Pattern UnRecognition: I am color blind

Sachit Gaudi
Department of Computer Science
Michigan State University
gaudisac@msu.edu

Abstract

Large models can accurately model complex decision boundaries but may not generalize well to new or out-of-distribution samples. This is a challenge for autonomous cars because it's impossible to collect data from every road in the world. Researchers - Ribeiro et.al and Arjovsky et.al - have studied the problem of spurious correlations in image classification, where, for example, wolves are more likely to be spotted in snowy backgrounds than dogs. We want to know how well current classification techniques address this issue and propose new ideas to overcome spurious correlations.

1 Introduction

In this work we attempt to study, How does the distributional shift in data affect both the parametric and non-parametric approaches? We work with colored MNIST from the lens of classical algorithms. Then we present 2 techniques that are designed to remove unwanted bias, The two approaches take the inspiration from signal processing and privacy. One of them is designed if we know the bias associated with the sample, for colored MNIST its the value of color. The other approach is more general where we do not consider any bias other than classifying the sample.

This document is organised as follows, In chapter 2, We introduce the dataset and study some properties of the dataset. In Chapter 3, We study the associated clusters - PCA, MDA - with respect to the label and spuriously co-related color and propose some improvements to address the challenges. To capture the non-linearity we extend the same clustering algorithms to the ResNet18 features. In chapter 4 we define the metrics and experimental setup to study various classification techniques. In Chapter 5 and 6 we study various classical parametric and non-parametric machine learning classification algorithms and study the captured unwanted bias captured by these algorithms. In chapter 7 we analyse fully connected Deep Networks and also training ResNet18 model and propose our methods and compare them to the existing baselines.In Chapter 8 we provide a detailed insights into our results.

2 Dataset

Various studies have attempted to study the spurious correlation, one of the works that popularised the problem was the paper by Kim et.al. To synthesize the color bias, they have selected ten distinct colors and assigned them to each digit category as their mean color and then, for each training image, they randomly sampled a color from the normal distribution of the corresponding mean color and provided variance, σ, In our example we take $\sigma = 0.04$, but in testing phase the colors of digits are uniformly distributed. We are mainly interested in how the model performs on non-correlated color samples.

Code is available at sachit3022/project802

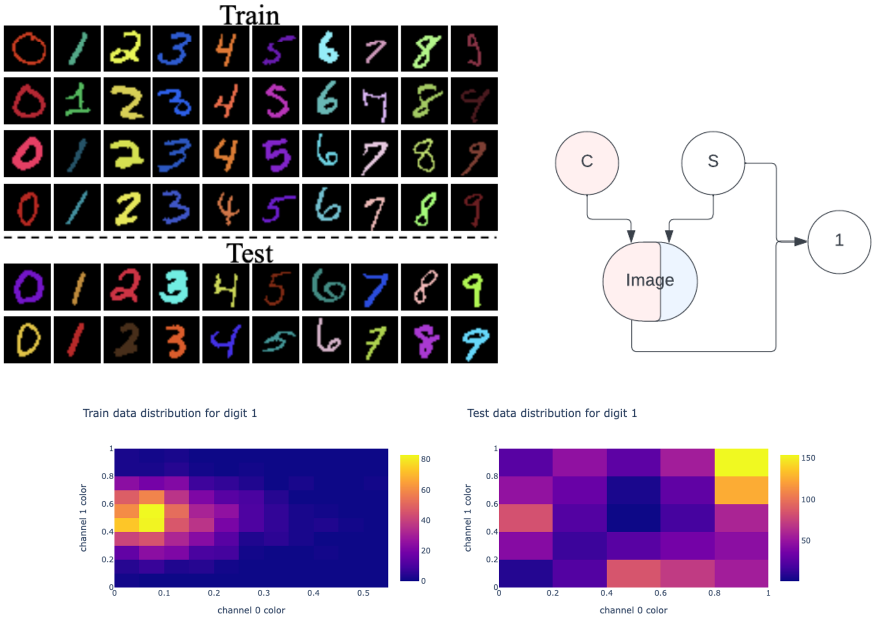


Figure 1: (a) sample snapshot of Dataset. (b) Causal Diagram (c) Train histogram of color pixels (d) Test histogram of color pixels

We assume that the color bias is easy to learn bias than the edges and curves of the digit. This data is interesting because, the color is concentrated for each digit but relying only on color of the training data will 83% train accuracy when trained a linear regression model. This will set our baseline for train accuracy. To motivate further, This behaviour is not limited to just the shallow models, even the ResNet18 (5) based models also suffers from the shift in data distribution below is the gap between the training and test accuracy. In case of same data distribution train and test accuracy reach 100% but In case of biased data test accuracy saturates at 83%.

