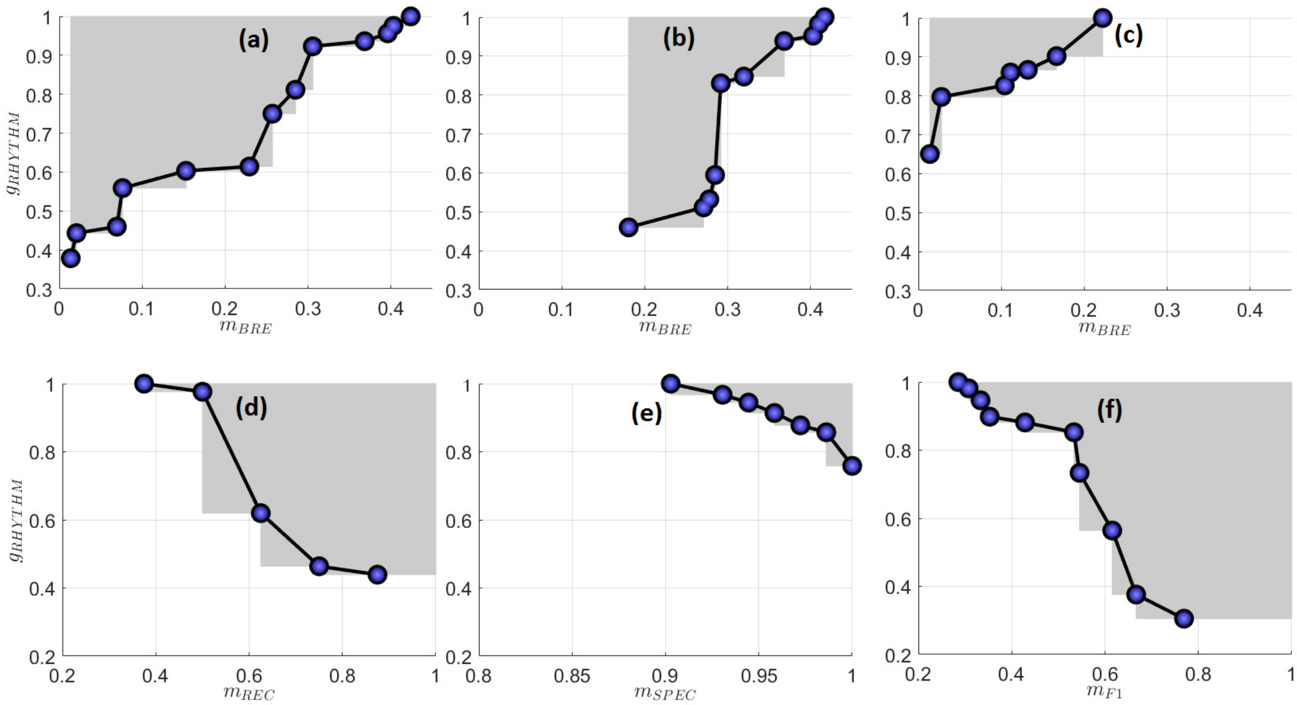

Figure 1. Non-dominated fronts after multi-objective feature selection for the identification of Traditional Blues in the SLAC dataset [2], with rhythmic descriptors the feature group being focused on. Top row: the share of rhythmic descriptors g_{RHYTHM} is maximized and the balanced relative error m_{BRE} is minimized using random forest (a), k-nearest neighbors (b), or support vector machine (c) classifiers. Bottom row: random forest classifiers are used to maximize both g_{RHYTHM} and recall m_{REC} (d), specificity m_{SPEC} (e), or F1-measure m_{F1} (f).

3. DESIGN OF EXPERIMENTS

We conducted our experiments on two pre-existing datasets, namely LMD-aligned [32] and SLAC [2], involving a total of 20 genres and sub-genres. While LMD-aligned is a larger dataset, it has problems with noisiness and structure (e.g., it is unbalanced, with an overly large representation of Pop/Rock songs). SLAC is smaller but carefully designed with, for example, an equal number of pieces belonging to each genre, each of which can be broken into two equally represented sub-genres. We made use of multi-modal features extracted from 1,575 Lakh and 250 SLAC tracks, which have been made publicly available. [1]

Table 1 shows sample jSymbolic features, divided into feature groups; the full list of symbolic (and other) features is available online. [2] We excluded the dynamics features used in [10], as dynamics are often inconsistently encoded.

In the first part of this study we measure the importance of eight groups of symbolic features (relative to features from all six modalities considered) not only with the random forests (RF) classifiers used in [10], but also with k-nearest neighbor (kNN) and support vector machine (SVM) classifiers. As can be observed in Figure 1, subfigures (a)-(c), notable differences in importance can be evident when the classification method is changed. Of course, a limitation of this study is that all classifiers are applied with default hyperparameters in the AMUSE framework [33]: 100 trees for RF, $k = 1$ for kNN and a linear kernel for SVM. In practice, varying hyperparameters may well also introduce meaningful variance in measured feature importances.

We chose to omit deep neural networks from our experiments, despite their popularity in MIR, for two reasons. First, they typically learn their own features, which can make it difficult to analyze the relative relevance of interpretable semantic descriptors for a given musical category. Second, the typically very large number of parameters they involve can lead to overfitting in situations with limited available data, such as when a musical category is defined by a small number of “positive” and “negative” tracks, as in real-world situations where a listener may wish to provide only a few labeled examples to train a supervised classification model.

In the second part of this study we vary the evaluation measures used in the two-objective feature selection. Although m_{BRE} is generally a good choice for binary classification tasks, as it can help to measure performance with unbalanced test sets [34, p.344], it is not often used in MIR classification studies. We have therefore extended the setup from [10] with three additional evaluation criteria. Recall and specificity measure classification errors associated with, respectively, instances annotated in the ground truth as belonging or not belonging to a category [34, p.342]. F1-measure is a weighted combination of precision and recall that, like m_{BRE}, can be useful for less

[1] `https://zenodo.org/record/5651429`
[2] `https://doi.org/10.5334/tismir.67.s1`

Group	Feature Examples
Pitch	First pitch, last pitch, major or minor, pitch class histogram, pitch variability, range
Melodic	Amount of arpeggiation, direction of melodic motion, melodic intervals, repeated notes
Chords	Chord type histogram, dominant seventh chords, variability of number of simultaneous pitches
Rhythm	Initial time signature, metrical diversity, note density per quarter note, prevalence of dotted notes
Tempo	Initial tempo, mean tempo, minimum and maximum note duration, note density and its variation
Instrument presence	Note prevalences of pitched and unpitched instruments, pitched instruments present
Instrument prevalence	Prevalences of individual instruments/instrument groups: acoustic guitar, string ensemble, etc.
Texture	Average number of independent voices, parallel fifths and octaves, voice overlap

Table 1. Sample jSymbolic [24] features grouped into eight semantically meaningful groups.

balanced datasets [34, p.344]. The potential for differences in measured importances based on recall, specificity, and F1 is demonstrated in subfigures (d), (e), and (f).

