FROM WEST TO EAST: WHO CAN UNDERSTAND THE MUSIC OF THE OTHERS BETTER?

Charilaos Papaioannou[1,2] **Emmanouil Benetos**[2] **Alexandros Potamianos**[1]
[1] School of ECE, National Technical University of Athens, Greece
[2] Centre for Digital Music, Queen Mary University of London, UK
cpapaioan@mail.ntua.gr

ABSTRACT

Recent developments in MIR have led to several benchmark deep learning models whose embeddings can be used for a variety of downstream tasks. At the same time, the vast majority of these models have been trained on Western pop/rock music and related styles. This leads to research questions on whether these models can be used to learn representations for different music cultures and styles, or whether we can build similar music audio embedding models trained on data from different cultures or styles. To that end, we leverage transfer learning methods to derive insights about the similarities between the different music cultures to which the data belongs to. We use two Western music datasets, two traditional/folk datasets coming from eastern Mediterranean cultures, and two datasets belonging to Indian art music. Three deep audio embedding models are trained and transferred across domains, including two CNN-based and a Transformer-based architecture, to perform auto-tagging for each target domain dataset. Experimental results show that competitive performance is achieved in all domains via transfer learning, while the best source dataset varies for each music culture. The implementation and the trained models are both provided in a public repository.

1. INTRODUCTION

As the time passes by, more and more pre-trained models are being made available in the MIR field. These models can be used in a variety of tasks by providing informative deep audio embeddings for music pieces. In correspondence with publicly available datasets, the vast majority of these models are trained on the so called "Western" [1] musical tradition [1]. While studying world, folk, or traditional music, that fact arises two research questions; on the one hand what is the potential of these models when they are being used in the realm of a different culture and on the other hand how capable can a model be when trained on a specific music tradition on providing meaningful audio embeddings.

There are several experimental setups one can employ in order to derive answers to the above questions. By taking into account the importance of the auto-tagging task in the MIR field [2], it becomes clear that transferring knowledge between domain-specific models to perform this task may lead us to valuable insights. Automatic content-based tagging aims to predict the tags of a music piece given its audio signal. The audio signal includes the acoustic characteristics and some of them are responsible for the occurrence of a tag in a piece, forming a multiple instance problem [3].

A variety of models have been proposed to cope with the automatic tagging of music pieces. They can be divided, according to the input data they process, into the ones that utilize time-frequency representations and the others that accept the raw audio signal. In the first category, CNN-based models which are adopted by the computer vision field can be found, such as VGG-ish [4] as well as specifically developed architectures for music, like Musicnn [5]. A Transformer-based architecture was recently proposed in [6] called Audio Spectrogram Transformer (AST). Regarding the models that process audio, the TCNN [7] and the Wave-U-Net [8] architectures are being commonly used. For the purposes of our study, it is essential to use models of the same category with respect to the input they accept and, thus, we selected the ones that process time-frequency representations because of their popularity in the MIR field.

While using deep neural networks, transfer learning of a trained model can lead to a significant performance improvement on the target domain, compared to one that starts from a random state in the parameters space [9]. Typically, the weights of the target domain model are initialized with the ones of a pre-trained model and then fine-tuning is applied. During this step, one has to determine which of the layers will be trainable and which ones will be kept frozen [10]. In general, it is not clear which part of the network should be allowed to be trained in the target task and, thus, experimentation with different setups is necessary. Standard methods include the fine-tuning of the whole network, as suggested in [11], as well as only the last few layers or a part of the network, as in [12]. We

[1] we use the term "Western" to denote music styles which mostly originate from Western cultures, including pop, rock, and Western classical.

experiment with both setups to derive valuable insights on knowledge transfer across domains.

Even though under-represented in general, datasets from specific music cultures are evident in the MIR field and a set of the aforementioned methods have been used to perform several tasks. In [13] a classification of Indian art music was conducted using deep learning models while automatic makam recognition in Turkish music was carried out in [14, 15]. With respect to Western music, there are several research works performing auto-tagging via deep learning models, as in [16] and [17].

