Evaluating the Potential and Realized Impact of Data Augmentations

David Heise
Science, Technology & Mathematics
Lincoln University
Jefferson City, Missouri, USA
heised@lincolnu.edu

Helen L. Bear
YLB Tech, Ltd
London, United Kingdom
ylb.tech.ltd@gmail.com

***Abstract*—Data augmentations have been shown to improve predictive performance of machine learning models in many domains. Augmentations are typically used to improve classification performance, but augmentations can distort the intrinsic properties of the original data, thus reducing the utility of a model for real-world applications. Because augmentations directly affect the training data, and thus also affect the machine learning models trained with said data, intelligent selection of augmentations is as critical as the selection of input features and other options in the machine learning pipeline. Such an approach will enable greater transferability of trained models from the research lab to products and services.**

This paper presents two metrics to evaluate the potential and realized impact of data augmentations. The first metric, eff-score, assesses the relative efficacy of prospective data augmentations before model training. To observe augmentation effects on the intrinsic properties of the training data, the second metric, nirvana distance, measures the effect of data augmentations beyond overall predictive performance after model training. These metrics are tested with a well known multi-purpose audio data set and augmentations from the domain of environmental sound scene analysis. The relative eff-scores correlate with classification results from predictive models trained on the augmented data sets, and the distance components of the nirvana distance explain results observed but not previously understood from output confusion matrices. These results demonstrate promise for data-driven, efficient selection of data augmentations whilst exposing previously hidden impacts on machine learning models. Furthermore, since eff-score and nirvana distance are domain-independent, these metrics have widespread applicability.

***Index Terms*—data augmentation, augmentation evaluation, eff-score, nirvana distance, visualization, audio**

I. Introduction

Data augmentations are frequently used to increase predictive performance of trained machine learning (ML) models [1]–[3]. Training on augmented data is often "low-hanging fruit" in the quest to develop a model that out-performs other models [4]. This singular focus, though, potentially limits the utility of developed models beyond the research lab. A model that performs well on a single test set, or indeed with some cross-validation, is not guaranteed to perform similarly on samples from a larger, more diverse real-world environment [5] – especially if the training data does not reflect all inherent qualities of real-world data [6]. Thus, in reality, researchers have a hierarchy of three objectives:

1) to build predictive models which accurately recognize all classes equitably;
2) to ensure models continue to predict fairly in an open environment;
3) to ensure that any data processing at inference is minimal to reduce latency when models are served.

During training, a predictive model learns a function with which it transforms data observations (either pre- or post-augmentation) into something toward a perfect classifier of target classes. Fig. 1 represents this transformation as a four-stage simple pipeline. The intrinsic properties of the real-world data should be preserved when (the optional) stage two, augmentation, is performed, lest the data be distorted such that the model loses equitable treatment of all target classes.

Augmentations are primarily evaluated after stage three in Fig. 1 (e.g., [7]), and such an approach to determining appropriate augmentations is costly and time-consuming [8]. If augmentations could be evaluated *prior* to stage three, much development could be optimised by cost and time reduction [9]. Thus, pre-determining the best augmentation strategy is preferable to simply "using them all" [8], and simply adding more data is less beneficial than adding relevant data to a training corpus. The first contribution of this paper is a novel approach to understand and predict the efficacy of data augmentations (Section IV-A). The second contribution of this paper addresses equal recognition of *all* classes, with a metric and a visualization process for evaluating the degree to which augmentations alter per-class performance (Section IV-B). Since the validity of machine learning predictions depends upon the degree to which augmented data retains the properties of real data [6], augmentations should not skew the data in a way that gives "preferential treatment" to one or more classes at the expense of others. Preserving fair performance of a

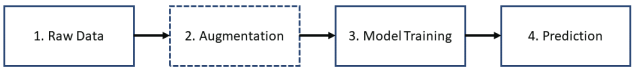

Fig. 1. Machine learning classifier pipeline.

This work is supported by the National Science Foundation, award 2100874.

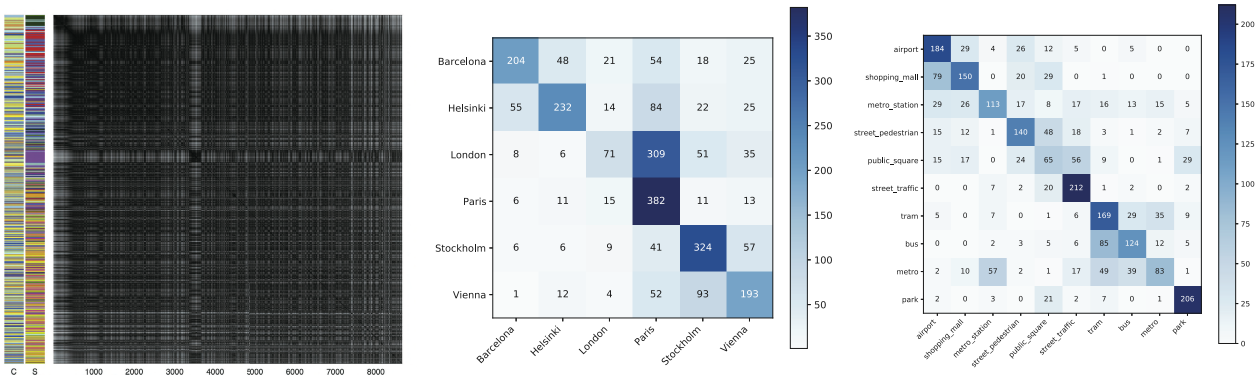

Fig. 2. (a) VAT image of original data from [10]; (b) & (c) confusion matrices from [11] for city and scene classification, respectively.

model on all classes is just as important as obtaining high overall classification accuracy when seeking models that are useful to society.