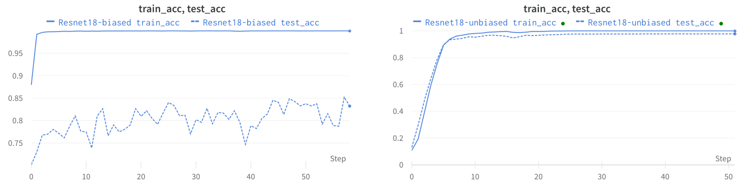


Figure 2: (a) biased data test-train accuracy (b) unbiased data test-train accuracy

2.1 Mathematical Preliminaries

$\mathbf{X}$ represents image data in 3 channels $X_{1000\times3\times28\times28}$, and y is label assoiated with the image $\mathbf{y_{1000\times1}}$ and s is the sensitive attribute associated with the image $\mathbf{s_{1000\times3}}$, As the Image has 3 channels, $\mathbf{s}$ is a color represented in the 3 channels. C is a covariance matrix

3 Dimension Reduction

In section we want to study what are the aspect that are captured by various dimensionality reduction techniques, We start with capturing linear relation - PCA, MDA and we will analyse the issues with both the approaches and suggest an improvement to MDA. To incorporate the non-linearity in the feature space, we use ResNet18 hidden layer features, this would be a common theme across the paper.

3.1 Projections

To projection onto the column space of a matrix $\mathbf{W}$, We need to find the orthogonal basis of $\mathbf{W}$, either by Gram Schmidt Orthonormalization or we can take the eigen vectors of the matrix as they from orthogonal basis. To put it more formally, A matrix $\mathbf{W}$ is considered as a projection matrix if $\mathbf{W}\mathbf{W}^{\mathbf{T}} = \mathbf{I}$. Here the matrix $\mathbf{W}$ is called projection matrix and $\mathbf{W}^{\mathbf{T}}$ is called reconstruction matrix. In this section we explore various to find $\mathbf{W}$ based on the basis space.

3.2 PCA

In PCA, we want to project onto the direction of maximum variance. To put it more formally,

$$\max_{\mathbf{w}} \frac{\mathbf{w}^{\mathbf{T}}\mathbf{C}\mathbf{w}}{\mathbf{w}^{\mathbf{T}}\mathbf{w}} \tag{1}$$

Where C is the covariance matrix. This is generally solved using Lagrangian. $\max \mathbf{w}^{\mathbf{T}}\mathbf{C}\mathbf{w} - \lambda\mathbf{w}^{\mathbf{T}}\mathbf{C}\mathbf{w}$. As C is a positive semi definite matrix, the maximum values can be achieved in the direction of eigen vectors.

3.3 MDA

The limitation to PCA is it finds the maximum variance line but not the maximum separable line. To address this, MDA technique has been proposed. MDA modifies the constraint by decomposing the C matrix into Inter-class and Intra-class covariance matrices, and maximises Inter-class covariance. These maximisation constraints have no upper bound. To make this bounded, we have to add constraint on Intra-class covariance matrix.

$$\max_{\mathbf{w}} \frac{\mathbf{w}^{\mathbf{T}}\mathbf{S_b}\mathbf{w}}{\mathbf{w}^{\mathbf{T}}\mathbf{S_w}\mathbf{w}} \tag{2}$$

This can be solved using lobpcg optimisation or can simply solve for eigen vectors of $\mathbf{S_w^{-1}S_b}$.

This optimisation is not possible if the matrix $\mathbf{S_w}$ is ill-conditioned. This approach also suffers from generalization, In colored MNIST, The maximum separability will be along the direction of color in the training data but maximum variance will be along the edge shape of the digits.

3.3.1 Improvement to MDA

To overcome the limitations of both PCA, and MDA, We propose a approach combined to overcome both the limitations mentioned above. In this approach We further add the decompose $\mathbf{S_b}$ into 2 components color component and edge component. $\mathbf{S_b} = \Sigma_{i=1}^{c}\Sigma_{j=1}^{color}\Sigma_{t=1}^{N}(x_t^{ij} - \mu_{ij})(x_t^{ij} - \mu_{ij})^T$ By solving the same way a MDA, We get $S_T = S_{be} + S_{bc} + S_w$ The objective function is modified as