In this paper, we incorporate a mosaic of different cultures by including six datasets from Western to Mediterranean and Indian music. Three music audio embedding models, two that mainly consist of convolutional layers and a Transformer-based architecture, are utilized on both single-domain and transfer learning experimental setups for music tagging. Results indicate that any model, despite the music culture that it is trained on, has the potential to adapt to another and achieve competitive results. When comparing the contributions of cross-domain knowledge transfers, we notice that they vary for each music culture and we suggest which one is the best candidate to outperform the single-domain approach. To the authors' knowledge, this is the first study which attempts to explore whether existing music audio embedding models can be used to transfer or learn representations for non-Western cultures. For reproducibility, we share the implementation in a public repository [2] .

2. DATASETS

The selection of the datasets is a prominent theme in the current study and it is constrained by the available corpora that reflect different music cultures. By basing our intuition on the location of each culture, we pursue to include three distinct geographic regions each one represented by two corpora.

Even though spread in several continents, we consider the "West" as a single entity and utilize the MagnaTagATune [18] and FMA-medium [19] datasets that mainly belong to this culture. The second region is the eastern Mediterranean represented by the traditions of Greece and Turkey in our study with Lyra [20] and Turkish-makam [21] datasets. The Indian subcontinent is also incorporated with Hindustani and Carnatic corpora [22], corresponding to the music traditions of the Northern and Southern areas of India respectively.

2.1 MagnaTagATune

MagnaTagATune [18] is a publicly available dataset that is commonly used for the auto-tagging problem in the MIR field. It consists of more than 25,000 audio recordings, summing to 210 hours of audio content at total. Each audio recording is annotated with a subset of the unique 188 tags. Typically, only the top 50 most popular tags are used, which include annotations about genre, instruments and mood. In Table 1, the most frequent tags for MagnaTagATune are presented along with the ones of the other datasets.

2.2 FMA-medium

The Free Music Archive [19] is an open and easily accessible dataset that is used for evaluating several tasks. It contains over 100,000 tracks which are arranged in a hierarchical taxonomy of 161 genres. In order to keep the durations of the datasets balanced whenever possible, and to include genres belonging to Western music styles, we use FMA-medium that consist of 25,000 tracks of 30 seconds each. That means that its total duration is 208 hours, almost equal to the one of MagnaTagATune. With regards to the metadata, we include the top-20 hierarchically related genres of the music pieces.

2.3 Lyra

Lyra [20] is a dataset for Greek traditional and folk music that comprises 1570 pieces and metadata information with regards to instrumentation, geography and genre. Its total duration is 80 hours which makes it the only dataset with duration less than 200 hours in our study. We incorporate the top-30 tags retrieved from columns "genre", "place" and "instruments" to form our multi-label classification setup.

2.4 Turkish-makam

The Turkish makam corpus [21, 23] includes thousands of audio recordings covering more than 2,000 works from hundreds of artists. It is part of CompMusic Corpora [3] [24] which comprises data collections that have been created with the aim of studying particular music traditions. Using Dunya [25] and the related software tool [4] , we were able to get access to 5297 audio recordings, summing in 359 hours, along with their metadata. In order to keep the dataset sizes similar, we set a maximum audio duration equal to 150 seconds which reduced the total length to 215 hours. For the tags, the top-30 most popular with regards to "makam", "usul" and "instruments" information have been included.

2.5 Hindustani

The Hindustani corpus [22] is also part of CompMusic Corpora. It includes 1204 audio recordings, with a total duration of 343 hours, covering a plethora of artists and metadata categories. By setting the maximum audio duration to 780 seconds, the size of the dataset has been decreased to 206 hours for the needs of our study. Furthermore, information about "raga", "tala", "instruments" and "form" has been used to form the labels of each piece. The top-20 most frequent tags have been incorporated to our study as the target of the classification models.