We conducted 24,000 feature selection experiments involving 20 genres and sub-genres $\times$ 3 cross-validation folds $\times$ 8 feature groups $\times$ 5 new combinations of classifiers and measures $\times$ 10 statistical repetitions, with runtimes between about 2 and 50 hours per experiment.

4. DISCUSSION OF RESULTS

Tables 2 and 3 show multi-group feature importances for all 20 classes, with each value indicating results aggregated over three folds. In each row, the importance values across the eight feature groups are shown with color: the group with the highest importance (compared to the other feature groups in the same row) is marked in deep red, and the group with the lowest importance in deep blue. The four "more important" feature groups are shown in shades of red, and four "less important" groups in shades of blue.

The instrument presence feature group seems to be the most important group in most of the experiments, with the instrument prevalence group being the most important in a few others. Features measuring pitch statistics are the most important in five experiments (e.g., RnB prediction using kNN and m_{BRE}). Interestingly, instrument prevalence is the least important group for almost all LMD-aligned genres (even when instrument presence is the most important group). Another intriguing result is that melodic features seem to be particularly unimportant for the Classical SLAC genre and its ClassBaroq and ClassRomant sub-genres.

As anticipated in Figure 1, these results reveal differences between importances when the classifier is varied. Such difference may be particularly meaningful when a cell's color changes from red to blue or vice versa: for example, chords are among the four more important groups for predicting Country using RF and SVM, but are among the less important groups when kNN is used. On the other hand, for this same genre tempo-based features become more important using kNN, not only relatively (the cell goes from blue to red), but also with respect to mean importance values (0.861 instead of 0.656 and 0.651, respectively). This supports our preliminary suspicion that the choice of classifier can have a strong impact, even for "robust" features groups (e.g., symbolic characteristics arguably describe musical properties in concise and clearly understood ways compared to audio descriptors). However, this is certainly not always the case: for 480 combinations of 20 classes, 8 feature groups, and 3 classifiers, in 62 cases (12.92%) the cell shade remains the same when the classifier is changed for a fixed genre and feature group.

When the evaluation measures are varied for a fixed classifier (note that here results were restricted to RF), the changes seem to be slightly less impactful: for 640 related combinations (20 classes, 8 feature groups, 4 measures), the shade changed in only 53 cases (8.28%).

A randomly chosen decision will assign a feature group to be either "more" or "less" relatively important with an expected probability of 25% for three classifiers (1 case with all more important values, 1 case with all less important values, and 6 remaining cases), and with 12.5% for the four measures. Although this interpretation is not perfect, as the change of a light red shade to a light blue shade will indicate a switch between "more" and "less" important, our main inference from the complete study is that it is not enough to claim that some feature group is "generally" more or less important for a particular musical category. Our results suggest that it is indeed necessary to accompany feature group evaluations with specification of the classifier and evaluation measure used. More general claims about the suitability of particular feature groups should be substantiated with broader experiments involving multiple classifiers and evaluation measures (and, perhaps, classifier hyper-parameters).

5. CONCLUSIONS

This paper studied the stability of multi-group feature importance when classifiers and evaluation measures are varied. Eight symbolic feature groups were focused on within a multi-modal classification context involving features extracted from six modalities (symbolic and five others). Various combinations of features and classification approaches were used to predict genres and sub-genres for two datasets with publicly available pre-extracted feature values. The results show that, although in most cases the relative importance of individual feature groups is not affected by the choice of classification method or evaluation measure, either or both of these do nonetheless have an influence in a non-negligible number of cases. Multiple parameters of the feature importance estimation chain can impact determination of which musical properties are more or less relevant for a particular musical category.