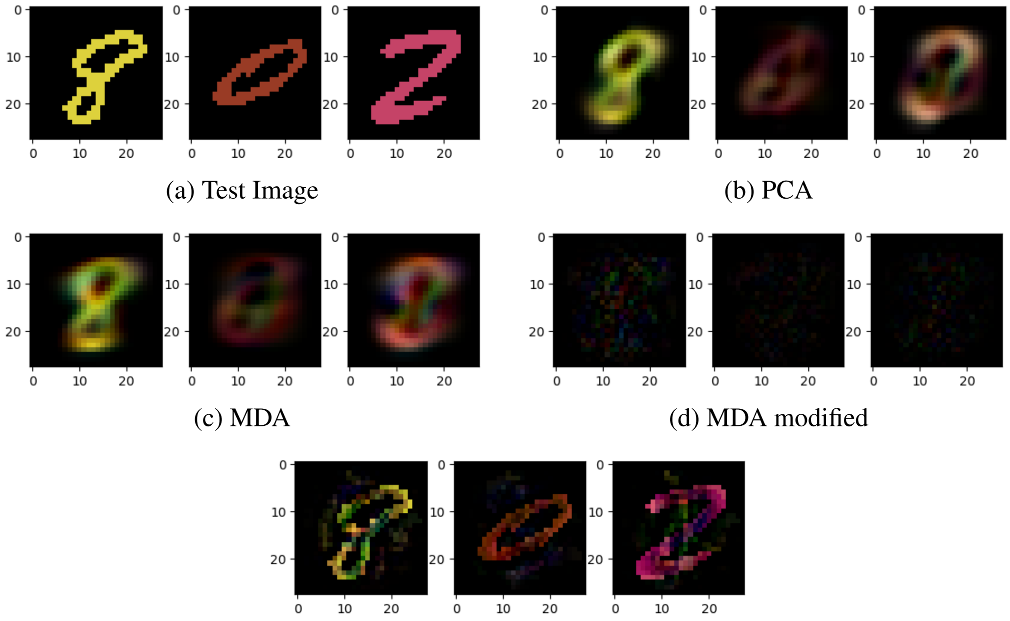
$$\max_{\mathbf{w}} \frac{\mathbf{w}^{\mathbf{T}}\mathbf{S_{be}}\mathbf{w}}{\mathbf{w}^{\mathbf{T}}(\mathbf{S_w} + \gamma\mathbf{S_{bc}})\mathbf{w}} \tag{3}$$

Where S_{bc}, S_{be} is scatter across the color space, and edge space respectively. and γ is the hyper parameter to control the trade off.

In colored MNIST, as most of the background pixels for example, (0,0,0) are always of 255. All the matrices have null space. General approach to improve the condition number of $S_w + \gamma S_{bc}$ we can add small identity matrix to make semi-positive definiteness to positive definiteness. So $S_w + \gamma S_{bc} + 1e^{-5}I$. but this approaches gets complete invariant space and the digits are mapped to black image in reconstruction. as the eigen values of the new matrix will be $\frac{1}{0+1e-5}$ which is still large value. We know that we don't want to project on invariant to both S_b and S_w. We rather take the pseudo Inverse that will create the null space.

3.3.2 General Idea of projections

The projection ideas boil down to find a good basis for the subspace. The goal is to find a matrix Which have the above properties, To find a color invariant subspace, We construct the subspace of color by averaging the pixels of a single color. $I - \mathbf{W}$ will project onto the invariant subspace of color. This approach is more robust compared to MDA.



(a) Test Image

(b) PCA

(c) MDA

(d) MDA modified

(e) projection

Figure 3: Reconstruction

3.4 Reconstruction

Does the principal components of - PCA, MDA, modified MDA - have information only about the spurious attribute or about the shape which is causally related to the digit recognition. To understand the Dimensionality reduction techniques, the representation of the latent can be projected back to the Image domain by $\mathbf{W^T}$. In all the cases, there is a strong relation with color, as the positively related samples to color are better reconstructed compared to the out of distribution images. Results of the reconstructed images are presented below.

3.5 Impact of whitening transform

The centering approach is taken in many deep learning training techniques, batch norm is proven to be effective in most of the case, in terms of better convergence of the Deep learning models. Ioffe, (6), have argued that after every update, there will be shift in the data that passes through the layer. As with our case where there is a distribution shift in our train and test, we investigate if this improves our results. We want to center data on all the features, We go beyond the centering, We also want to give every pixel equal importance in the distance metric. This argument can be debated based on the results, In some cases this argument failes, as we donot want to give background pixels equal importance as the digit / boundary pixels.

Does centering equal true improve the performance? To support this claim, we have produced the results on a dummy dataset, results are available in Supplementary material.

