 1, 2 and 3) of controlled class-dependent noise to this dataset in terms of increased average grayscale levels to images of one of the classes. This shift is visually perceived as a slight increase in brightness of the overall image. As expected, higher levels of EAPP are noticeable in Figure 5 as the increase gets larger (*sigma* represents the number of standard deviations of brightness added to one class of the dataset), with a good separability when the gray levels are increased at least by 2σ.

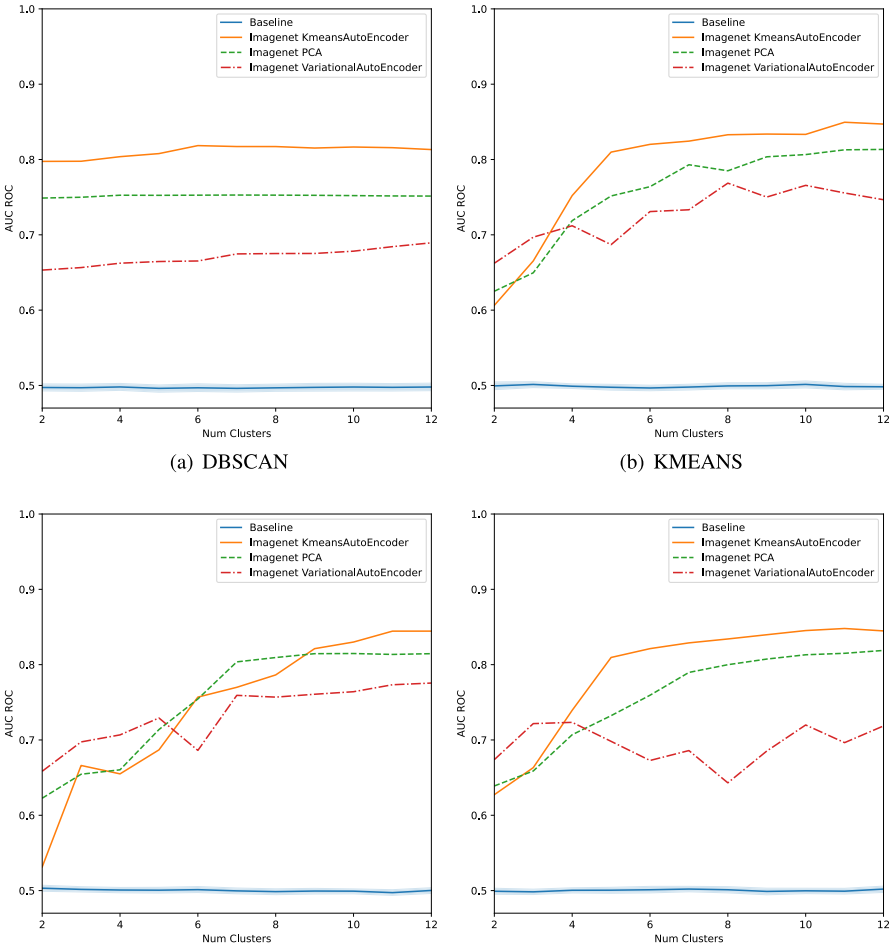
C. NUMERICAL DATASETS

The EAPP calculation was then performed on the numerical, structured nCOV dataset to evaluate the difficulty of a binary classification task not dealing with image data.

In this case, p_0, which accounts for the *a priori* probability of the minority class, is 0.28, being p_1 the *a priori* probability of the majority class equal to 0.72. As can be seen in Figure 6, the EAPP value using 4 clusters is around 0.86, which is significantly higher than the one expected from a random classifier represented by the line and the shaded zone plotted in blue.

Figure 6 shows how the 2D latent representation of the nCOV2019, as k increases, can be automatically classified by a set of clusters built without taking into account the class labels. This EAPP value confirms the analysis of [28], where a high bias in the dataset is reported.