In the future, we plan to continue our experiments by

Genre		Pitch	Melodic	Chords	Rhythm	Tempo	Instr. Pres.	Instr. Prev.	Texture
Country	RF	0.796±0.05	0.675±0.07	0.798±0.02	0.820±0.04	0.656±0.03	0.945±0.01	0.614±0.02	0.700±0.03
	kNN	0.924±0.08	0.704±0.07	0.832±0.03	0.865±0.04	0.861±0.06	0.952±0.03	0.645±0.02	0.779±0.03
	SVM	0.756±0.03	0.688±0.04	0.785±0.04	0.730±0.03	0.651±0.05	0.861±0.03	0.633±0.01	0.665±0.02
	RF-F1	0.610±0.09	0.439±0.03	0.576±0.10	0.654±0.05	0.413±0.11	0.889±0.01	0.285±0.12	0.381±0.08
	RF-Rec	0.849±0.01	0.722±0.06	0.823±0.03	0.882±0.02	0.755±0.08	0.965±0.01	0.668±0.06	0.733±0.08
	RF-Spec	0.760±0.03	0.678±0.06	0.768±0.04	0.785±0.04	0.618±0.06	0.927±0.02	0.604±0.09	0.674±0.03
Electronic	RF	0.825±0.04	0.774±0.02	0.815±0.01	0.871±0.02	0.754±0.07	0.951±0.02	0.697±0.03	0.746±0.02
	kNN	0.867±0.07	0.748±0.11	0.861±0.06	0.914±0.06	0.954±0.01	0.964±0.02	0.700±0.04	0.708±0.05
	SVM	0.824±0.04	0.773±0.01	0.770±0.05	0.836±0.05	0.768±0.00	0.916±0.01	0.683±0.03	0.758±0.03
	RF-F1	0.695±0.05	0.583±0.05	0.681±0.02	0.771±0.02	0.542±0.08	0.934±0.01	0.450±0.01	0.553±0.02
	RF-Rec	0.877±0.03	0.751±0.08	0.830±0.05	0.911±0.03	0.738±0.12	0.944±0.01	0.718±0.04	0.759±0.04
	RF-Spec	0.757±0.09	0.738±0.05	0.748±0.02	0.835±0.02	0.742±0.04	0.956±0.01	0.660±0.03	0.733±0.01
Pop	RF	0.752±0.05	0.694±0.07	0.735±0.04	0.806±0.04	0.701±0.04	0.927±0.02	0.626±0.06	0.713±0.04
	kNN	0.946±0.04	0.898±0.01	0.908±0.01	0.982±0.01	0.961±0.03	0.993±0.00	0.837±0.06	0.906±0.08
	SVM	0.799±0.08	0.718±0.08	0.777±0.02	0.780±0.04	0.781±0.01	0.875±0.02	0.738±0.02	0.795±0.04
	RF-F1	0.788±0.02	0.668±0.06	0.782±0.02	0.786±0.03	0.732±0.05	0.916±0.04	0.649±0.05	0.748±0.01
	RF-Rec	0.803±0.10	0.729±0.04	0.769±0.11	0.870±0.07	0.722±0.13	0.959±0.02	0.643±0.03	0.745±0.09
	RF-Spec	0.575±0.08	0.510±0.10	0.593±0.11	0.612±0.06	0.497±0.04	0.899±0.03	0.475±0.09	0.569±0.08
RnB	RF	0.808±0.04	0.689±0.10	0.807±0.05	0.844±0.03	0.716±0.02	0.958±0.02	0.621±0.09	0.697±0.07
	kNN	0.921±0.06	0.601±0.08	0.778±0.00	0.892±0.06	0.818±0.20	0.920±0.06	0.402±0.02	0.743±0.07
	SVM	0.764±0.07	0.699±0.03	0.780±0.04	0.788±0.04	0.735±0.01	0.889±0.03	0.663±0.10	0.691±0.07
	RF-F1	0.711±0.06	0.418±0.16	0.603±0.03	0.661±0.03	0.448±0.17	0.892±0.00	0.306±0.07	0.417±0.03
	RF-Rec	0.818±0.01	0.778±0.03	0.834±0.03	0.861±0.04	0.738±0.06	0.972±0.00	0.665±0.04	0.679±0.08
	RF-Spec	0.829±0.06	0.683±0.02	0.738±0.02	0.832±0.02	0.704±0.04	0.930±0.01	0.601±0.07	0.637±0.02
Rock	RF	0.726±0.06	0.602±0.03	0.713±0.07	0.748±0.03	0.655±0.03	0.917±0.02	0.598±0.02	0.657±0.05
	kNN	0.949±0.05	0.856±0.03	0.919±0.02	0.961±0.04	0.983±0.01	0.984±0.01	0.813±0.05	0.948±0.02
	SVM	0.768±0.03	0.756±0.01	0.766±0.02	0.780±0.03	0.710±0.04	0.886±0.04	0.625±0.03	0.686±0.04
	RF-F1	0.767±0.03	0.641±0.08	0.742±0.01	0.773±0.05	0.679±0.08	0.928±0.02	0.608±0.03	0.663±0.04
	RF-Rec	0.808±0.04	0.693±0.10	0.766±0.05	0.887±0.01	0.784±0.06	0.968±0.01	0.641±0.07	0.731±0.01
	RF-Spec	0.612±0.08	0.446±0.04	0.586±0.08	0.603±0.05	0.482±0.07	0.859±0.06	0.509±0.12	0.508±0.05
Blues	RF	0.923±0.03	0.742±0.07	0.858±0.06	0.902±0.03	0.759±0.05	0.975±0.00	0.769±0.10	0.723±0.13
	kNN	0.743±0.06	0.650±0.20	0.659±0.11	0.841±0.04	0.725±0.14	0.945±0.00	0.576±0.04	0.624±0.07
	SVM	0.792±0.02	0.557±0.14	0.665±0.20	0.795±0.08	0.584±0.04	0.916±0.01	0.621±0.13	0.627±0.03
	RF-F1	0.806±0.09	0.551±0.20	0.656±0.11	0.801±0.07	0.534±0.04	0.953±0.02	0.645±0.03	0.525±0.12
	RF-Rec	0.984±0.01	0.865±0.07	0.910±0.10	0.982±0.02	0.891±0.05	0.994±0.00	0.904±0.09	0.865±0.12
	RF-Spec	0.943±0.03	0.826±0.10	0.954±0.02	0.899±0.01	0.894±0.01	0.984±0.00	0.898±0.05	0.857±0.06
Classical	RF	0.934±0.02	0.782±0.04	0.962±0.01	0.949±0.01	0.872±0.02	0.996±0.00	0.945±0.05	0.885±0.05
	kNN	0.858±0.04	0.690±0.08	0.844±0.04	0.829±0.04	0.811±0.01	0.990±0.01	0.945±0.05	0.791±0.08
	SVM	0.872±0.01	0.706±0.06	0.915±0.01	0.836±0.04	0.837±0.04	0.979±0.02	0.963±0.03	0.812±0.03
	RF-F1	0.848±0.04	0.679±0.04	0.910±0.04	0.878±0.04	0.722±0.04	0.989±0.01	0.889±0.10	0.738±0.08
	RF-Rec	0.987±0.01	0.925±0.07	0.989±0.02	0.982±0.02	0.952±0.03	1.000±0.00	1.000±0.00	0.983±0.03
	RF-Spec	0.953±0.02	0.836±0.07	0.981±0.02	0.960±0.00	0.892±0.04	0.992±0.01	0.925±0.07	0.905±0.06
Rock	RF	0.933±0.00	0.887±0.03	0.911±0.07	0.917±0.03	0.833±0.05	0.995±0.00	0.915±0.04	0.905±0.02
	kNN	0.791±0.15	0.716±0.09	0.753±0.08	0.802±0.13	0.596±0.32	0.995±0.00	0.804±0.01	0.702±0.11
	SVM	0.808±0.02	0.827±0.09	0.816±0.09	0.797±0.04	0.758±0.03	0.982±0.00	0.866±0.06	0.749±0.06
	RF-F1	0.799±0.03	0.705±0.06	0.843±0.05	0.804±0.02	0.628±0.05	0.984±0.01	0.787±0.04	0.743±0.03
	RF-Rec	0.960±0.03	0.971±0.03	0.982±0.03	0.969±0.03	0.910±0.07	1.000±0.00	1.000±0.00	0.974±0.02
	RF-Spec	0.956±0.03	0.878±0.01	0.955±0.01	0.947±0.01	0.870±0.02	0.994±0.00	0.928±0.02	0.941±0.06
Jazz	RF	0.967±0.01	0.924±0.02	0.965±0.03	0.966±0.00	0.900±0.02	0.996±0.00	0.908±0.02	0.888±0.06
	kNN	0.886±0.03	0.566±0.05	0.764±0.10	0.772±0.11	0.754±0.11	0.985±0.01	0.606±0.11	0.618±0.23
	SVM	0.898±0.03	0.832±0.04	0.888±0.05	0.777±0.05	0.821±0.04	0.985±0.00	0.820±0.01	0.813±0.04
	RF-F1	0.913±0.01	0.815±0.05	0.929±0.05	0.898±0.04	0.837±0.10	0.990±0.00	0.805±0.02	0.748±0.10
	RF-Rec	0.972±0.02	0.966±0.04	0.991±0.01	0.988±0.02	0.933±0.04	0.998±0.00	0.981±0.03	0.927±0.01
	RF-Spec	0.970±0.02	0.952±0.01	0.983±0.02	0.967±0.01	0.971±0.01	0.999±0.00	0.940±0.01	0.941±0.01
Rap	RF	0.967±0.01	0.920±0.02	0.930±0.05	0.928±0.02	0.861±0.02	0.996±0.00	0.849±0.03	0.912±0.04
	kNN	0.955±0.01	0.731±0.10	0.839±0.04	0.828±0.04	0.797±0.07	0.979±0.00	0.711±0.09	0.555±0.06
	SVM	0.880±0.04	0.705±0.07	0.887±0.06	0.793±0.01	0.788±0.04	0.955±0.01	0.798±0.07	0.823±0.07
	RF-F1	0.918±0.03	0.796±0.01	0.818±0.10	0.847±0.07	0.748±0.06	0.992±0.01	0.762±0.03	0.737±0.21
	RF-Rec	0.990±0.02	0.936±0.01	0.988±0.02	0.959±0.06	0.936±0.03	0.999±0.00	0.960±0.04	0.969±0.04
	RF-Spec	0.975±0.02	0.945±0.01	0.939±0.06	0.950±0.03	0.898±0.03	0.999±0.00	0.938±0.02	0.910±0.05