4 Metrics

We are interested to study the bias associated with the color. We compute the bias as the Mahalanobis distance across the channels [RGB]. We are interested in studying how the accuracy of the classifier changes as we go away from the mean of that class. We say our model is robust to color if the accuracy does not change as we move away from the distribution of color.

We compute the accuracy score, f1 score and confusion matrix of a classifier, The confusion matrix are presented in the Supplementary. We layer our improvements on top of the base classifier and discuss how each of the discussed improved on par with the data.

4.1 Experimental Procedure

MNIST data has 70,000 Images, we divide the data into 60,000 train and 10,000 test images, where 10,000 images are generated from a class conditional color distribution. We have labels y associated with each data point and also the sensitive attribute associated with the image, s, 3 channel color vector. We further use the 5-fold split as a validation strategy and find the best hyper parameters for the non-parametric methods.

5 Parametric classification

5.1 Logistic Regression

5.2 SVM and kernalised SVM

Support Vector Machines (SVM) is a powerful supervised learning algorithm used for classification. SVM works by finding the optimal hyperplane that separates two classes in a high-dimensional feature space, such that the margin between the classes is maximized. We hypothesise that the margin is like a weighting technique, where it weights the points on the decision boundary more compared to similar colored samples. This feature of hinge loss in SVM will help in increase the accuracy on the test set.

Kernel SVM is an extension of the SVM algorithm that allows for non-linear classification by mapping the input features into a higher-dimensional space using a kernel function. The kernel function computes the dot product between the transformed features, allowing SVM to separate non-linearly separable classes. (7) paper discuss the relation of memorization and modeling complex features as a function of graph and distances associated with each training sample. (10) paper also uses kernel functions in similar setting.

Kernel Normalisation centering none of the approaches workin in imporving the accuracy, only reconstruction and sensitive conditional appraches are the way to go if we dont have full distribution in training. becuase we need to carry invariant information from the non-sensitive classes. If we had information or atleast few samples then complex models will try to improve performacnce for the last few outliers which indeed improves testing accuracy as we have seen the samples below.

6 Non-parametric classification

In this section we will explore Non-parametric classification techniques, As we will see that non-parametric models are not robust to the distribution shift, this will establish our baseline. Due to the huge shift In the data for train and test, and ResNet features are more sensitive to the color as it is the easy feature to learn when pretraining on massive data. but when classifying on the test dataset direction of the features entirely change.

6.1 KNN

KNN is considered as a very effective baseline for understanding the memorisation aspect of the models, we explore the KNN with both on the raw pixels and the ResNet features.

6.2 Parzen Window

Parzen Window classifiers are first introduced in (11). As demonstrated in (4), Parzen window systems are used in current day GAN models to evaluate the generated distribution. As the Parzen windows are differentiable we can use them in training the neural network with KL-loss, but In this section we will study the memorisation aspect of Parzen windows.

7 Deep Learning Methods

7.1 ResNet18 backbone

We train end-end ResNet18 base model, The model's depth and structure allow it to learn complex features and patterns, which makes it suitable for our task as want to identify the complex features as opposed to the simple feature color. The model learns the simple feratures first so in the first few epocs the train accuracy will reach 99.98% but as the eopcs increases the models works on the outliers with respect to the color and test acccuracy improves and saturates as the train accuracy reaches 100% and the model saturates.

7.1.1 How are models performing compared to increase in σ

Model performance on test set increases as we increase the σ. we have performed 2 experiments on increasing the σ the test accuracy increases from 42% to 78%.

7.2 ARL and re-weighting technique

We want to incorporate diverse yet of same class samples in our training procedure. We can model diversity as a loss function of sensitive class,s when trained on a linear model for predicting y. This approach was first used in (8). This can be related the SVM algorithm in the sense that hinge loss is as weighting all the samples except the ones on the boundary.

The other ARL setting we can use from the performance gains achieved in the modified MDA and can explicit train the latent to contain no information about the color as an adversary.