Additional experiments are reported on the Appendix. They show the results for different latent space sizes, where it can be seen that this parameter does not significantly affect the results. Furthermore, different clustering methods have been explored, such as DBSCAN [31], while agglomerative



(a) DBSCAN (b) KMEANS

(c) Spectral Clustering (d) Agglomerative Clustering

FIGURE 10. Comparison between different clustering methods and feature extraction algorithms for ImageNet dataset.

clustering and hierarchical clustering [32] have also been used with worse results than other methods like K-means. Similarly, agglomerative clustering and hierarchical clustering have also been explored, yielding similar results to the ones obtained by K-means (see Figs. 8, 9, 10 and 11). It is also worth mentioning that PCA can be a good alternative to KmeansAutoEncoder, as it is faster to train and achieves similar results. Finally, an alternative approach to reduce the evaluation's computational cost is commented.

In short, a correlation between the perceived difficulty of the task and EAPP values can be noticed. This behavior is consistent with our hypothesis. However, if a significant amount of hidden bias is present in the data, EAPP values could also potentially rise. Notwithstanding that, the difference between ease and bias is not yet detected by the algorithm.

IV. DISCUSSION

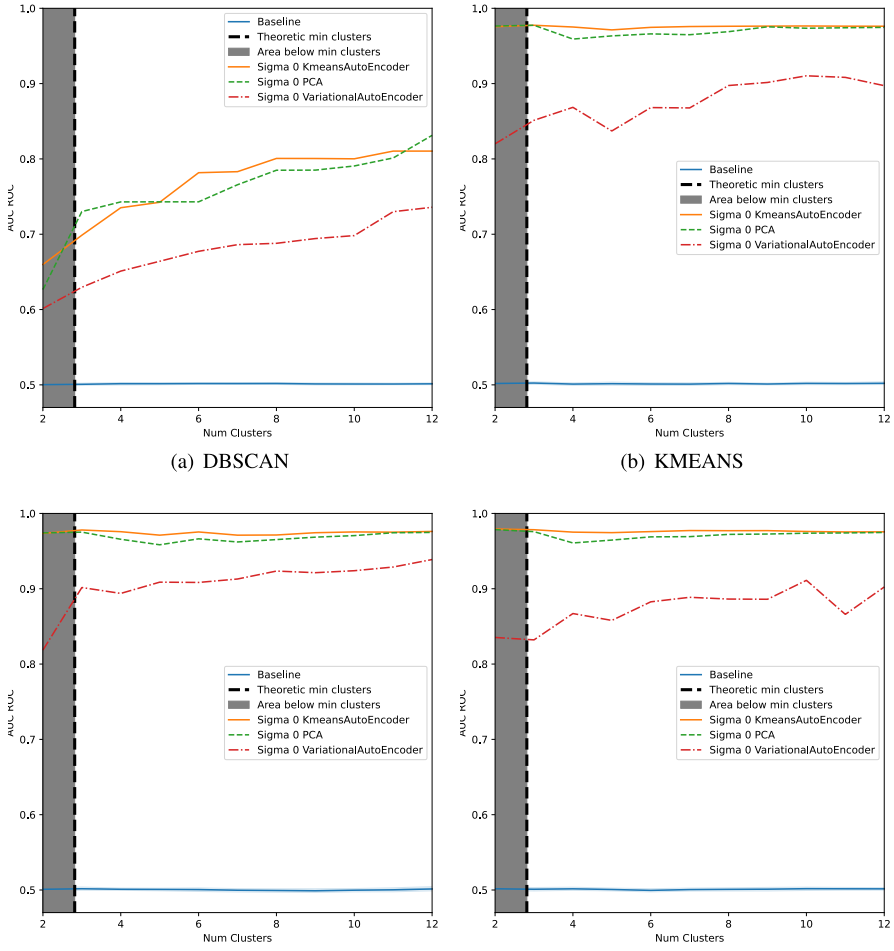
A. STATE OF THE ART COMPARISON

Our contribution aims to establish a new and more informative baseline for the performance of binary classification tasks. Although any internal metric might be used, we selected the AUC ROC as the performance metric. Due to the nature of our approach, the results are not meant to compete with the performance of supervised algorithms, but to offer a lower bound.