[2] https://github.com/pxaris/ccml

[3] https://compmusic.upf.edu/corpora

[4] https://github.com/MTG/pycompmusic

MagnaTagATune		FMA-medium		Lyra		Turkish-makam		Hindustani		Carnatic	
guitar	18.76%	Rock	28.41%	Voice	76.21%	Voice	63.33%	Voice	83.90%	Voice	82.35%
classical	16.52%	Electronic	25.26%	Traditional	76.05%	Kanun	31.09%	Tabla	53.03%	Violin	78.45%
slow	13.71%	Punk	13.28%	Violin	57.34%	Tanbur	27.93%	Khayal	41.33%	Mridangam	75.65%
techno	11.42%	Experimental	9.00%	Percussion	53.71%	Ney	27.56%	Harmonium	39.25%	Kriti	70.87%
strings	10.55%	Hip-Hop	8.80%	Laouto	51.69%	orchestra	26.38%	Teentaal	35.35%	adi	51.88%
drums	10.05%	Folk	6.08%	Guitar	37.34%	Oud	24.36%	Tambura	27.88%	Ghatam	30.32%
electronic	9.74%	Garage	5.67%	Klarino	31.05%	kemence	22.79%	Ektaal	21.58%	Khanjira	17.65%
rock	9.17%	Instrumental	5.40%	Nisiotiko	26.85%	Cello	17.83%	Pakhavaj	7.88%	rupaka	11.98%
fast	8.92%	Indie-Rock	5.17%	place-None	25.16%	Violin	17.62%	Sarangi	7.30%	mishra chapu	7.27%
piano	7.95%	Pop	4.74%	Bass	24.76%	Hicaz	10.63%	Dhrupad	7.05%	Tana Varnam	5.21%

Table 1. Relative frequencies of the top 10 most popular tags in each dataset.

2.6 Carnatic

The Carnatic corpus [22] comprises 2612 audio recordings, summing in more than 500 hours of content. As with the previous datasets, by setting a maximum duration cut equal to 330 seconds, the total duration has been decreased to 218 hours. Identical to Hindustani, the top-20 most popular annotations regarding "raga", "tala", "instruments" and "form" have been included for the metadata.

3. METHOD

In this section, the models which are used for the purposes of this study are first presented. We, then, describe how transfer learning is utilized to infer similarities between the music cultures by employing knowledge from the domain adaptation field.

3.1 Models

3.1.1 VGG-ish

All of our models use the mel-spectrogram as their input, a commonly used feature for MIR tasks such as automatic tagging [26]. This selection enables the utilization of CNN-based architectures which have been successfully used in computer vision tasks. The Visual Geometry Group (VGG) network [27] and its variants consist of a stack of convolutional layers followed by fully connected layers.

We use a VGG-ish architecture, similar to the one implemented by the authors in [28], that is a 7-layer CNN, with 3×3 convolution filters and 2×2 max-pooling, followed by two fully-connected layers. It accepts mel-spectrograms that correspond to short chunks of audio as its input, with duration equal to 3.69 seconds.

3.1.2 Musicnn

Musicnn [17] is a music inspired model that uses convolutional layers at its core. Its first convolutional layer consists of vertical and horizontal filters in order to capture timbral and temporal features respectively. These features are, then, concatenated and fed to 1D convolutional layers followed by a pair of dense layers that summarize them and predict the relevant tags. Similar to VGG-ish, it uses mel spectrograms from short audio chunks at its input with duration 3 seconds.

3.1.3 Audio Spectrogram Transformer

As its name indicates, Audio Spectrogram Transformer (AST) is a purely attention-based model for audio classification [6]. Based on the Transformer architecture [29], AST splits the input mel-spectrogram to 16×16 patches in both time and frequency dimensions that are, in turn, flattened to 1D embeddings of size 768 using a linear projection layer. A trainable positional embedding is also added to each patch embedding so that the model will capture the spatial structure of the input 2D spectrogram. The resulting sequence is fed to the Transformer, where only the encoder is utilized since AST is designed for classification tasks. The output of the encoder is followed by a linear layer that predicts the labels. As the authors that introduced the architecture suggest, we set a specific cut to the input length of the AST model that is equal to 8 seconds in all our experiments.

3.2 Transfer Learning

The purpose of transfer learning is to improve the performance of the models on target domains by transferring knowledge from different but related source domains [30]. In the field of MIR, both transferring feature representations to the target domain from a pre-trained model on a source task [31] as well as learning shared latent representations across domains [32] have been proposed in the past. Yet, these methods have not been applied to non-Western music datasets neither by adapting an existing model to them nor by studying to what end these cultures can be valuable source domains for widely developed models, two aspects which are both studied in this work.