7.3 Class Conditional permutation

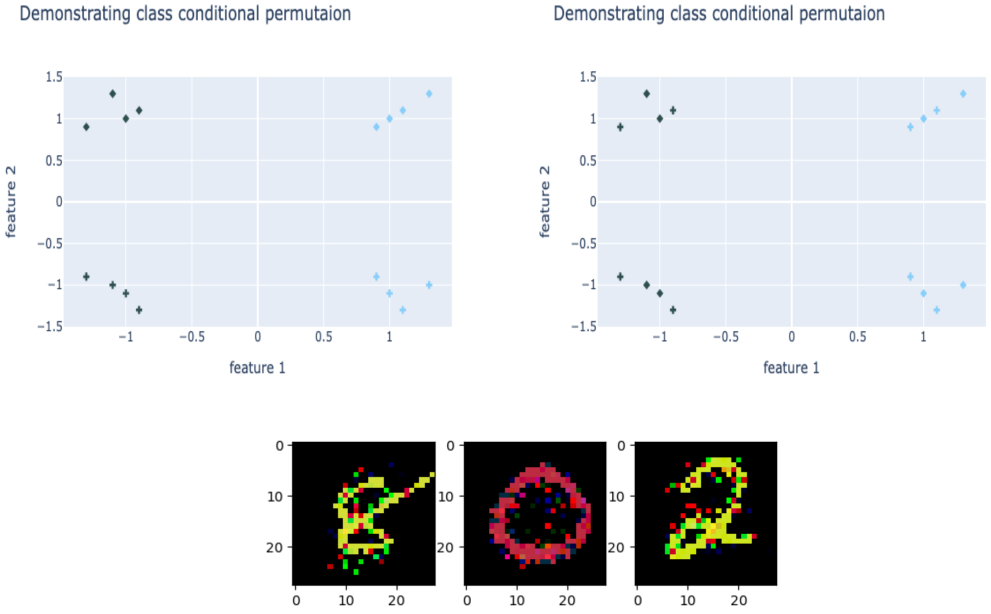


Figure 4: (a) Biased data (b) Un-biased data after permutation (c) permuation example

Suppose here colors represent the sensitive attribute and the shape indicate the different classes. we can remove the information without knowing the sensitive information by randomly permuting. This gives us flexibility of not knowing the sensitive attribute.

8 Results

We compare our approach to the different classifiers with different to data.

Table 1: Results

Model	train accuracy	test accuracy	f1 score	precision	Recall
LogisticRegression on Color Information	83.38%	10.5 ± 0.6%	6.7 ± 0.4%	10.3 ± 3%	10.5 ± 0.6%
PCA + Linear SVM	98%	42.02 ± 1.9%	42.65 ± 1.7%	44.60 ± 1.7%	42.0 ± 1.9%
PCA + RBF SVM	97.80%	39.56 ± 1.9%	40.17 ± 1.7%	46.27 ± 1.6%	39.56 ± 1.9%
MDA + Linear SVM	98.04%	38.50 ± 1.9%	38.98 ± 1.7%	40.50 ± 1.4%	38.50 ± 1.9%
MDA + RBF SVM	97.91%	38.17 ± 1.7%	38.68 ± 1.7%	40.33 ± 1.1%	38.17 ± 1.7%
Center + MDA + Linear SVM	98.05%	39.05 ± 1.9%	39.54 ± 1.9%	41.19 ± 1.6%	39.05 ± 1.9%
Whitening + MDA + Linear SVM	98.10%	41.56 ± 2.7%	42.06 ± 2.5%	44.26 ± 2.3%	41.56 ± 2.7%
*** Center + modified MDA + Linear SVM**	90.67%	**49.62 ± 2.5%**	49.46 ± 2.5%	50.71 ± 2.7%	49.62 ± 2.87%
*** Center + projection + Linear SVM**	-	-	-	-	-
Linear SVM	98.8%	42.42 ± 0.9%	43.20 ± 0.6%	45.16 ± 0.6%	42.42 ± 0.9%
RBF SVM	98.8%	42.42 ± 0.9%	43.23 ± 0.6%	45.16 ± 0.6%	42.42 ± 0.9%
Logistic Regression	98.57%	34.56 ± 1.4%	35.5 ± 1.4%	37.95 ± 1.4%	34.46 ± 1.2%
P-ResNet + Logistic Regression	89.58%	41.5 ± 0.7%	41.81 ± 0.8%	42.5 ± 0.9%	41.5 ± 0.9%
P-ResNet + Linear SVM	93.06%	42.20 ± 0.8%	42.69 ± 0.9%	43.50 ± 1.1%	42.2 ± 8.2%
Parzen Window	100%	24.33%	25.22%	27.55%	24.33%
P- ResNet + Parzen Window	65%	18.70%	12.93%	20.79%	18.70%
KNN	100%	45.34%	46.22%	63.48%	45.33%
P- ResNet + Whitening + KNN	77.13%	25.20%	20.5%	41.50%	23.00%
P-ResNet + KNN	77.13%	37.63 ± 0.6%	37.22%	38.63%	37.63%
ResNet End-2-End	99.98%	43.62%	43.62%	43.62%	43.62%
*** ResNet + ARL**	-	-	-	-	-
*** ResNet + Class cond Permutation**	99.99%	41.05	41.05%	41.05%	41.05%

***contributions**

9 Conclusion

Based on the observed training pattern, it becomes apparent that the training accuracy quickly reaches a high level of 95%. However, as the number of epochs increases, the model begins to struggle with hard samples, which are at the ends of the color distribution of the class that contribute to a larger loss. Therefore, increase in breadth of the sensitive class σ, which can be considered from a causal perspective as a randomized experiment on color.