Table 2. Multi-group symbolic feature importances for five LMD-aligned genres (top half of the table) and five SLAC parent genres (bottom half of the table), aggregated over 3 folds. F1: F1-measure; Rec: recall; Spec: specificity. The first three rows of each genre block were evaluated with balanced relative error m_{BRE}.

measuring the impact of other feature types and modalities. We will also further examine the effects of varying other parameters in the experimental setup, such as classifier hyper-parameters, and also systematically consider as-

		Pitch	Melodic	Chords	Rhythm	Tempo	Instr. Pres.	Instr. Prev.	Texture
BluesModern	RF	0.922±0.05	0.711±0.07	0.775±0.04	0.701±0.18	0.732±0.19	0.965±0.04	0.826±0.05	0.837±0.14
	kNN	0.762±0.07	0.772±0.13	0.631±0.08	0.698±0.03	0.773±0.08	0.966±0.04	0.536±0.07	0.774±0.16
	SVM	0.865±0.03	0.602±0.02	0.660±0.16	0.758±0.12	0.641±0.13	0.946±0.03	0.684±0.17	0.614±0.07
	RF-F1	0.916±0.07	0.488±0.39	0.603±0.40	0.704±0.33	0.511±0.29	0.929±0.09	0.510±0.27	0.711±0.22
	RF-Rec	0.769±0.20	0.629±0.17	0.640±0.07	0.498±0.30	0.895±0.10	0.991±0.02	0.885±0.20	0.794±0.07
	RF-Spec	0.984±0.01	0.966±0.02	0.970±0.03	0.982±0.02	0.963±0.01	0.997±0.00	0.928±0.02	0.962±0.04
BluesTradit	RF	0.836±0.10	0.664±0.10	0.909±0.02	0.949±0.03	0.778±0.18	0.971±0.01	0.844±0.09	0.673±0.14
	kNN	0.798±0.12	0.619±0.14	0.755±0.09	0.851±0.07	0.768±0.10	0.935±0.04	0.805±0.18	0.537±0.10
	SVM	0.844±0.05	0.672±0.14	0.618±0.15	0.773±0.14	0.748±0.10	0.964±0.05	0.598±0.13	0.608±0.11
	RF-F1	0.852±0.04	0.595±0.23	0.645±0.13	0.852±0.04	0.629±0.24	0.960±0.04	0.694±0.19	0.505±0.16
	RF-Rec	0.872±0.10	0.747±0.17	0.943±0.10	0.905±0.14	0.944±0.10	0.999±0.00	0.624±0.45	0.782±0.38
	RF-Spec	0.993±0.01	0.911±0.05	0.916±0.05	0.971±0.03	0.918±0.10	0.991±0.01	0.922±0.09	0.855±0.14
ClassBaroq	RF	0.914±0.07	0.747±0.26	0.921±0.06	0.804±0.11	0.815±0.14	0.997±0.00	0.816±0.03	0.860±0.14
	kNN	0.821±0.06	0.486±0.40	0.754±0.14	0.654±0.03	0.572±0.18	0.985±0.01	0.964±0.04	0.421±0.12
	SVM	0.901±0.02	0.595±0.04	0.856±0.05	0.774±0.06	0.625±0.04	0.985±0.01	0.909±0.08	0.740±0.10
	RF-F1	0.851±0.21	0.520±0.14	0.701±0.30	0.688±0.21	0.593±0.08	0.987±0.02	0.728±0.18	0.539±0.17
	RF-Rec	0.995±0.01	0.861±0.13	0.971±0.05	0.875±0.22	0.778±0.29	1.000±0.00	0.883±0.11	0.976±0.02
	RF-Spec	0.970±0.02	0.922±0.08	0.986±0.01	0.988±0.01	0.954±0.04	1.000±0.00	1.000±0.00	0.934±0.07
ClassRomant	RF	0.831±0.12	0.710±0.14	0.940±0.04	0.893±0.07	0.781±0.12	1.000±0.00	0.895±0.14	0.670±0.09
	kNN	0.888±0.09	0.670±0.25	0.789±0.12	0.947±0.05	0.817±0.14	0.983±0.01	0.804±0.13	0.722±0.23
	SVM	0.817±0.07	0.723±0.18	0.952±0.04	0.860±0.05	0.738±0.04	0.983±0.02	0.971±0.02	0.746±0.11
	RF-F1	0.744±0.04	0.510±0.23	0.864±0.09	0.855±0.03	0.579±0.13	1.000±0.00	0.731±0.10	0.621±0.09
	RF-Rec	0.792±0.14	0.724±0.24	0.974±0.02	0.848±0.16	0.480±0.25	0.999±0.00	1.000±0.00	0.806±0.09
	RF-Spec	0.979±0.02	0.942±0.05	0.999±0.00	0.993±0.01	0.980±0.01	0.999±0.00	0.986±0.01	0.971±0.02
JazzBop	RF	0.957±0.01	0.767±0.15	0.855±0.04	0.824±0.08	0.863±0.03	0.966±0.03	0.818±0.10	0.865±0.03
	kNN	0.761±0.11	0.604±0.07	0.838±0.04	0.751±0.02	0.711±0.08	0.994±0.00	0.713±0.16	0.572±0.06
	SVM	0.844±0.01	0.720±0.06	0.840±0.02	0.784±0.06	0.816±0.06	0.973±0.03	0.713±0.12	0.759±0.05