Section II provides background on sound scene analysis (the application test domain), summarizing pertinent literature before describing the data set used in this work. Section III describes the preparation of data and overall approach to experimentation (in the context of the machine learning development pipeline). Section IV details the novel metrics, and Section V analyses the results of applying these metrics to a case study. The paper concludes with a summary of contributions and observations. A link to code for metrics and visualizations used is provided to encourage and assist others interested in investigating and evaluating data augmentations.

II. Background

This work was initially inspired by observations from literature in the domain of sound scene analysis. In particular, the authors observed from papers and conferences (e.g., [12]) that often researchers only state that standard augmentations were applied before training, or simply name the augmentations used, without justification or rationale for augmentation choices made. Such papers focus on detailing model design parameters, but the model parameters may matter less than the data processing applied [8] (in the same way that feature selection is critical).

A. Case study domain: Sound scene analysis

Sound scenes are recordings of any environment, such as an office or a street. Understanding these scenes is the task of environmental sound analysis, which contains many possible challenges: audio description (captioning) [13], scene classification [14], event detection and/or classification [15], bioacoustic recognition [16], geotagging [11], and synthetic generation [17], to name a few. For practical use, real-world data is desired to ensure a training corpus is representative of real-world conditions. This data is often multi-faceted and frequently contains noise (where noise is any component not pertinent to a model's primary task or purpose). The multi-facetedness of real-world data can be exploited by multi-task models, which infer higher-order features that contribute to improve predictions on each task [18]. In the sound scene analysis domain (and others), data collections of recordings can be used for more than one task [10]. Even so, multi-purpose data sets are not always of the size needed to train predictive models; consequently, many researchers and developers take advantage of data augmentation.

City classification is also known as audio geotagging [19]. The seminal audio geotagging study [11] provides a baseline for city classification from audio; confusion matrices of their best model are shown in Figs. 2(b)-(c). That work demonstrated minor variation in the predictions over all target classes without data augmentation and a robust predictive accuracy with a multi-task CNN for both scene and city classification.

An investigation [10] into understanding the data space of this multi-purpose audio data (a requirement for multi-task models) used an unsupervised data-driven visualization technique, the Visual Assessment of cluster Tendency (VAT) [20]. [10] revealed some structure relating to different target classifications. With multi-purpose audio data, VAT clusters corresponding to scene labels appeared more obvious than for city labels. The authors speculated the emergent clusters explained common misclassifications seen in prior predictive work. Fig. 2 shows a VAT image from [10] with ground truth labels for separate scene/city classification tasks shown as colored stacked bars on the left, along with the confusion matrices from models trained on the same data (from [11]).

City classification rates reported in [11] were improved in [7] through the use of data augmentations. Significant improvement in accuracy was achieved, though limited evaluation suggested that augmentation performance was not uniform over all possible target classes. When broken down by target classes, accuracy varied from 47% in a park to 91% on a bus. Thus, the impact of an augmentation depends on the particular task, model, and class at hand.

B. Augmentation methods for audio classification

[22] and [23] provide excellent reviews of many augmentations used in the audio domain. The augmentations described here were chosen for this study due to the minimal number of required parameters, their preservation of original target labels (versus, for instance, *mixup*, which creates augmented

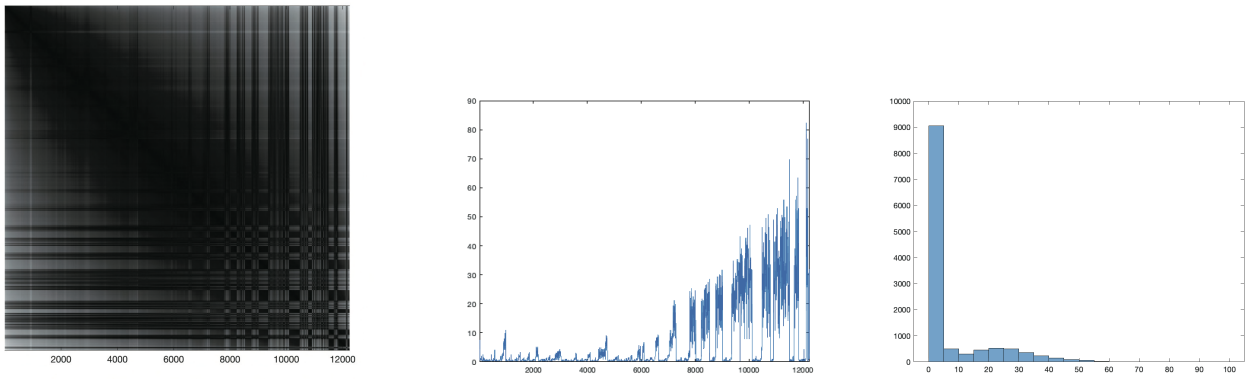

Fig. 3. Visualization of selected stages to compute *eff-score* for the *cyclic* augmentation on data from [21]: (a) VAT dissimilarity image; (b) diversity values ($\delta_{1\ldots n}$) of neighborhoods along the diagonal of (a); and (c) histogram of $\delta_{1\ldots n}$. x-axes for (a) and (b) represent sample numbers after VAT reordering.

samples with soft target labels [24]), and their use in prior related work. A *time stretch* augmentation changes the tempo and length of the original sample without changing its pitch. Experimentation to discover a data appropriate stretch parameter γ is needed, or a random value from a uniform distribution works. *Frequency stretch* is analogous to time stretch in the frequency dimension; here a random value between 10 and 100 is appropriate for γ. *Stretch* can be a third option where both time and frequency is stretched on the same sample to generate one new training sample. If a waveform is considered a loop, a *cyclic* augmentation alters the start and end point of a sample such that, to maintain the sample length, the original start of the sample is appended to the end of the new sample. No values are needed to be experimentally found as new starting points can be random. Finally, *drop* is a method where random frames are cut from the original sample.