The only idea that is robust to distribution shift is to explicitly instruct the model not to learn the color, either by projecting it into a color-invariant space or by using a regularizer with the class color scatter matrix, as suggested in the proposed Improvement to MDA. As you can see in the reconstruction of the color-invariant projection, each digit blends into the other. In both the situations we explicitly use the sensitive attribute information, In our example, color. This fashion of Adversarial training is found in (3), where the authors have used the external information to control the influence of sensitive attribute. To my knowledge, the idea of variance space constraint to LDA is novel.

Research such as DRO and other weighting techniques (1) (2) and teacher-student as ARL as proposed by the paper (9), are successful in making the model fair with respective sensitive attribute. These papers work on (12) dataset, where there are few samples (small number compared to the majority class) of sensitive classes present in the labels. However, these approaches are prone to fail when dealing with completely new data shifts, where the training data does not contain single outlier sample. Whereas our work on MDA correction will causally capture the color relation.In the case of colored MNIST, the task is challenging because there are no outliers of other classes, and the sensitive attribute is closely correlated with the target label. This case can be made firm by observing that as the value of σ increases, these techniques becomes relevant. Moving the σ value from 0.02 to 0.04 results in an almost double increase in performance. We have also seen that our own method class conditional permutation has failed as there are no sensitive attributes on tail in the training data.

As it is known that the standard non-parametric approaches are reliant on the memorization aspect, these are the models that drop the most accuracy on test set. Parzen window is has highest accuracy when the distribution of the data is same is not disturbed, and the papers have shown a connection for PNNs and the Parzen windows in the context of memorisation of the neural network models.

As there is a equal drop in the accuracy with and without dimensionality reduction techniques where we used number of dimensions as 10, we can say that all the techniques are learning to separate on color. So the top 10 dimensions capture the most information about the class variance, we also see this in the reconstructed images, human eye can make out the difference in digits. but as we have explicitly added our modification the major components now doe snot contain color.

We can say that pixels are linearly related as we do not see an improved performance when switched to a RBF kernel. We also can add that our idea class conditional permutation has not worked because of the above mentioned limitations, as in the MNIST dataset we don't have outliers with respective to color pertaining to a single class.

The performance of all the algorithms have degraded if we use the ResNet18 pretrained model weights, We learn that the resnet weights do not generalise to any tasks well. (It generalises mostly to the datasets where edges are more predominant.)

We started with Whitening as an important improvement but no much change or degradation in some cases has been observed because whitening considers the information in all the pixels as equal, but in our case the background pixels do not contributed to the distance much, but on the other hand the pixels on the edges contribute a lot and they should get higher weight compared to the other pixels. (in accordance with their variance).

In this paper, we have presented various classification techniques. The general comments about the Non parametric approaches in the sense of memorisation and this comments are true for the Deep Learning frameworks as well. We also observe that increase in the capacity of the model does not improves the performance as it always comes to the curse of diamensionality case. This paper provides insights on distribution shifts in data affects the models. We also provide the a metric, we call it test-train difference which helps in understanding the generalisation capacity of the model, which is very important in the realm of large models. We challenge the existing frameworks in existing fairness research which are focused on re-weighting and DRO.

10 Future work

We need more data on Ablation study and also confusion matrix to identify which sufferes, the idea is to use mahalanobis distance to catagorize the outliers plan to implement in the future.

Why would our ResNet based ARL works? As the improvement to MDA showed impressive results and its is a improvement to the linear space (We can say our data is linearly separable as adding kernels to the classifiers didnot improve the performance). because the only way we Can incorporate the explicit information of color in variance is by conditioning on the color or to put it explicitly training to not contain variance in color. We can also look for the improvement to these problems from the causality perspective as well.

References

[1] Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models, 2023.

[2] Junyi Chai and Xiaoqian Wang. Self-supervised fair representation learning without demographics. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[3] Jinwoo Choi, Chen Gao, Joseph C. E. Messou, and Jia-Bin Huang. Why can't i dance in the mall? learning to mitigate scene bias in action recognition. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[4] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[6] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*, abs/1502.03167, 2015.