According to the categorization conducted by the authors in [33], these methods belong to *parameter sharing* category of the model-based transfer learning techniques. In the deep learning realm, it is, thus, common to use a trained network for a source task, share its parameters and in turn fine-tune some or all layers to produce a target network. While following this method, one expects it to lead to better results when the participating domains are similar

Model	VGG-ish		Musicnn		AST	
Metric / Dataset	ROC-AUC	PR-AUC	ROC-AUC	PR-AUC	ROC-AUC	PR-AUC
MagnaTagATune	0.9123	0.4582	0.9019	0.4333	**0.9172**	**0.4654**
FMA-medium	**0.8889**	0.4949	0.8766	0.4473	0.8886	**0.5024**
Lyra	0.8097	0.4806	0.7391	0.4042	**0.8476**	**0.5333**
Turkish-makam	**0.8696**	0.5639	0.8505	0.5299	0.8643	**0.5669**
Hindustani	**0.8477**	**0.6082**	0.8471	0.6016	0.8307	0.5786
Carnatic	0.7392	0.4278	0.7496	0.4182	**0.7706**	**0.4394**

Table 2. ROC-AUC and PR-AUC scores of the models on single domain auto-tagging tasks.

to each other. Indeed, by studying the prior work on domain adaptation, one will find that the main strategy consists of minimizing the difference between the source and target feature distributions, when transferring representations from a labeled dataset to a target domain where labeled data is sparse or non-existent [34, 35].

By adapting the above rationale to our study, where the participating domains are all rich in labeled data, we expect that when applying transfer learning by parameter sharing, the more the similarity between the participating domains the better the performance of the target domain on its supervised learning task.

In order to study to what end this hypothesis stands in computational musicology with deep neural networks, we utilize the previously presented models which are widely used in the MIR field and consist of different cores, namely convolutional layers (VGG-ish and Musicnn) and a Transformer module (AST). Having the models trained on each single dataset, we apply all the cross-domain knowledge transfers for each architecture by fine-tuning only the output layer as well as the whole network. We then aggregate the results across the models seeking to derive insights with regards to the similarities between the domains as well as specifying which source is the best candidate for each target dataset.

4. EXPERIMENTS

As already mentioned, we use mel spectrograms as the input of all our models. In order to convert the audio recordings of the datasets to this representation, we use Librosa [36] to re-sample them to 16 kHz sample rate. Then, 512-point FFT with a 50% overlap is applied, the maximum frequency is set to 8 kHz and number of Mel bands to 128. Our intention, in this study, is not the optimization of the performance of the single-domain tasks but rather studying the knowledge transfer across the domains. So, we keep our training setup as close as possible to the literature, at each single domain task, in order to have a sanity check for the implementation.

For VGG-ish and Musicnn models, we use a mixture of scheduled Adam [37] and stochastic gradient descent (SGD) for the optimization method, identical to what the authors at [28] have used. The batch size is set to 16 and the learning rate to $1e-4$ for both models while the maximum number of epochs are 200 for VGG-ish and 50 for Musicnn. With regards to the AST model, we follow the setup proposed in [6], namely batch size 12, Adam optimizer, learning rate scheduling that begins from $1e-5$ and is decreased by a factor of 0.85 every epoch after the 5th one as well as pre-trained on Imagenet Transformer weights.

All our models accept a fixed size audio chunk at their input but need to predict song-level tags. During the evaluation phase, we aggregate the tag scores across all chunks by averaging them to acquire the label scores for the whole audio. We use the area under receiver operating characteristic curve (ROC-AUC), a widely used evaluation metric on multi-label classification problems and the area under precision-recall curve (PR-AUC), a suitable metric for unbalanced datasets [38].

During transfer learning, we initialize all parameters of the target model, except for the output layer, from each source dataset and (i) allow only the output layer to be trained and (ii) train the whole network. In both settings, we use the same hyper-parameters and evaluation procedure with the single-domain setups across all datasets for each model architecture.