C. Data

This work uses the same multi-purpose public evaluation data as [11] which has predefined training, validation, and test subsets [21]. Data were recorded from ten different scenes (*airport*, *bus*, *metro*, *metro_station*, *park*, *public_square*, *shopping_mall*, *street_pedestrian*, *street_traffic*, and *tram*) in six different cities (*Barcelona*, *Helsinki*, *London*, *Paris*, *Stockholm*, and *Vienna*). Each acoustic scene has 864 ten-second segments (giving 8640 segments across ten scenes). These were recorded using a binaural Soundman OKM II Klassik/studio A3 electret in-ear microphone and a Zoom F8 audio recorder using 24-bit resolution.

III. Experimental Design and Data Preparation

This experiment addresses evaluation of data augmentations with respect to two stages in the end-to-end machine learning pipeline (Fig. 1). First, to estimate how well a data augmentation may improve classification performance, an unsupervised data-driven metric is presented for use pre-training, at stage two (augmentation). Second, a metric to evaluate how an augmentation impacts the per-class prediction performance of a model is presented for use post-training, in stage four (prediction). These metrics are applied to a case study that builds upon the work of [7].

To prepare the data for evaluating both new metrics, the librosa [25] python library is used to extract log mel spectrogram features with parameters: 128 mel bands, 2048-point STFT, input sampling rate = 22050 Hz, and hop length = 512. Feature-wise normalization is completed separately after applying each augmentation. Log mel spectrograms are commonly used for sound scene analysis.

Three data augmentations – *cyclic*, *stretch*, and *drop* – are separately applied to produce three augmented data sets. For *cyclic*, each sample is shifted in time by 25%, 50% and 75% of its length; for *drop*, the dropped frames are random; and for *stretch*, a random number of columns and rows at a random position are resized by bi-linear interpolation four times on each audio sample each with different random number of columns and rows to stretch. Augmentations are applied to each channel separately and then averaged into a single-channel feature matrix. These augmentation choices were based upon their common use in the audio domain [22] and in relevant prior work [7].

IV. Metrics for Augmentation Evaluation

In Section IV-A, an efficacy score, *eff-score*, is defined to assess the potential efficacy of an augmentation on a given data set, designed to be used pre-model training. In Section IV-B, the *nirvana distance*, ND, is defined as a means of assessing how far from ideal a particular augmentation performs on a given data set, on a particular model (post-training).

A. Efficacy score

Selected stages to assess augmentation efficacy are shown in Fig. 3. Broadly, this process includes: augmenting the original data, reordering and visualizing the (augmented) data using VAT [Fig. 3(a)], computing the diversity of pixel neighborhoods along the diagonal [Fig. 3(b)], charting a histogram of the computed diversities [Fig. 3(c)], and computing a score to quantify the effectiveness of the augmentation. Before

detailing *eff-score*, VAT is described as it forms the foundation of our method.

1) Visual assessment of cluster tendency: VAT [20] is an unsupervised method to visualize the degree to which data points may cluster. A key feature of VAT is that the target number of clusters (denoted as k in other methods) need not be specified; VAT can be used to predict the number of data clusters. Variations of VAT exist for different data contexts [26], though VAT has been rarely applied to audio data (and within audio, mostly on speech utterances) [10]. Visualization of data at different stages of processing can assist with understanding changes to the data at each stage [27].

VAT calculates the distance between all pairs of data points prior to reordering them to produce a dissimilarity matrix [rendered as a grayscale image, e.g., Fig. 3(a)]. The two furthest data points are used as the start and end points, and the resulting path between them after reordering is equivalent to a minimal spanning tree of the complete graph per Prim's algorithm [28]. In typical use, dark blocks visible on the diagonal *suggest* the number of clusters in the data. Here, each audio sample was transformed into a single vector by taking the feature-wise mean over all time frames before VAT is applied. Euclidean distance was used to measure the distance between feature vectors, but other measures are possible since VAT is distance-measure independent. Patterns observed along the diagonal of VAT dissimilarity images produced from augmented data motivated *eff-score*.

2) eff-score: The *eff-score*, or efficacy score, is intended as a *relative* value for comparative analysis of likely augmentation benefit. As a relative score, augmented data sets are *from the same source data*. Once created, the following is evaluated for each augmentation ξ in the set of augmentations Ξ:

1) Generate a VAT dissimilarity matrix D for the data set (i.e., the augmented data) of dimensions $n \times n$, where n is the number of samples. D is rendered as an image, so the matrix elements are referred to as pixels.
2) Extract the $\kappa \times \kappa$ neighborhood N of pixels (κ must be $>= 3$ and odd) surrounding each of the n pixels on the diagonal of D. For instance, if $\kappa = 9$, each resulting neighborhood N is given by:

$$N_i = \begin{bmatrix} p_{(i-4,i-4)} & p_{(i-4,i-3)} & \cdots & p_{(i-4,i+4)} \\ p_{(i-3,i-4)} & p_{(i-3,i-3)} & \cdots & p_{(i-3,i+4)} \\ \cdots & \cdots & \cdots & \cdots \\ p_{(i+4,i-4)} & p_{(i+4,i-3)} & \cdots & p_{(i+4,i+4)} \end{bmatrix} \quad (1)$$

   where p is a pixel from D and i is the index along the diagonal, beginning with the upper left. [Note that for $i < (\kappa/2)$ and $i > (n-(\kappa/2))$, N_i will have dimensions less than $\kappa \times \kappa$ due to the limits of D.]
3) Compute the diversity δ for each neighborhood N, with $\delta_i = max(N_i) - min(N_i)$. *max*() and *min*() return the maximum and minimum values, respectively, within the specified neighborhood of pixels.
4) Group the diversity values $\delta_{1\ldots n}$ into b class intervals (or bins) for creating a histogram to visualize the distribution. Here, $b = 20$ was empirically found to effectively span the range of $\delta_{1\ldots n}$ computed for all neighborhoods $N_{1\ldots n}$. To facilitate comparison, class intervals should be consistent for every augmentation ξ and span the maximum range of $\delta_{1\ldots n}$ for any ξ in Ξ.
5) Count the number of elements η grouped within each class interval ι.
6) Compute the efficacy score, *eff-score*, for a given augmentation ξ by:

$$\textit{eff-score} = \eta_1 / \sum_{\iota=1}^{b} \eta_\iota. \quad (2)$$

B. Nirvana distance

The *nirvana distance* (ND) evaluates the performance of different augmentations beyond overall classification accuracy *after* model training. Data in feature space is not always uniformly distributed, and some confusions between classes can be explained through nearness of feature vectors [10]. Therefore, this metric combines the classification performance of a given augmentation (by class) with information known from the initial data.

Let F represent a matrix of distances in feature space between classes, where $F_{i,j}$ represents the distance between class i and class j. F should be inferred from the original (non-augmented) data and may be computed prior to training of classification models. For a model trained on ξ-augmented data, let $\mathrm{FN}_\xi(g,p)$ represent the number of samples with ground-truth label g that are misclassified as class p. Let C represent a cardinality vector, where C_λ gives the number of samples from class (label) λ for the data set under evaluation.

A distance component dc is defined for a given augmentation ξ and class λ as follows:

$$\mathrm{dc}_{\xi,\lambda} = \sum_{\iota \in \Lambda} \frac{F_{\lambda,\iota} * \mathrm{FN}_\xi(\lambda,\iota)}{C_\lambda} \quad (3)$$

where Λ is the set of all classes within the data set under evaluation. The *nirvana distance* ND for augmentation ξ is thus computed:

$$\mathrm{ND}_\xi = \sum_{\lambda \in \Lambda} \mathrm{dc}_{\xi,\lambda}. \quad (4)$$

The term *nirvana distance* is coined because ND_ξ may be thought of as the distance of ξ from the "ideal" augmentation. That is, finding an augmentation that yields (near) perfect classification performance such that $\mathrm{ND} \approx 0$, whilst distributing any misclassifications in a way that reflects the original feature space, would be akin to reaching "nirvana".

The method here is motivated by the Earth Mover's Distance (EMD) [29]. EMD is posed as a problem of moving a mass of earth from some number of sources to fill some number of holes. In this construction, $\mathcal{I}$ represents a set of suppliers, $\mathcal{J}$ a set of consumers, and $c_{i,j}$ the cost to ship a unit of earth from $i \in \mathcal{I}$ to $j \in \mathcal{J}$. A flow $f_{i,j}$ represents the number of units of earth shipped from supplier i to consumer j. A key constraint in EMD is that the amount of earth moved from suppliers must be the same as the amount given to consumers.

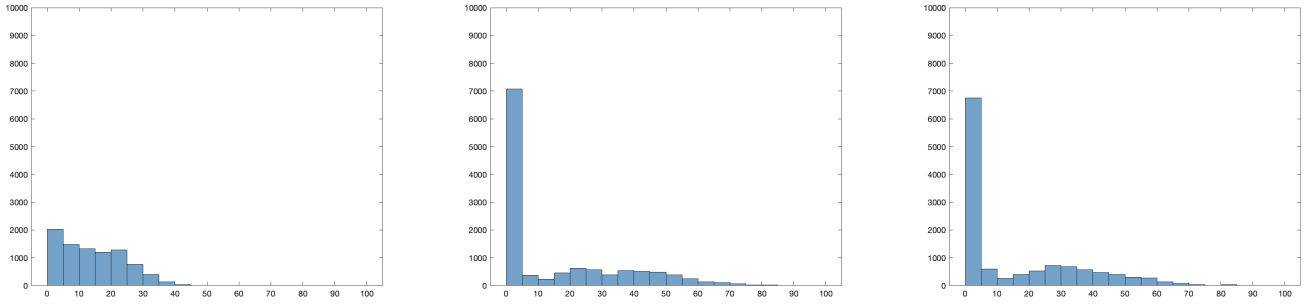

Fig. 4. Histograms of diversity values ($\delta_{1\ldots n}$) for: (a) no augmentations; (b) *stretch* augmentation; and (c) *drop* augmentation.

Many elements of EMD translate to ND. Here, the analog of cost c is distance F, and flow f is represented by the number of misclassified samples FN to be "moved" to correct classification. A key difference between ND and EMD is that, unlike with EMD, with ND there is only one valid "destination" for each misclassified sample, and each misclassified sample *must* go to that destination. Thus, the "flows" $f_{i,j}$ are already defined by the classification failures and need not be computed via an optimization problem. An additional difference is the normalization factor C_λ. With EMD, the normalization factor is the total flow from all sources $i \in \mathcal{I}$ to all destinations $j \in \mathcal{J}$; for ND, the focus is on characterizing the performance of an augmentation ξ, so each $\mathrm{dc}_{\xi,\lambda}$ is normalized by the total number of samples within each λ; a greater number of correct classifications by a model trained with ξ-augmented data will yield a smaller numerator with same denominator (compared to other augmentations) for dc.