[7] R. Jenssen, D. Erdogmus, J.C. Principe, and T. Eltoft. Towards a unification of information theoretic learning and kernel methods. In *Proceedings of the 2004 14th IEEE Signal Processing Society Workshop Machine Learning for Signal Processing, 2004.*, pages 93–102, 2004.

[8] Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed H. Chi. Fairness without demographics through adversarially reweighted learning. *CoRR*, abs/2006.13114, 2020.

[9] Jun Hyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: Training debiased classifier from biased classifier. *CoRR*, abs/2007.02561, 2020.

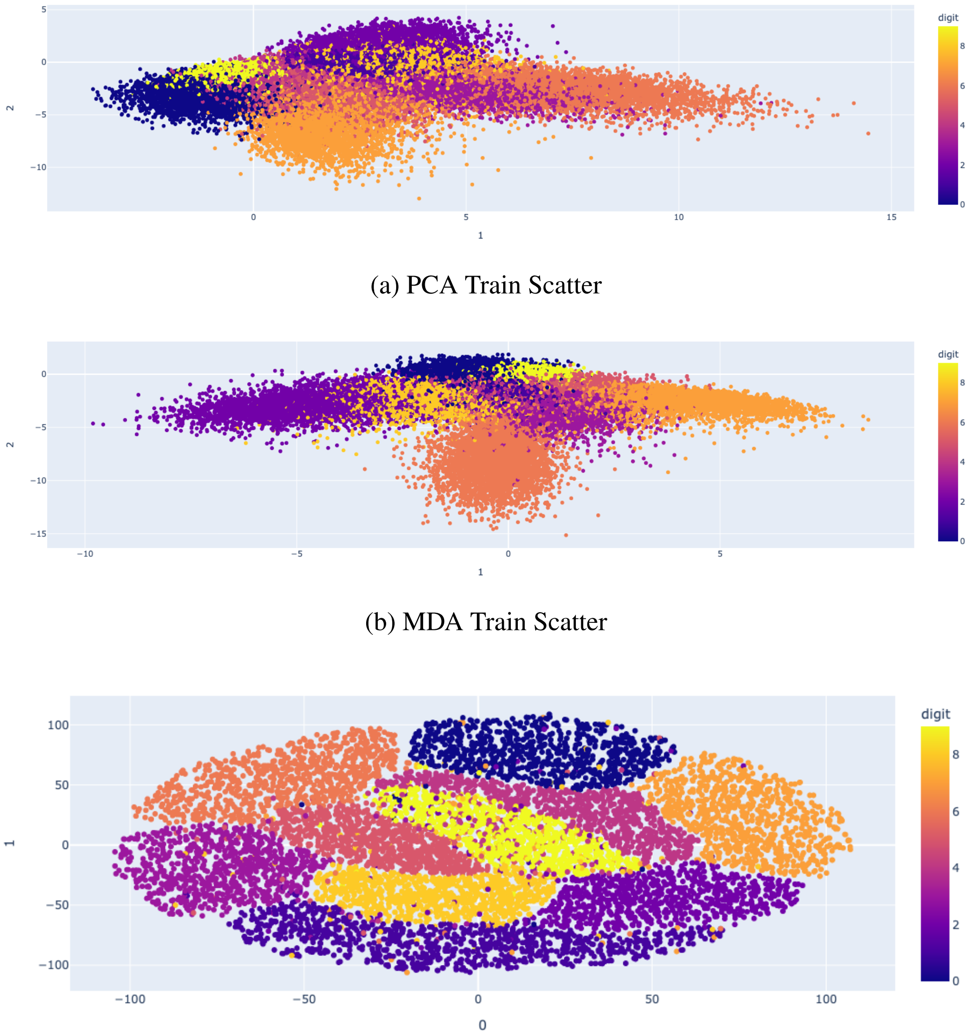
[10] Bashir Sadeghi and Vishnu Boddeti. On the fundamental trade-offs in learning invariant representations. *CoRR*, abs/2109.03386, 2021.

[11] Donald F Specht. Probabilistic neural networks. *Neural networks*, 3(1):109–118, 1990.

[12] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge J. Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

Supplementary Material

Visualisation of training data



(a) PCA Train Scatter

(b) MDA Train Scatter

(b) tSNE Train Scatter

Figure 5: Scatter 2D data after projection

Effectiveness of Whitening transform

To study the impact of whitening transform, we created a linearly separable data, with 3 independent features of different variance and one feature which is dependent on the other three. On this dataset, we observed that whitening resulted in 3.3% increase in performance (96.7% to 100%) on best performing hyper parameters. and also more robust to hyper parameter to h.