5. RESULTS

The performance of the three models on all single-domain tasks can be seen in Table 2. The performance of the Musicnn and VGG-ish models on MagnaTagATune is similar to the reported metrics in [28], which indicates the validity of our implementation. In general, the AST model shows the best performance followed by VGG-ish and then Musicnn. This result should not be taken into account solidly, because no hyper-parameter tuning has been taken place for each domain and in order to keep the duration of the training to less than 24 hours for each task, the number of epochs for Musicnn was significantly less than VGG-ish. On the other hand, one should consider that the AST [6] and VGG-ish [28] models may, indeed, perform better for limited time resources.

In Table 3, one can see the ROC-AUC scores in all single-domain and cross-domain setups. The rows are the source datasets while the columns are the target datasets. A sub-table is constructed for each model architecture and for a transfer from domain *A* to *B*, the result of the fine-tuning of only the output layer ('output') as well as all the layers ('all') are reported. The single-domain setup is

Target domain	**MagnaTagATune**		**FMA-medium**		**Lyra**		**Turkish-makam**		**Hindustani**		**Carnatic**	
trainable layer(s) / Source domain	output	all	output	all	output	all	output	all	output	all	output	all
VGG-ish												
MagnaTagATune	-	91.23	88.11	**92.39**	74.69	**85.40**	76.79	86.84	76.09	85.04	67.19	**74.71**
FMA-medium	85.82	**91.29**	-	88.89	68.56	84.04	75.40	**87.78**	75.77	84.39	67.03	74.56
Lyra	84.34	90.93	82.84	92.10	-	80.97	76.98	87.21	77.41	84.24	67.30	73.52
Turkish-makam	85.19	90.90	84.41	91.74	70.93	82.38	-	86.96	77.54	**85.32**	67.16	73.50
Hindustani	84.24	91.02	83.83	91.91	66.27	79.71	77.25	87.63	-	84.77	66.72	74.63
Carnatic	84.18	91.00	82.62	91.73	61.59	76.72	77.07	87.40	78.19	84.81	-	73.92
Musicnn												
MagnaTagATune	-	90.19	87.34	**91.03**	71.79	78.74	74.72	**85.96**	75.87	84.18	66.12	75.57
FMA-medium	85.52	**90.35**	-	87.66	65.94	77.59	75.51	85.13	73.16	**85.49**	66.38	75.77
Lyra	81.38	90.03	82.23	90.80	-	73.91	74.11	85.20	78.10	83.29	65.09	75.51
Turkish-makam	84.35	90.11	83.79	90.81	61.87	**79.83**	-	85.05	75.67	83.75	67.49	74.09
Hindustani	82.38	89.86	83.42	90.85	64.48	78.95	74.60	85.58	-	84.71	65.25	**76.95**
Carnatic	83.02	90.05	82.78	90.74	61.83	77.92	75.09	85.43	75.34	84.19	-	74.96
AST												
MagnaTagATune	-	**91.72**	89.25	91.99	75.68	83.77	76.28	87.20	74.67	**86.57**	66.03	75.43
FMA-medium	88.63	91.62	-	88.86	65.72	82.17	76.37	**87.43**	74.51	85.76	67.33	75.98
Lyra	87.49	91.44	87.44	**92.43**	-	**84.76**	77.08	86.80	72.24	83.73	68.47	76.59
Turkish-makam	87.33	91.40	86.31	91.95	72.70	77.95	-	86.43	70.13	83.56	67.10	75.23
Hindustani	87.40	91.35	87.11	92.26	71.74	84.60	75.70	86.90	-	83.07	67.75	75.85
Carnatic	87.42	91.45	86.83	91.75	63.33	81.44	76.87	87.14	74.11	82.91	-	**77.06**

Table 3. ROC-AUC scores (%) when applying transfer learning using the models VGG-ish, Musicnn and Audio Spectrogram Transformer. Rows are the source domains and columns the target domains. After initializing the network with the parameters of the trained (at the source dataset) model, fine-tuning on the output layer as well as on the whole network is applied. The diagonal values (under the "all" columns) correspond to the respective single-domain models (no transfer learning) where the experimentation with only the output layer trainable has no meaning.

when source and target is the same dataset and, thus, only training of the whole network has meaning. The table is better parsed column-wise, e.g., by inspecting the results of VGG-ish model on MagnaTagATune when transferring knowledge from the other domains at the upper-left pair of columns in the table.