V. Case Study Results

A. Applying *eff-score*

Table I shows scene and city classification accuracy related to the computed *eff-score* for each data augmentation, calculated from a complete set of models. The single-task scene models were trained for this paper consistent with the method from [7], and other scores are duplicated from [7] with the authors' consent. Evaluation of *eff-score* included empirical testing of different neighborhood sizes, and whilst results were not significantly different, $\kappa = 3$ (i.e., 3×3, the minimum-sized neighborhood) and $\kappa = 9$ were selected to show the relative effects. For single-task models trained for city classification, the *eff-score* has strong positive correlation with model performance ($r = 0.82$ and $r = 0.99$ for $\kappa = 3$ and $\kappa = 9$, respectively). The multi-task city models have lower but still positive ($r = 0.69$ and 0.59, respectively) correlation due to *drop* negatively impacting scene classification, which in turn impacted city accuracy. This illustrates the need to find the appropriate augmentations for the task, or tasks (in the case of multi-task predictors).

TABLE I
Results showing the *eff-score* for each data augmentation compared with classification accuracy (%) from CNN models

	single-task		multi-task		*eff-score*	
Augment	city	scene	city	scene	$\kappa = 3$	$\kappa = 9$
None	56	59	56	57	0.540	0.235
Drop	69	26	47	19	0.619	0.551
Stretch	71	72	75	63	0.633	0.578
Cyclic	75	77	79	70	0.829	0.739

Figs. 4(a)-(c) show histograms for no augmentations, *stretch*, and *drop*, respectively [*cyclic* is already presented in Fig. 3(c)]. The ratio of the counts in the first bin to the total number of samples is a useful metric because with augmented data, more uniform neighborhoods are sought along a VAT diagonal; [30] discusses the value of variance reduction with respect to data augmentations. A higher *eff-score* indicates greater uniformity.

As noted above, *drop* negatively impacts scene classification, suggesting that the results include an "exception" with respect to likely performance of evaluated augmentations. This is to be expected. Data augmentations do not perform equally well on all models, as seen by the confusion matrices in Fig. 5. Further, data augmentations are not the sole factor responsible for a model's performance; its design, size, parameters, training strategy, etc. all have a significant impact [7]. Nevertheless, *eff-score* can predict the *potential* efficacy of an augmentation, as supported by the data in Table I. As will be explored in Section V-B, augmentations do not always equally improve predictions for all classes within a target schema, demonstrating that: a) the "best" augmentation is not absolute and depends on the objective task, and b) an increase in prediction accuracy can mask an augmented-induced bias that has skewed the intrinsic properties of the original data.

B. Measuring nirvana distance

A set of trained CNN models from [7], [11] were used to generate predicted labels for augmented and non-augmented data. The results are presented in Fig. 5 as a series of confusion matrices in order to: a) compare predictions of multi-task models to single-task models, and b) discover and compare changes in predictive behavior from different augmentations. Confusion matrices serve as a primary tool for evaluating model performance, but they are limited in conveying nuance. To address this, data are plotted as bar charts (as illustrated by Fig. 6 for single-task model, no augmentation).