B. STRENGTHS AND WEAKNESSES

The algorithm proposed is able to give an estimate of the hardness of any dataset for a binary classification task further than the naive *a priori* probability, which is simply the proportion of the majority class to the dataset size. Therefore, it may be used as a baseline for the AUC ROC obtained from any binary classification algorithm. Also, as could be seen in the results, our methodology can be applied both to images and other structured data.

Nevertheless, that hardness can be intrinsic or due to a particular bias in the dataset that allows correct classification of a vast majority of examples without considering the real nature of the evidently meaningful attributes. This effect may be especially interesting in images, in which a particular classification algorithm (i.e. convolutional deep neural networks)



(a) DBSCAN (b) KMEANS

(c) Spectral Clustering (d) Agglomerative Clustering

FIGURE 11. Comparison between different clustering methods and feature extraction algorithms for BIMCV sigma 3 dataset.

could take enormous advantage of the fine-grain, complex variables extracted from the image, which would be almost indistinguishable for a human. These cases are frequently related to some bias in the dataset that allows these powerful algorithms to clearly outperform the expected outcomes for the classification task with no real basis on the real predictive attributes.

C. FURTHER WORK

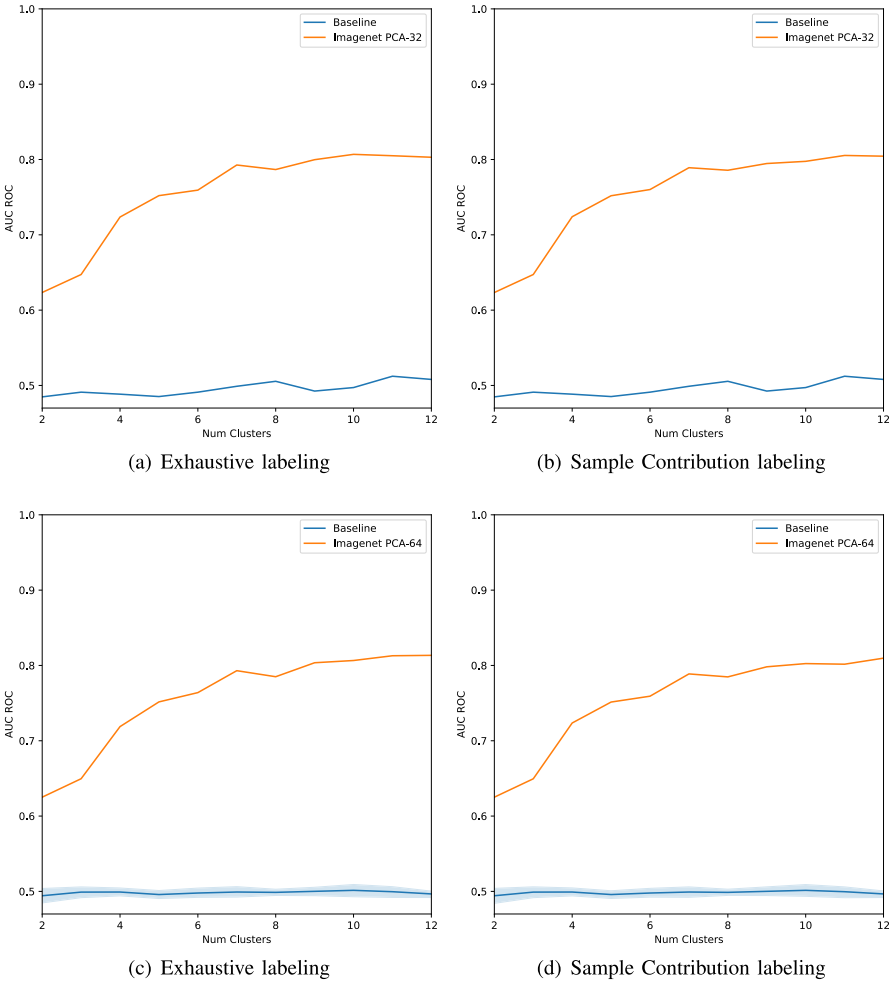
The current proposal is based on a particular internal representation and clustering technique, but our paradigm can accommodate different methods. Therefore, internal representations for data, such as those obtained with other dimensionality reduction techniques (PCA, ZCA whitening…), may be tested. Similarly, other clustering methods like Gaussian Mixtures or Ward, may be used. Further research is needed in order to gain more insight into the difference between ease of the classification task and bias presence in the dataset, so as to infer this information automatically, without human intervention. Furthermore, our method could be extended to other problems such as multiclass classification, multilabel classification or regression.