	RF-F1	0.788±0.03	0.609±0.35	0.722±0.16	0.811±0.03	0.681±0.03	0.962±0.05	0.571±0.15	0.568±0.09
	RF-Rec	0.949±0.04	0.653±0.33	0.774±0.06	0.904±0.11	0.965±0.03	0.954±0.08	1.000±0.00	0.839±0.18
	RF-Spec	0.994±0.01	0.985±0.02	0.995±0.01	0.994±0.01	0.991±0.01	1.000±0.00	0.960±0.05	0.930±0.05
JazzSwing	RF	0.957±0.01	0.863±0.07	0.911±0.06	0.946±0.05	0.932±0.04	0.947±0.04	0.876±0.05	0.861±0.08
	kNN	0.939±0.07	0.840±0.10	0.887±0.08	0.825±0.12	0.707±0.14	0.983±0.01	0.738±0.05	0.749±0.22
	SVM	0.870±0.06	0.823±0.05	0.903±0.07	0.840±0.07	0.807±0.06	0.946±0.03	0.822±0.07	0.800±0.10
	RF-F1	0.868±0.07	0.658±0.05	0.788±0.10	0.888±0.03	0.604±0.16	0.943±0.03	0.759±0.16	0.650±0.08
	RF-Rec	0.988±0.02	0.942±0.06	0.840±0.16	0.943±0.08	0.901±0.12	0.962±0.03	0.867±0.13	0.828±0.26
	RF-Spec	0.994±0.01	0.966±0.01	0.978±0.02	0.996±0.01	0.964±0.02	0.999±0.00	0.978±0.00	0.979±0.02
RapHardcore	RF	0.845±0.00	0.790±0.07	0.898±0.06	0.880±0.03	0.688±0.18	0.965±0.01	0.693±0.11	0.883±0.11
	kNN	0.816±0.09	0.685±0.14	0.801±0.13	0.725±0.07	0.661±0.06	0.969±0.03	0.677±0.24	0.686±0.02
	SVM	0.768±0.10	0.779±0.05	0.820±0.13	0.718±0.18	0.712±0.12	0.913±0.07	0.719±0.10	0.813±0.06
	RF-F1	0.842±0.14	0.727±0.05	0.649±0.23	0.949±0.05	0.838±0.09	0.978±0.01	0.576±0.39	0.842±0.16
	RF-Rec	0.660±0.27	0.700±0.24	0.981±0.02	0.985±0.03	0.888±0.07	0.970±0.01	0.829±0.15	0.954±0.05
	RF-Spec	0.993±0.01	0.959±0.01	0.985±0.01	0.987±0.02	0.964±0.03	1.000±0.00	0.972±0.02	0.989±0.01
RapPop	RF	0.851±0.08	0.637±0.03	0.926±0.04	0.883±0.06	0.787±0.19	0.953±0.01	0.644±0.16	0.658±0.17
	kNN	0.907±0.09	0.629±0.11	0.790±0.12	0.736±0.04	0.732±0.12	0.987±0.01	0.643±0.17	0.668±0.08
	SVM	0.854±0.07	0.776±0.08	0.863±0.09	0.682±0.10	0.625±0.32	0.895±0.06	0.511±0.17	0.756±0.09
	RF-F1	0.835±0.12	0.626±0.38	0.862±0.11	0.755±0.22	0.637±0.32	0.971±0.03	0.710±0.25	0.582±0.21
	RF-Rec	0.972±0.05	0.761±0.18	0.910±0.16	0.529±0.09	0.603±0.41	0.957±0.07	0.734±0.06	0.681±0.48
	RF-Spec	0.982±0.01	0.965±0.04	0.961±0.02	0.961±0.04	0.960±0.02	1.000±0.00	0.935±0.04	0.944±0.03
RockAltern	RF	0.787±0.06	0.768±0.10	0.815±0.07	0.891±0.02	0.824±0.08	0.971±0.01	0.728±0.09	0.681±0.05
	kNN	0.658±0.14	0.678±0.07	0.639±0.10	0.830±0.14	0.751±0.08	0.952±0.05	0.423±0.24	0.593±0.19
	SVM	0.670±0.12	0.679±0.13	0.797±0.08	0.804±0.14	0.751±0.11	0.903±0.07	0.695±0.10	0.592±0.04
	RF-F1	0.535±0.20	0.458±0.05	0.764±0.14	0.740±0.05	0.597±0.13	0.911±0.07	0.442±0.15	0.498±0.24
	RF-Rec	0.578±0.19	0.794±0.19	0.855±0.04	0.884±0.03	0.847±0.22	0.942±0.04	0.698±0.08	0.765±0.10
	RF-Spec	0.977±0.03	0.918±0.06	0.987±0.02	0.980±0.03	0.907±0.11	1.000±0.00	0.963±0.04	0.934±0.01
RockMetal	RF	0.914±0.03	0.840±0.06	0.873±0.09	0.847±0.11	0.840±0.04	0.988±0.00	0.950±0.03	0.840±0.09
	kNN	0.986±0.02	0.797±0.12	0.797±0.18	0.835±0.04	0.720±0.30	0.985±0.01	0.795±0.17	0.601±0.17
	SVM	0.758±0.25	0.788±0.05	0.915±0.04	0.826±0.03	0.750±0.08	0.981±0.02	0.946±0.04	0.810±0.03
	RF-F1	0.943±0.04	0.605±0.14	0.742±0.15	0.739±0.18	0.630±0.15	0.985±0.01	0.802±0.04	0.775±0.04
	RF-Rec	0.965±0.03	0.703±0.23	0.795±0.26	0.757±0.16	0.823±0.05	0.991±0.01	1.000±0.00	0.807±0.16
	RF-Spec	0.995±0.01	0.972±0.02	0.985±0.00	0.992±0.01	0.965±0.03	1.000±0.00	0.992±0.01	0.990±0.01