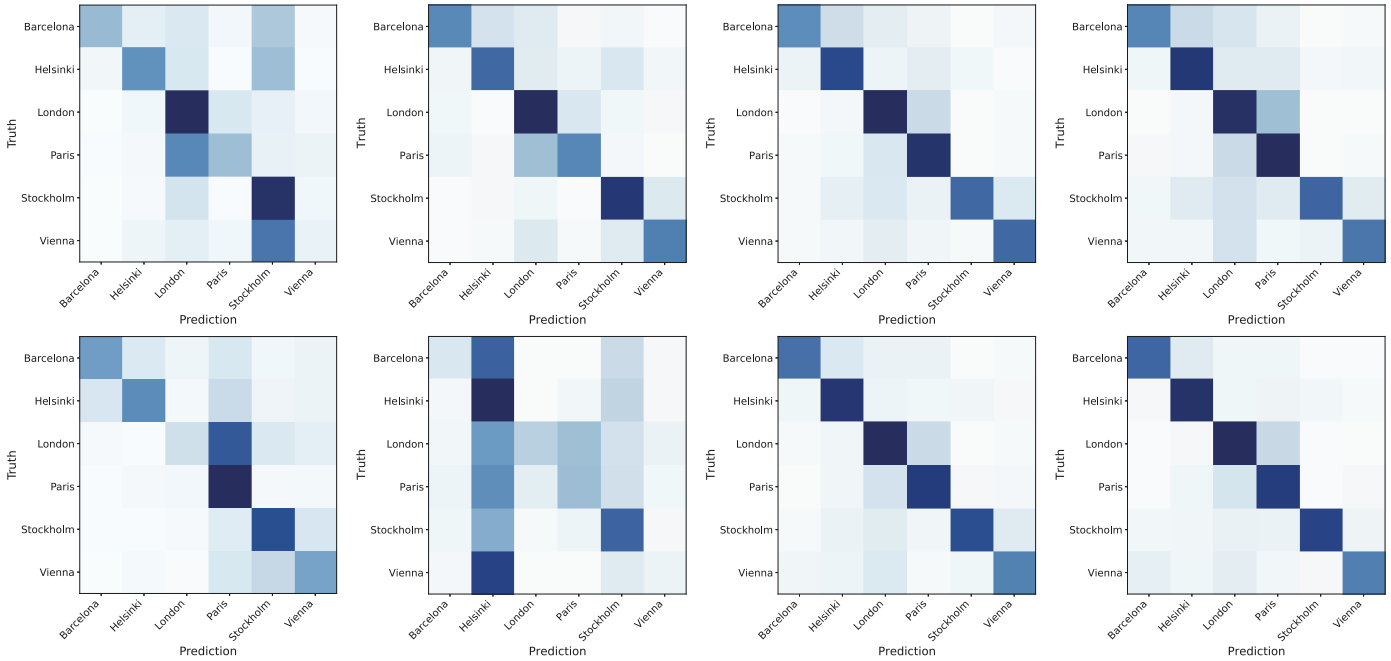

Fig. 5. Confusion matrices for trained models, city schema. Row 1: single-task models. Row 2: multi-task models. Column 1: no augmentation. Column 2: *drop* augmentation. Column 3: *stretch* augmentation. Column 4: *cyclic* augmentation.

Table II contains class-wise city classification accuracy for each model, per augmentation, separated between single-task and multi-task models. Although it is desirable all target classes are equally affected by an augmentation, this is not the case, as evidenced by the wide-ranging scores between classes in each model type and augmentation pair.

The values of ND are presented in Table III for each model by augmentation used, with class-wise distance components representing how much each class contributed to the overall distance. To assess the degree to which the properties are maintained and classes are treated fairly, the variance of the distance components that comprise each ND are given in Table IV. Here, a smaller value indicates more uniformity across target classes. Without augmentations, a single-task classifier has significantly greater between-class variation. This is noticeably reduced with each augmentation, indicating that the augmentations have improved the ease with which the classes can be discriminated. This observation further suggests that these models depend upon the augmentations to transform the original data in order for the models to score accurately. With the exception of the model trained with *drop*-augmented data, multi-task models outperformed single-task models for city classification, and Table IV multi-task values show that the

TABLE II
CLASSIFICATION ACCURACY (%), BY CITY, FOR MODELS TRAINED ON DATA USING SPECIFIED AUGMENTATION

City\Aug→	none	*drop*	*stretch*	*cyclic*
Single-task models				
Barcelona	40.0	65.1	62.2	58.9
Helsinki	50.0	66.9	77.3	73.6
London	76.5	81.0	79.6	69.8
Paris	32.4	56.2	83.8	77.9
Stockholm	80.8	82.2	64.3	58.5
Vienna	7.6	72.1	80.3	66.8
Multi-task models				
Barcelona	55.1	11.9	72.2	78.1
Helsinki	53.7	74.3	81.5	85.4
London	14.8	19.6	78.1	79.6
Paris	87.2	28.8	78.5	79.7
Stockholm	73.1	55.3	72.0	76.7
Vienna	54.4	5.9	69.0	71.3

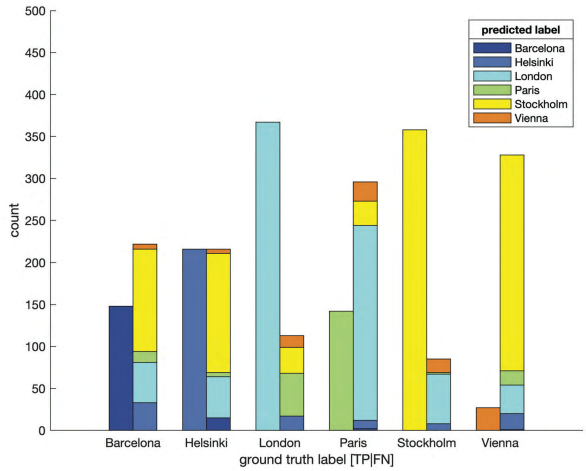

Fig. 6. Predictions by target class for single-task, no augmentation model. For each ground truth group, left bar indicates true positives (TP), right stack indicates false negatives (FN).

TABLE III
DISTANCE COMPONENT COMPUTED FOR EACH CITY AND AUGMENTATION WITH OVERALL NIRVANA DISTANCE (ND) FOR EACH MODEL

City\Aug→	none	*drop*	*stretch*	*cyclic*
Single-task models				
Barcelona	1.594	1.109	1.292	1.375
Helsinki	1.737	0.932	0.555	0.567
London	0.471	0.343	0.182	0.233
Paris	0.855	0.576	0.179	0.250
Stockholm	0.917	0.960	1.757	1.932
Vienna	4.868	1.094	0.511	1.072
ND	10.443	5.015	4.476	5.430
Multi-task models				
Barcelona	1.561	2.640	0.949	0.731
Helsinki	1.158	0.995	0.470	0.345
London	1.137	1.344	0.202	0.172
Paris	0.310	1.739	0.292	0.248
Stockholm	1.407	1.897	1.389	1.070
Vienna	2.055	2.447	1.014	0.932
ND	7.627	11.062	4.316	3.497

TABLE IV
VARIANCE IN COMPONENTS OF NIRVANA DISTANCE (ND) FOR EACH MODEL/AUGMENTATION

Model type	none	*drop*	*stretch*	*cyclic*
Single-task	2.58	0.10	0.41	0.46
Multi-task	0.33	0.40	0.22	0.14

tested augmentations have much less effect on the uniformity of per-class performance. In these instances, class uniformity remaining consistent *and* predictive performance increasing represents the most desirable outcome for models designed for real-world use.