Moreover, model interpretability is highly relevant in this procedure as features that affect the score most can be identified. This means that, should the features not be related to class information, bias can be detected.

V. CONCLUSION

This paper proposes a method to calculate a more informative metric set than the simple *a priori* probability for estimating the difficulty or bias of the data in the context of a binary classification task.

The method is based on the separability of the target classes in a given latent space of representation: it tries to find the assignation of clusters and classes that performs the best regarding the area under the ROC curve. This maximum value is registered for any number of clusters k, so a graph can be plotted for k, starting from the inverse of the minority class proportion to set a fair comparison among imbalanced datasets. Moreover, a cross-validation scheme is included



(a) Exhaustive labeling (b) Sample Contribution labeling

(c) Exhaustive labeling (d) Sample Contribution labeling

FIGURE 12. **Comparison between different cluster labeling methods.**

to avoid overfitting and assure independence between the training and testing stages.

By using some well-known datasets, our method has proven beneficial to preliminarily assess the difficulty of a binary classification task and suggest a certain level of bias in cases where the task is perceived to be easy, and a high EAPP is found. Thus, our metric represents a baseline beyond the *a priori* probability to assess the actual capabilities of binary classification models.

APPENDIX. COMPARISON OF DIFFERENT LATENT SPACE SIZES, CLUSTERING METHODS AND FEATURE EXTRACTORS

In this section, we present a comparison of several latent space sizes (32, 64 and 128), clustering methods (DBSCAN, K-means, Agglomerative Clustering and Spectral Clustering) and feature extractors (the proposed KmeansAutoEncoder, PCA and Variational AutoEncoders). Remark that, in contrast with the original experiments where the K-means is performed along feature extraction training, in this section, all clustering methods are applied after training the feature extraction algorithm.

A. LATENT SPACE STUDY

Figure 7 shows how different latent spaces affect the results in the ImageNet dataset. As can be seen, this hyperparameter does not significantly affect the results achieved by each feature extractor. Thus, for simplicity purposes, we selected a latent space size of 64 for the paper.

B. CLUSTERING METHOD AND FEATURE EXTRACTOR STUDY

Given the results above wihch are similar for 32, 64 and 128 sizes, we finally select a 64-dimensional latent space (except for nCov2019, which is has only 27 variables and the latent space size is set to 2 dimensions).

DBSCAN is not a clustering method that contains the number of clusters as a hyperparameter. However, it includes a distance hyperparameter that could be swept to achieve the same comparable values to other clustering methods. Nonetheless, in the following Figures 8, 9, 10 and 11, it can be seen that DBSCAN cannot be properly used for the EAPP score. Moreover, these figures show that PCA and KmeansAutoEncoder achieve the best results compared to Variational AutoEncoder (VAE). Therefore, PCA can be a

good alternative to KmeansAutoEncoder, as it is faster to train and achieves similar results.