Table 3. Multi-group symbolic feature importances for ten SLAC sub-genres, aggregated over 3 folds. F1: F1-measure; Rec: recall; Spec: specificity. The first three rows of each sub-genre block were evaluated with balanced relative error m_{BRE}.

pects like extraction times, statistical properties, and suitability for data augmentation. We will also run further trials in order to be able to apply more developed statistical significance testing.

6. ACKNOWLEDGMENTS

The experiments were carried out on the Linux HPC cluster at TU Dortmund (LiDO3), which was partially funded by the Large-Scale Equipment Initiative by the German Research Foundation (DFG) grant 271512359. The second author's work was funded by the Fonds de recherche du Québec – Société et culture, under grants 2021-CHZ-282456 and 2022-CHZ-309882.

7. REFERENCES

[1] C. McKay and I. Fujinaga, "Combining features extracted from audio, symbolic and cultural sources," in *Proc. of the 9th International Conference on Music Information Retrieval, ISMIR*, 2008, pp. 597–602.

[2] C. McKay, J. A. Burgoyne, J. Hockman, J. B. L. Smith, G. Vigliensoni, and I. Fujinaga, "Evaluating the genre classification performance of lyrical features relative to audio, symbolic and cultural features," in *Proc. of the 11th International Society for Music Information Retrieval Conference, ISMIR*, 2010, pp. 213–218.

[3] B. McFee and G. R. G. Lanckriet, "Hypergraph models of playlist dialects," in *Proc. of the 13th International Society for Music Information Retrieval Conference, ISMIR*, 2012, pp. 343–348.

[4] R. Panda, R. Malheiro, B. Rocha, A. Oliveira, and R. P. Paiva, "Multi-modal music emotion recognition: A new dataset, methodology and comparative analysis," in *Proc. of the 10th International Symposium on Computer Music Multidisciplinary Research, CMMR*. Springer, 2013.

[5] S. Oramas, O. Nieto, F. Barbieri, and X. Serra, "Multi-label music genre classification from audio, text and images using deep features," in *Proc. of the 18th International Society for Music Information Retrieval Conference, ISMIR*, 2017, pp. 23–30.

[6] S. Oramas, F. Barbieri, O. Nieto, and X. Serra, "Multimodal deep learning for music genre classification," *Transactions of the International Society for Music Information Retrieval*, vol. 1, no. 1, pp. 4–21, 2018.

[7] N. Orio, D. Rizo, R. Miotto, M. Schedl, N. Montecchio, and O. Lartillot, "Musiclef: a benchmark activity in multimodal music information retrieval," in *Proc. of the 12th International Society for Music Information Retrieval Conference, ISMIR*, 2011, pp. 603–608.

[8] G. Meseguer-Brocal, A. Cohen-Hadria, and G. Peeters, "DALI: A large dataset of synchronized audio, lyrics and notes, automatically created using teacher-student machine learning paradigm." in *Proc. of the 19th International Society for Music Information Retrieval Conference, ISMIR*, 2018, pp. 431–437.

[9] D. Bogdanov, A. Porter, H. Schreiber, J. Urbano, and S. Oramas, "The acousticbrainz genre dataset: Multi-source, multi-level, multi-label, and large-scale," in *Proc. of the 20th International Society for Music Information Retrieval Conference, ISMIR*, 2019, pp. 360–367.

[10] I. Vatolkin and C. McKay, "Multi-objective investigation of six feature source types for multi-modal music classification," *Transactions of the International Society for Music Information Retrieval*, vol. 5, no. 1, pp. 1–19, 2022.

[11] R. Mayer and A. Rauber, "Multimodal aspects of music retrieval: Audio, song lyrics - and beyond?" in *Advances in Music Information Retrieval*, Z. W. Ras and A. Wieczorkowska, Eds. Springer, 2010, pp. 333–363.

[12] P. Knees and M. Schedl, "A survey of music similarity and recommendation from music context data," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 10, no. 1, pp. 2:1–2:21, 2013.

[13] D. Jannach, I. Vatolkin, and G. Bonnin, "Music data: beyond the signal level," in *Music Data Analysis: Foundations and Applications*, C. Weihs, D. Jannach, I. Vatolkin, and G. Rudolph, Eds. CRC Press, 2017, pp. 197–215.

[14] F. Simonetta, S. Ntalampiras, and F. Avanzini, "Multimodal music information processing and retrieval: Survey and future challenges," in *Proc. of the International Workshop on Multilayer Music Representation and Processing, MMRP*, 2019, pp. 10–18.

[15] I. Guyon, M. Nikravesh, S. Gunn, and L. A. Zadeh, Eds., *Feature Extraction. Foundations and Applications*, ser. Studies in Fuzziness and Soft Computing. Berlin Heidelberg: Springer, 2006, vol. 207.

[16] I. Fujinaga, "Machine recognition of timbre using steady-state tone of acoustic musical instruments," in *Proc. of the International Computer Music Conference, ICMC*, 1998, pp. 207–210.

[17] R. Fiebrink and I. Fujinaga, "Feature selection pitfalls and music classification," in *Proc. of the 7th International Conference on Music Information Retrieval, ISMIR*, 2006, pp. 340–341.

[18] S. Doraisamy, S. Golzari, N. M. Norowi, M. N. Sulaiman, and N. I. Udzir, "A study on feature selection and classification techniques for automatic genre classification of traditional malay music," in *Proc. of the 9th International Conference on Music Information Retrieval, ISMIR*, J. P. Bello, E. Chew, and D. Turnbull, Eds., 2008, pp. 331–336.

[19] C. N. Silla Jr., A. L. Koerich, and C. A. A. Kaestner, "A feature selection approach for automatic music genre classification," *International Journal of Semantic Computing*, vol. 3, no. 2, pp. 183–208, 2009.