Fig. 7 visualizes the individual distance components of ND for each augmentation. The left plot reflects that the single-task model using no augmentation classifies *Vienna* poorly (as confirmed by Fig. 6). The right plot shows that *cyclic* is best for multi-task models (confirmed by Table III) given that it has the smallest and most uniform dc values. This illustrates that, for this model, *cyclic* appears to retain the intrinsic properties of the raw data that contribute to inter-class separation, doing so from an original data space where the class centroids are not equidistant from each other. For reference, Fig. 8 maps the original data space (i.e., F to compute ND) using Euclidean distance between class centroids (from [11]) as the edge weights of the graph. In this figure, London and Paris are the acoustically closest cities in the non-augmented data, explaining their confusions in the corresponding non-augmented confusion matrices (Fig. 5, column 1) and bar chart (Fig. 6).

VI. CONCLUSIONS AND FUTURE WORK

This work is a foundation for evaluating the potential and realized impact of augmentations on the performance of machine learning models. This paper presents metrics and visualizations to efficiently and intelligently select and evaluate augmentations, both pre- and post-model training. This represents an innovation for machine learning model development, with the potential to save development time and cost while simultaneously increasing the utility of developed models for real-world use. The effectiveness of an augmentation is data, task, model, and class dependent, so the validation presented here should be extended to try different machine learning model architectures, test in other domains, and expand the set of augmentations under analysis. Nevertheless, this work represents an important step forward in the evaluation of augmentations, which — despite their influence on model performance (on par with the impact of feature selection or model parameters) — has heretofore been understudied. MATLAB code for the presented metrics and visualizations are publicly available [31]; it is hoped that others will experiment with these methods and build upon this work for the benefit of the machine learning community and to improve the products of their efforts.

REFERENCES

[1] C. Khosla and B. S. Saini, "Enhancing performance of deep learning models with different data augmentation techniques: A survey," in *2020 International Conference on Intelligent Engineering and Management (ICIEM)*, 2020, pp. 79–85.

[2] A. Y. Choquenaira Florez, L. Scabora, S. Amer-Yahia, and J. F. Rodrigues Júnior, "Augmentation techniques for sequential clinical data to improve deep learning prediction techniques," in *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*, 2020, pp. 597–602.

[3] M. A. Bansal, D. R. Sharma, and D. M. Kathuria, "A systematic review on data scarcity problem in deep learning: solution and applications," *ACM Computing Surveys (CSUR)*, vol. 54, no. 10s, pp. 1–29, 2022.

[4] L. Taylor and G. Nitschke, "Improving deep learning with generic data augmentation," in *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2018, pp. 1542–1547.

[5] D. Heaven, "Deep trouble for deep learning," *Nature*, vol. 574, no. 7777, pp. 163–166, 2019.

[6] J. N. Kahlen, A. Wurde, M. Andres, and A. Moser, "Improving machine learning diagnostic systems with model-based data augmentation — part b: Application," in *2021 IEEE PES Innovative Smart Grid Technologies Europe (ISGT Europe)*, 2021, pp. 1–6.

[7] H. L. Bear, V. Morfi, and E. Benetos, "An evaluation of data augmentation methods for sound scene geotagging," in *Proc. Interspeech 2021*, 2021, pp. 581–585.

[8] A. Mumuni and F. Mumuni, "Data augmentation: A comprehensive survey of modern approaches," *Array*, vol. 16, pp. 100258, 2022.

[9] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020, pp. 3008–3017.

[10] D. Heise and H. L. Bear, "Visually exploring multi-purpose audio data," in *2021 IEEE 23rd International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 2021, pp. 1–6.

[11] H. L. Bear, T. Heittola, A. Mesaros, E. Benetos, and T. Virtanen, "City classification from multiple real-world sound scenes," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019, pp. 11–15.

[12] M. Mandel, J. Salamon, and D. P. W. Ellis, *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, New York University, NY, USA, October 2019.

[13] K. Drossos, S. Adavanne, and T. Virtanen, "Automated audio captioning with recurrent neural networks," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017, pp. 374–378.

[14] K. Imoto, "Acoustic scene classification using multichannel observation with partially missing channels," in *2021 29th European Signal Processing Conference (EUSIPCO)*, 2021, pp. 875–879.

[15] Ç. Bilen, G. Ferroni, F. Tuveri, J. Azcarreta, and S. Krstulović, "A framework for the robust evaluation of sound event detection," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 61–65.