Hyper parameter for KNN window

We have split the data into train and validation sets, where the validation set is used for setting the hyper parameter. We vary from window width (k) from 1-500 by incrementing 5.

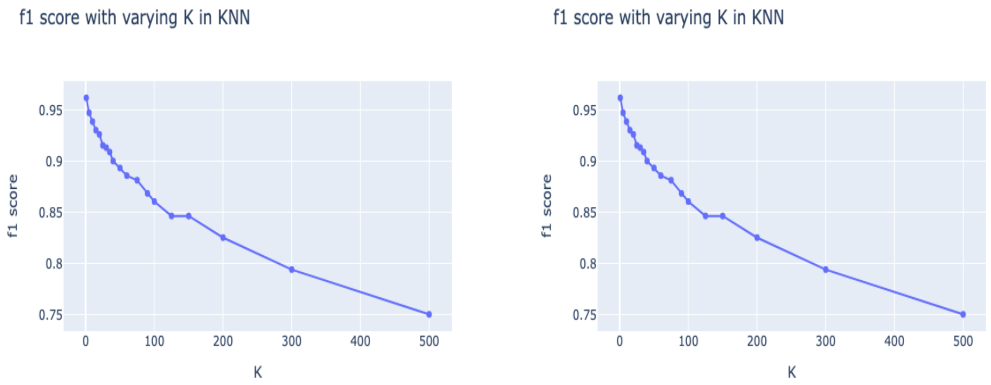


Figure 6: (a) KNN raw pixel K (b) P-ResNet18 + KNN

Model	K-value	validation accuracy	f1 score
KNN	1	95.26%	94.30%
KNN + Whittening	6	51.73%	50.63%
P-ResNet + KNN + Whittening	6	51.73%	50.63%

Hyper parameter for Parzen window

We have split the data into train and validation sets, where the validation set is used for setting the hyper parameter. We vary from window width (h) from 0-10 in a logarithmic fashion.

Model	K-value	validation accuracy	f1 score
Parzen	3.75	100%	100%
P-ResNet +Parzen	3.76	65%	63%

Confusion matrix

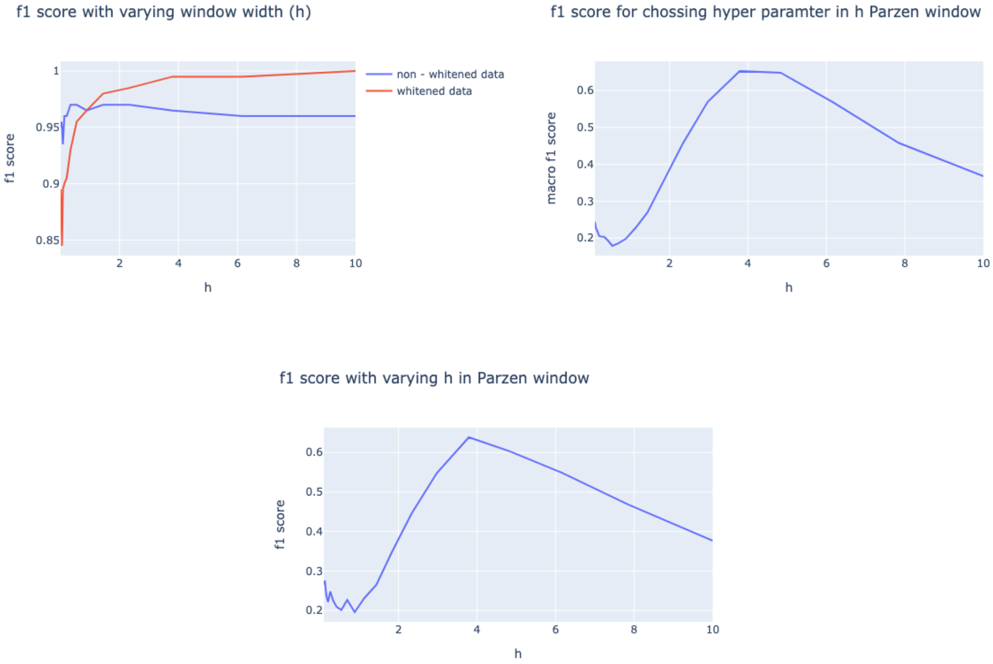


Figure 7: (a) Effectiveness of whitening (b) Window width hyper parameter search (c) Window width hyper parameter search