C. EXHAUSTIVE CLUSTER LABELING VS SAMPLE CONTRIBUTION CLUSTER LABELING

As described in this article, to assess if the samples were effectively clustered in an unsupervised manner, an exhaustive combinatory analysis was performed. This procedure evaluates all the possible labels each cluster could take, and the best combination is returned. However, this evaluation has an exponential cost that depends on the number of clusters. Another alternative approach is that each sample contributes to its closest cluster center with its class weighted by the distance. This way provides a good alternative to the exhaustive one, as seen in Figure 12, scaling linearly with the number of samples and obtaining almost the same results.

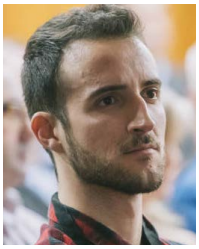
REFERENCES

[1] Sitender, S. Bawa, M. Kumar, and Sangeeta, "A comprehensive survey on machine translation for English, Hindi and Sanskrit languages," *J. Ambient Intell. Hum. Comput.*, vol. 2021, pp. 1–34, Sep. 2021.
[2] I. Lauriola, A. Lavelli, and F. Aiolli, "An introduction to deep learning in natural language processing: Models, techniques, and tools," *Neurocomputing*, vol. 470, pp. 443–456, Jan. 2021.
[3] V. Ortiz Castelló, O. del Tejo Catalá, I. S. Igual, and J.-C. Perez-Cortes, "Real-time on-board pedestrian detection using generic single-stage algorithms and on-road databases," *Int. J. Adv. Robotic Syst.*, vol. 17, no. 5, Sep. 2020, Art. no. 1729881420929175.
[4] V. O. O. Castelló, I. S. S. Igual, O. del Tejo Catalá, and J.-C. Perez-Cortes, "High-profile VRU detection on resource-constrained hardware using YOLOv3/v4 on BDD100K," *J. Imag.*, vol. 6, no. 12, p. 142, Dec. 2020.
[5] F. J. Pérez-Benito, F. Signol, J.-C. Perez-Cortes, A. Fuster-Baggetto, M. Pollan, B. Pérez-Gómez, D. Salas-Trejo, M. Casals, I. Martínez, and R. Llobet, "A deep learning system to obtain the optimal parameters for a threshold-based breast and dense tissue segmentation," *Comput. Methods Programs Biomed.*, vol. 195, Oct. 2020, Art. no. 105668.
[6] W. Xia, B. Hu, H. Li, W. Shi, Y. Tang, Y. Yu, C. Geng, Q. Wu, L. Yang, Z. Yu, D. Geng, and Y. Li, "Deep learning for automatic differential diagnosis of primary central nervous system lymphoma and glioblastoma: Multi-parametric magnetic resonance imaging based convolutional neural network model," *J. Magn. Reson. Imag.*, vol. 54, no. 3, pp. 880–887, Sep. 2021.
[7] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *Proc. CVPR*, Jun. 2011, pp. 1521–1528.
[8] C. Meske, E. Bunde, J. Schneider, and M. Gersch, "Explainable artificial intelligence: Objectives, stakeholders, and future research opportunities," *Inf. Syst. Manage.*, vol. 39, pp. 1–11, Jan. 2021.
[9] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang, "Men also like shopping: Reducing gender bias amplification using corpus-level constraints," 2017, *arXiv:1707.09457*.
[10] B. Kim, H. Kim, K. Kim, S. Kim, and J. Kim, "Learning not to learn: Training deep neural networks with biased data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9012–9020.
[11] C. Sáez, N. Romero, J. A. Conejero, and J. M. García-Gómez, "Potential limitations in COVID-19 machine learning due to data source variability: A case study in the nCov2019 dataset," *J. Amer. Med. Inform. Assoc.*, vol. 28, no. 2, pp. 360–364, Feb. 2021.
[12] O. D. T. Catala, I. S. Igual, F. J. Perez-Benito, D. M. Escriva, V. O. Castello, R. Llobet, and J.-C. Perez-Cortes, "Bias analysis on public X-ray image datasets of pneumonia and COVID-19 patients," *IEEE Access*, vol. 9, pp. 42370–42383, 2021.