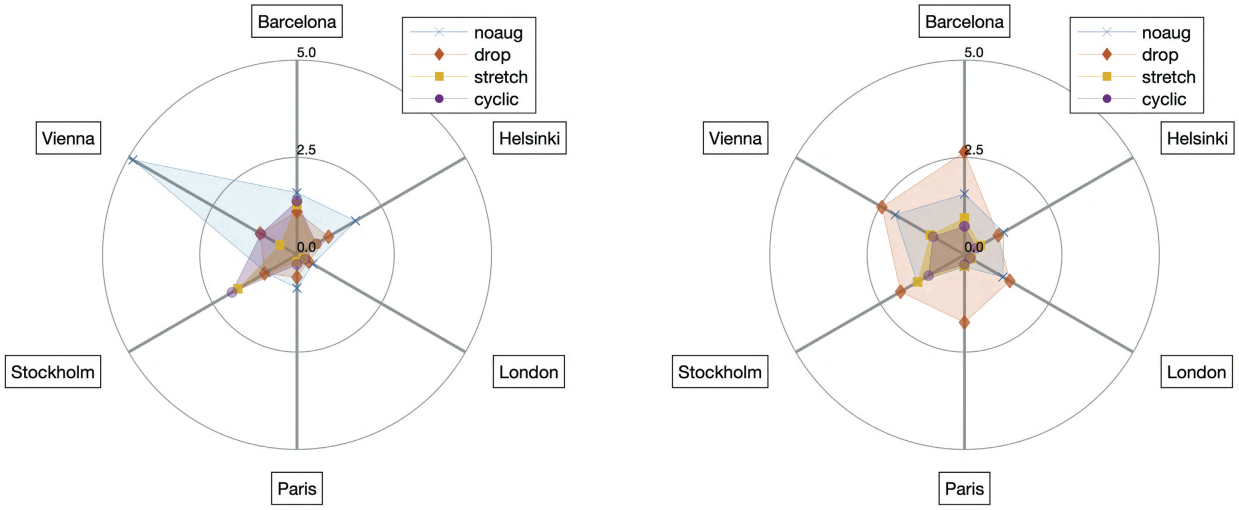

Fig. 7. Illustrating the distance components (dc) of the nirvana distance (ND) for each of the augmentations under study. Points on radial axes represent dc values for each class. Augmentations with smaller radii (and thus smaller shaded area) have lower ND and are thus closer to "ideal". Left: single-task models. Notice the large *Vienna* dc for no augmentation; compare to Fig. 5, row 1, column 1. Right: multi-task models. Notice the relatively large dc for all cities using *drop* augmentation; compare to Fig. 5, row 2, column 2.

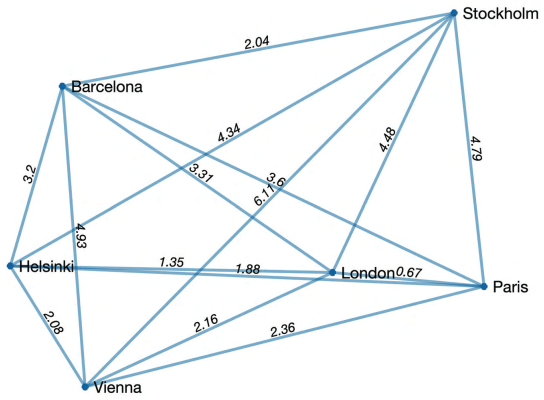

Fig. 8. Distances between city classes (F) in the feature space of the original (non-augmented) data; distances provided by [11].

[16] I. Nolasco et al., "Few-shot bioacoustic event detection at the DCASE 2022 challenge," in *Proceedings of the 7th Detection and Classification of Acoustic Scenes and Events 2022 Workshop (DCASE2022)*, Nancy, France, November 2022.

[17] F. Gontier, M. Lagrange, C. Lavandier, and J.-F. Petiot, "Privacy aware acoustic scene synthesis using deep spectral feature inversion," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 886–890.

[18] K. Imoto, N. Tonami, Y. Koizumi, M. Yasuda, R. Yamanishi, and Y. Yamashita, "Sound event detection by multitask learning of sound events and scenes with soft scene labels," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 621–625.

[19] A. Kumar, B. Elizalde, and B. Raj, "Audio Content Based Geotagging in Multimedia," in *Proc. Interspeech 2017*, 2017, pp. 1874–1878.

[20] J. C. Bezdek and R. J. Hathaway, "VAT: A tool for visual assessment of (cluster) tendency," in *Proc. 2002 Int. Joint Conf. Neural Netw. IJCNN'02 (Cat. No.02CH37290)*, 2002, vol. 3, pp. 2225–2230.

[21] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)*, November 2018, pp. 9–13.

[22] S. Wei, S. Zou, F. Liao, and W. Lang, "A comparison on data augmentation methods based on deep learning for audio classification," in *Journal of Physics: Conference Series*. IOP Publishing, 2020, vol. 1453, p. 012085.

[23] O. O. Abayomi-Alli, R. Damaševičius, A. Qazi, M. Adedoyin-Olowe, and S. Misra, "Data augmentation and deep learning methods in sound classification: A systematic review," *Electronics*, vol. 11, no. 22, 2022.

[24] L. Zhang, Z. Deng, K. Kawaguchi, A. Ghorbani, and J. Zou, "How does mixup help with robustness and generalization?," in *International Conference on Learning Representations*, 2021.

[25] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, 2015, vol. 8, pp. 18–25.

[26] D. Kumar and J. C. Bezdek, "Visual approaches for exploratory data analysis: A survey of the visual assessment of clustering tendency (VAT) family of algorithms," *IEEE Systems, Man, and Cybernetics Magazine*, vol. 6, no. 2, pp. 10–48, 2020.

[27] K. Rajendra Prasad, "Gaussian mixture model (GMM) based k-means method for speech clustering," *International Journal of Advanced Research in Computer Science*, vol. 8, no. 7, pp. 643–647, 2009.

[28] T. C. Havens, J. C. Bezdek, J. M. Keller, M. Popescu, and J. M. Huband, "Is VAT really single linkage in disguise?," *Annals of Mathematics and Artificial Intelligence*, vol. 55, pp. 237–251, 2009.

[29] Y. Rubner, C. Tomasi, and L. Guibas, "A metric for distributions with applications to image databases," in *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*, 1998, pp. 59–66.

[30] S. Chen, E. Dobriban, and J. H. Lee, "A group-theoretic framework for data augmentation," *J. Mach. Learn. Res.*, vol. 21, no. 1, January 2020.

[31] D. Heise, "MATLAB File Exchange community profile," https://www.mathworks.com/matlabcentral/profile/authors/4656280.