[20] R. Mayer, A. Rauber, P. J. P. de León, C. Pérez-Sancho, and J. M. Iñesta, “Feature selection in a cartesian ensemble of feature subspace classifiers for music categorisation,” in *Proc. of the 3rd International Workshop on Machine Learning and Music, MML*. ACM, 2010, pp. 53–56.

[21] P. Saari, T. Eerola, and O. Lartillot, “Generalizability and simplicity as criteria in feature selection: Application to mood classification in music,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 6, pp. 1802–1812, 2011.

[22] S.-C. Lim, J.-S. Lee, S.-J. Jang, S.-P. Lee, and M. Y. Kim, “Music-genre classification system based on spectro-temporal features and feature selection,” *IEEE Transactions on Consumer Electronics*, vol. 58, no. 4, pp. 1262–1268, 2012.

[23] Y. Huang, S. Lin, H. Wu, and Y. Li, “Music genre classification based on local feature selection using a self-adaptive harmony search algorithm,” *Data Knowledge Engineering*, vol. 92, pp. 60–76, 2014.

[24] C. McKay, J. Cumming, and I. Fujinaga, “jSymbolic 2.2: Extracting features from symbolic music for use in musicological and MIR research,” in *Proc. of the 19th International Society for Music Information Retrieval Conference, ISMIR*, 2018, pp. 348–354.

[25] B. L. Sturm, “A survey of evaluation in music genre recognition,” in *Adaptive Multimedia Retrieval: Semantics, Context, and Adaptation, 10th International Workshop, AMR*, 2012, pp. 29–66.

[26] ——, “Two systems for automatic music genre recognition: What are they really recognizing?” in *Proc. of the 2nd International ACM Workshop on Music Information Retrieval with User-Centered and Mulitmodal Strategies, MIRUM*, 2012, pp. 69–74.

[27] ——, “Classification accuracy is not enough,” *Journal of Intelligent Information Systems*, vol. 41, no. 3, pp. 371–406, 2013.

[28] ——, “Evaluating music emotion recognition: Lessons from music genre recognition?” in *IEEE International Conf. on Multimedia and Expo Workshops, ICMEW*, 2013, pp. 1–6.

[29] I. Vatolkin, M. Preuß, and G. Rudolph, “Multi-objective feature selection in music genre and style recognition tasks,” in *Proc. of the 13th Annual Genetic and Evolutionary Computation Conference, GECCO*, N. Krasnogor and P. L. Lanzi, Eds. ACM Press, 2011, pp. 411–418.

[30] I. Vatolkin, “Exploration of two-objective scenarios on supervised evolutionary feature selection: A survey and a case study (application to music categorisation),” in *Proc. of the 8th International Conference on Evolutionary Multi-Criterion Optimization, EMO*. Springer, 2015, pp. 529–543.

[31] I. Vatolkin, G. Rudolph, and C. Weihs, “Evaluation of album effect for feature selection in music genre recognition,” in *Proc. of the 16th International Society for Music Information Retrieval Conference, ISMIR*, 2015, pp. 169–175.

[32] C. Raffel, “Learning-based methods for comparing sequences, with applications to audio-to-midi alignment and matching,” Ph.D. dissertation, Graduate School of Arts and Sciences, Columbia University, 2016.

[33] I. Vatolkin, P. Ginsel, and G. Rudolph, “Advancements in the music information retrieval framework AMUSE over the last decade,” in *Proc. of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2021, pp. 2383–2389.

[34] C. Weihs, D. Jannach, I. Vatolkin, and G. Rudolph, Eds., *Music Data Analysis: Foundations and Applications*. CRC Press, 2016.

MULTI-PITCH ESTIMATION MEETS MICROPHONE MISMATCH: APPLICABILITY OF DOMAIN ADAPTATION

Franca Bittner **Marcel Gonzalez** **Maike Richter**
Hanna Lukashevich **Jakob Abeßer**
Semantic Music Technologies Group, Fraunhofer IDMT, Ilmenau, Germany
jakob.abesser@idmt.fraunhofer.de

ABSTRACT

The performance of machine learning (ML) models is known to be affected by discrepancies between training (source) and real-world (target) data distributions. This problem is referred to as domain shift and is commonly approached using domain adaptation (DA) methods. As one relevant scenario, automatic piano transcription algorithms in music learning applications potentially suffer from domain shift since pianos are recorded in different acoustic conditions using various devices. Yet, most currently available datasets for piano transcription only cover ideal recording situations with high-quality microphones. Consequently, a transcription model trained on these datasets will face a mismatch between source and target data in real-world scenarios. To address this issue, we employ a recently proposed dataset which includes annotated piano recordings covering typical real-life recording settings for a piano learning application on mobile devices. We first quantify the influence of the domain shift on the performance of a deep learning-based piano multi-pitch estimation (MPE) algorithm. Then, we employ and evaluate four unsupervised DA methods to reduce domain shift. Our results show that the studied MPE model is surprisingly robust to domain shift in microphone mismatch scenarios and the DA methods do not notably improve the transcription performance.

1. INTRODUCTION

Recent advances in Automatic Music Transcription (AMT) enable its practical application in musical education applications where students can record themselves while playing a musical instrument and retrieve a performance feedback in near real-time. The underlying algorithms are driven by deep learning and commonly trained on audio data (source domain), which was gathered in specific and ideal recording setups such as music studios with high-quality microphones [1–3]. In real-life scenarios however, the recording setups may vary from user to user due to different recording devices, room acoustics, and music instrument timbres (target domain). Due to the resulting distribution discrepancy between both domains (domain shift), AMT algorithms might exhibit performance degradation. To overcome this issue, one approach would be to fine-tune pre-trained transcription models using labeled data recorded in real-world settings [4]. This comes with two main drawbacks. First, this procedure would have to be repeated for each user. Second, it requires a lot of effort to obtain perfectly aligned score annotations by manually transcribing audio recordings. For this reason, domain adaptation (DA) methods are used to bridge the gap between different data domains and ensure a good model performance even on previously unseen data. These methods can align the target data distribution to the source data distribution (or vice-versa) [5].

 Attribution: F. Bittner, M. Gonzalez, M. Richter, H. Lukashevich, and J. Abeßer, "Multi-pitch Estimation meets Microphone Mismatch: Applicability of Domain Adaptation", in *Proc. of the 23rd Int. Society for Music Information Retrieval Conf.*, Bengaluru, India, 2022.