[13] J. Attenberg, P. Ipeirotis, and F. Provost, "Beat the machine: Challenging humans to find a predictive model's 'unknown unknown,'" *J. Data Inf. Qual.*, vol. 6, no. 1, pp. 1–17, Mar. 2015.
[14] G. Bansal and D. Weld, "A coverage-based utility model for identifying unknown unknowns," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, 2018, pp. 1463–1470.
[15] M. Alvi, A. Zisserman, and C. Nellåker, "Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings," in *Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops*, 2018, pp. 556–572.
[16] A. Khosla, T. Zhou, T. Malisiewicz, A. A. Efros, and A. Torralba, "Undoing the damage of dataset bias," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2012, pp. 158–171.
[17] T. Tommasi, N. Patricia, B. Caputo, and T. Tuytelaars, *A Deeper Look at Dataset Bias*. Cham, Switzerland: Springer, 2017, pp. 37–55.
[18] C. Clark, M. Yatskar, and L. Zettlemoyer, "Don't take the easy way out: Ensemble based methods for avoiding known dataset biases," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, K. Inui, J. Jiang, V. Ng, and X. W., Eds. Hong Kong, Nov. 2019, pp. 4067–4080.
[19] J. Hoffman, E. Tzeng, T. Darrell, and K. Saenko, "Simultaneous deep transfer across domains and tasks," in *Domain Adaptation in Computer Vision Applications*. Berlin, Germany: Springer, 2017, pp. 173–187.
[20] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7167–7176.
[21] M. Jia, J. Wang, Z. Zhang, B. Han, Z. Shi, L. Guo, and W. Zhao, "A novel bearing transfer fault diagnosis method based on MMD guided domain adversarial mechanism," *Meas. Sci. Technol.*, vol. 33, no. 1, p. 015109, 2021.
[22] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars, "Unsupervised visual domain adaptation using subspace alignment," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2960–2967.
[23] B. Gong, Y. Shi, F. Sha, and K. Grauman, "Geodesic flow kernel for unsupervised domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2066–2073.
[24] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "DeCAF: A deep convolutional activation feature for generic visual recognition," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 647–655.
[25] Y. LeCun, C. Cortes, and C. J. C. Burges. (Jan. 2022). *The MNIST Database of Handwritten Digits*. [Online]. Available: http://yann.lecun.com/exdb/mnist/
[26] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[27] A. Bustos, A. Pertusa, J.-M. Salinas, and M. de la Iglesia-Vayá, "PadChest: A large chest X-ray image dataset with multi-label annotated reports," *Med. Image Anal.*, vol. 66, Dec. 2020, Art. no. 101797.
[28] B. Xu, B. Gutierrez, S. Mekaru, K. Sewalk, L. Goodwin, A. Loskill, E. L. Cohn, Y. Hswen, S. C. Hill, and M. M. Cobo, "Epidemiological data from the COVID-19 outbreak, real-time case information," *Sci. data*, vol. 7, no. 1, pp. 1–6, 2020.
[29] S. Wolfram and M. Gad-el Hak, "A new kind of science," *Appl. Mech. Rev.*, vol. 56, no. 2, pp. B18–B19, 2003.
[30] S. Bubeck and U. von Luxburg, "Overfitting of clustering and how to avoid it," V1, Tech. Rep. inria-00185780, Nov. 2007. [Online]. Available: https://hal.inria.fr/inria-00185780
[31] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. KDD*, vol. 96, 1996, pp. 226–231.
[32] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. Hoboken, NJ, USA: Wiley, 2009.