In order to aggregate all the cross-domain knowledge transfers, we follow the subsequent procedure: for each target task that consists of a specific model, target dataset and fine-tuning method, min-max normalization is applied to the $N-1$ transfer learning results, where N is the number of all datasets. The previous step leads to the construction of $M \times F$ matrices, M the number of the models and F the number of fine-tuning methods, where rows are the source domains, columns the target domains and diagonal elements are empty. Each cell has a value in the range $[0, 1]$, as a result of the normalization step, while the value 1 corresponds to the knowledge transfer that led to the best performance in the target domain. By calculating the element-wise mean of the produced $M \times F$ matrices, we reach to the result that can be seen in Figure 2.

6. DISCUSSION

The results indicate that knowledge transfer both from Western to non-Western cultures and the opposite can be beneficial when deep learning models are used to perform automatic music tagging. Indeed, by inspecting Table 3, the general take-home message one should acquire is that regardless of the model architecture, all datasets have the potential to contribute as a source to a target domain by providing their deep audio embeddings. To investigate how valuable knowledge transfers from widely used datasets to non-Western music cultures can be, we focus on the last four datasets, i.e., the last eight columns of the table, and parse the two first rows, corresponding to MagnaTagATune and FMA datasets, at each model architecture. For instance, we notice that for Lyra, when Musicnn is used and fine-tuning only of the output layer is applied, the model coming from MagnaTagATune has the greater ROC-AUC score, namely 71.79%. Additionally, the AST model trained on the FMA-medium dataset, outperforms the others when totally fine-tuned to the Turkish-makam dataset, scoring 87.43%.

In order to study the inverse transfer direction, we center our interest to the first four columns of the entire table. Even though MagnaTagATune and FMA are almost always the best source for each other, the deep audio embeddings provided by the other datasets achieve competitive performance. For example, when MagnaTagATune is the target domain and fine-tuning is restricted to the output layer of the network, we observe that transferring from Turkish-

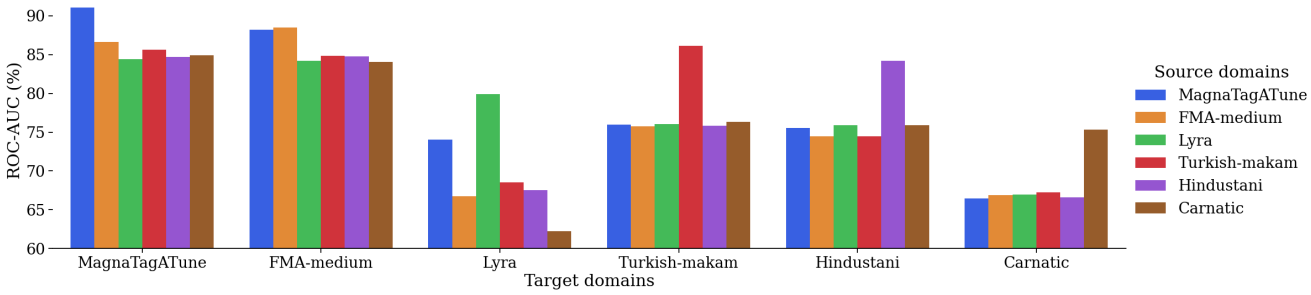

Figure 1. Average, over the three models, ROC-AUC scores of all cross-domain transfers when fine-tuning of the output layer is applied. The highest bar at each group corresponds to the respective single-domain model.