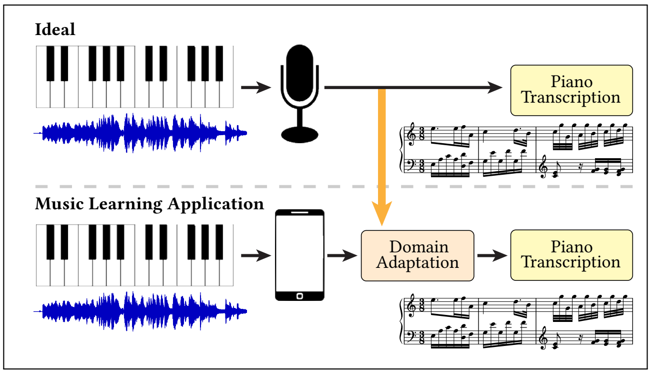

Figure 1. Illustration of the piano transcription process including domain adaptation. Piano music is captured with different recording devices (high quality microphone, mobile devices), adapted to the source domain, and transcribed by an MPE model.

The contributions of the paper are as follows: We study the task of piano multi-pitch estimation (MPE), i. e., the estimation of simultaneously sounding note pitches, from audio recordings captured with different mobile devices. We first analyze and quantify the mismatch of recording devices by comparing their frequency responses. Then, we study the effectiveness of four different DA methods as a pre-processing step to improve MPE algorithms. We do this by investigating to what extent the microphone mismatch impacts the performance of a deep learning-based

MPE model with and without DA. We use the representation shift metric to assess whether the domain shift was reduced by DA.

2. DOMAIN ADAPTATION

2.1 Application in Audio Domain

DA was successfully applied for various tasks in different audio domains. Several acoustic scene classification (ASC) algorithms were proposed during the Detection and Classification of Acoustic Scenes and Events (DCASE) challenge [6,7] to compensate domain shift caused by mismatched recording devices. Furthermore, Yang et al. [8] proposed a two-stage domain adaptation approach to improve the robustness of sound event detection (SED) models by aligning synthetic and real audio data distributions in feature space during training. DA was also applied in Music Information Retrieval (MIR) tasks, such as expressivity analysis in piano recordings [9], representation learning for music processing [10], and instrument activity detection [11], and in automatic speech recognition, DA methods are employed to avoid overfitting of a model which was trained on limited data by transferring knowledge from a source model [12].

2.2 Selected Methods

2.2.1 Zero-mean Unit Variance (ZMUV) Normalization

A common pre-processing step for machine learning (ML) models is the standardization of input features to zero mean and a standard deviation of one. We implemented four variants of the zero-mean unit variance (ZMUV) standardization process from [4] as unsupervised DA methods, which do not require data annotations. The statistics used to standardize the target domain data (mean and standard deviation coefficients) are computed either from the source dataset (*global* variants) or from the target dataset (*adaptive* variants). As a consequence, global variants require access to the source domain data whereas adaptive variants only need fractions of the target domain data.The statistics can be computed either over all available files of the selected domain or individually per file. Furthermore, it is possible to have a finer resolution when computing frequency-wise, i. e., by averaging over all time frames and obtaining coefficients per frequency bin, or patch-wise statistics, i. e., by averaging all frequencies within a patch of 16 time frames. These methods are summarized in Table 1. Finally, in addition to standardizing the target domain features, the source domain data will also be standardized before training the model.

Table 1. Overview of the implemented standardization methods w.r.t. their type (global or adaptive), data scope (whole domain or per file), and resolution (subdivision by frequency bins or patches).

Type	Data Scope	Resolution
Global	Domain (all)	Frequency
Adaptive	Domain (all)	Frequency
Adaptive	File	Frequency
Adaptive	File	Patch

2.2.2 Band-wise Statistics Matching (BWSM)

Band-Wise Statistics Matching (BWSM) is an unsupervised DA method, which involves a band-wise alignment of the first and second statistical moments of the target domain data to the ones of the source domain data [13]. Similarly to standardization, the method is applied on the data level and avoids a re-training of a machine learning model. First, the sample mean and standard deviation values are computed over source and target domain data per frequency bands. Then, a band-wise standardization is applied to the target domain in a similar way as in the adaptive ZMUV normalization per frequency as discussed in the previous section. At the final stage, the adapted features in the target domain $\mathbf{X}_\mathrm{T}$ are aligned to the source domain $\mathbf{X}_\mathrm{S}$, sharing the same means and standard deviations. In contrast to other assessed DA methods, source domain data remains unchanged and the originally trained ML models can be re-used.

2.2.3 Correlation Alignment (CORAL)

Sun et al. [14] proposed CORrelation ALignment (CORAL) as an unsupervised DA method to align the statistical moments of source and target data. After whitening the source domain data (i. e., removing correlation among features), the covariance matrix of the target domain data distribution $\mathbf{C}_\mathrm{T}$ is transferred to the source distribution (recoloring). Then, a new model needs to be trained on the adapted source domain data.

Given that the distribution of target domain data varies with each mobile device, room, and piano type in our given scenario, it is not feasible to train a new MPE model each time one of these parameters changes. In contrast to the original publication, we implemented CORAL such that the target domain data is adapted using the statistics obtained from the source domain data. We follow [14] and use a regularization parameter $\lambda = 1$ for traditional whitening without a singular value decomposition (SVD). The target domain data $\mathbf{D}_\mathrm{T}$ is whitened as in [14] and then re-colored with the source domain covariance matrix $\mathbf{C}_\mathrm{S}$ as $\mathbf{D}_\mathrm{T} \leftarrow \mathbf{D}_\mathrm{T} * \mathbf{C}_\mathrm{T}^{-\frac{1}{2}} * \mathbf{C}_\mathrm{S}^{\frac{1}{2}}$. This way, the source domain data remains unchanged and the same classifier can be used for all target domains. However, Sun et al. [14] observed a lower performance with this approach and supposed that a model trained on adapted source domain data may benefit from the knowledge inherited by the target data distribution. This is not possible if only the target domain data is modified by DA, as the model is trained on the original source domain data independent of the target domain data.

2.2.4 Feature Projection-Based Unsupervised Domain Adaptation (FPDA)

Mezza et al. [15] proposed Feature Projection-based Unsupervised Domain Adaptation (FPDA) to address the mis-