VICENT ORTIZ CASTELLÓ was born in Oliva, València, Spain, in 1991. He received the bachelor's and master's degrees in industrial engineering from the Universitat Politècnica de València, València, in 2015, the bachelor's degree in telecommunications engineering from the Universitat Oberta de Catalunya, Barcelona, in 2022, and the master's degree in artificial intelligence from Universidad Internacional Menéndez Pelayo, Santander, in 2022. He is currently pursuing the master's degree in telecommunications engineering with the Universitat Oberta de Catalunya. He worked as a Research Assistant at the AI2—Instituto de Automática e Informática Industrial, Universitat Politècnica de València, in 2015. From 2016 to 2018, he was a Researcher in biomechanics at the Institut de Biomecànica de València (IBV). Since 2018, he has been a Researcher in artificial intelligence and computer vision at the Instituto Tecnológico de Informática (ITI). His research interests include artificial intelligence, machine learning, deep learning, model interpretability, and computer vision.

FRANCISCO JAVIER PÉREZ-BENITO was born in Salamanca, Spain, in 1988. He received the degree in mathematics and the Technical Engineering degree in computer systems from the Universidad de Salamanca, Salamanca, in 2011 and 2015, respectively, and the Ph.D. degree in mathematics from the Universitat Politècnica de València, Valencia, Spain, in 2020.

He worked as the Project Manager for a biotechnology enterprise, Immunostep S.L., from 2012 to 2016, where his interest in research includes the biomedical domain arose. From that moment, he collaborated with the Universitat Politècnica de València, and finally, he joined the Instituto Tecnológico de la Informática, in 2018. He focused on the characterization of data variability in a clinical environment and how this variability may influence the machine learning model's performance. These interests drove the publication of several scientific articles in highly cited journals. The topics of his scientific contributions mainly cover applied mathematics, computer science, and artificial intelligence.



OMAR DEL TEJO CATALÁ was born in Valencia, Spain, in 1995. He received the master's degree in computer science engineering. He is currently a Computer Science Engineer with the Polytechnic University of Valencia (UPV). Since 2018, he has been researching at the Pattern Recognition and Artificial Intelligence Group, Instituto Tecnologico de Informatica (ITI). Moreover, he is currently performing his doctoral studies therein. He is deeply keen on deep learning techniques applied to several fields, such as object detection, medical applications, reinforcement learning, and image classification.

ISMAEL SALVADOR IGUAL received the Advanced Studies Diploma degree in the field of pattern matching, in 2002. He is currently a Computer Science Engineer with the Polytechnic University of Valencia (UPV). Since 2003, he has been working at the Pattern Recognition and Artificial Intelligence Group (PRAIA), Instituto Tecnológico de Informática (ITI), where he has become a Specialist in artificial vision systems for biometrics, medical imaging, and 3D inspection. He has also led and participated in research and development projects for public institutions and for private companies and has published more than 15 articles in the field of machine learning.

RAFAEL LLOBET received the Ph.D. degree in computer science from the Universitat Politècnica de València (UPV), Spain, in 2006.

He has worked with the Instituto de Biomecánica de Valencia (IBV) and Instituto Tecnológico de Informática (ITI). Since 2000, he has been working at UPV, where he works as an Assistant Lecturer with the Department of Information Systems and Computation. He also collaborates with ITI, where he develops his research. His current research interests include machine learning and its application to healthcare area. His research is mainly focused on medical image processing, genomics data analysis, and computer-aided diagnosis. He has published works in 14 international journals and 13 international conferences.

JUAN-CARLOS PEREZ-CORTES received the Ph.D. degree in computer science from the Polytechnic University of Valencia. He is currently a Full Professor. He is also the Director of the Pattern Recognition and Image Analysis (PRAIA) Research Group, Instituto Tecnológico de Informática (ITI). He has leaded and coordinated research projects funded by public national and international entities in the field of medical imaging, industrial software, computer vision, pattern recognition, and free software. He teaches master's degree and Ph.D. courses in computer systems artificial vision and pattern recognition. He has published works in journals (15), books (three), books, and conferences (17), and has been awarded by public and private entities.