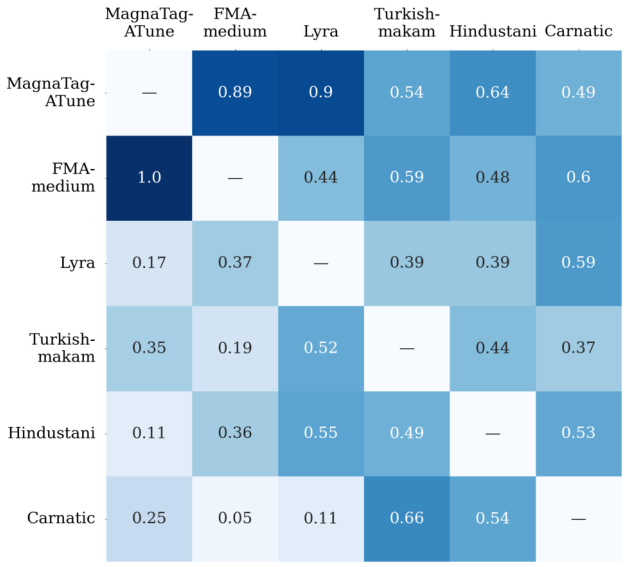

Figure 2. Cross-cultural music transfer learning results. Rows correspond to the source datasets and columns to the target datasets. The value of each cell (knowledge transfer) is normalized and averaged across all models and fine-tuning methods.

makam leads to a performance that is comparable to the best source (FMA-medium) for all models.

By considering all cross-domain knowledge transfers, one can specify the best candidate to provide a trained model, with a specific architecture, for each target dataset. We, thus, notice that the model that is transferred from Hindustani outperforms the others at the Carnatic dataset, when fine-tuning on the whole Musicnn architecture is applied. A holistic picture of the cross-cultural music transfer learning is depicted in Figures 1 and 2.

In Fig. 1 the scores of all cross-domain transfers when fine-tuning the output layer, can be seen, averaged across the three models. The uniformity of the performances of different sources at each target dataset can be examined. We, thus, recognize that the most unbalanced performances are spotted on the Lyra target domain, a result that is probably related to the smaller size of this dataset compared to the others. By exploring Fig. 2 in a column-wise fashion, we observe that for MagnaTagATune as the target domain, FMA-medium is the best source with a value equal to 1. This means that in all transfer learning setups, this source performed better than the others in this domain.

Both figures show that MagnaTagATune and FMA-medium perform consistently well across the domains, something that possibly indicates their appropriateness for the auto-tagging task. However, as we move to the Eastern cultures, we notice that their contribution is somehow decreased and other domains tend to contribute similarly or even more in those targets. The values at Fig. 2 should not be considered solidly as similarity metrics between the domains because other factors may also affect the results we notice. It is, although, a first step towards studying different music cultures using deep learning methods.

7. CONCLUSIONS

In this paper, the transferrability of music cultures by utilizing deep audio embedding models is studied. To that end, six datasets and three models were employed while experimentation with two fine-tuning methods took place. The automatic tagging of music pieces served as the supervised learning task where all cross-domain knowledge transfers were applied and evaluated.

The results show that state-of-the-art models can benefit from knowledge transfer not only from Western to non-Western cultures but also the opposite too. By aggregating the scores across all models and fine-tuning methods, the suitability of each source domain for a target task was calculated and, thus, which domain can be the best candidate to transfer knowledge from for each dataset was proposed. Based on the literature, we suggest that this result can be interpreted to a degree as a similarity metric between the music cultures.

We identify that the current study has limitations. In the future, the semantic similarities between the labels of the involved domains will be examined. More datasets and models, like those that process raw audio signals, will be considered as well as semi-supervised and unsupervised learning techniques. Other tasks may be employed such as mode estimation, assuming that key in Western cultures functions in a similar way with makam or raga in other cultures. All datasets can also be utilized to learn music embeddings in order to unveil cross-cultural links between acoustic features and tags.

8. ACKNOWLEDGEMENTS

The authors would like to thank Sertan Şentürk, Alastair Porter and the Universitat Pompeu Fabra for their willingness to provide us with the data without which this study would not have been possible. We would like to also thank Charalampos Saitis and the reviewers for their valuable and constructive comments that helped us improve our work.

9. REFERENCES

[1] E. Gómez, P. Herrera, and F. Gómez-Martin, "Computational Ethnomusicology: perspectives and challenges," *Journal of New Music Research*, vol. 42, no. 2, pp. 111–112, June 2013.

[2] K. Choi, "Deep Neural Networks for Music Tagging," Ph.D. dissertation, Queen Mary University of London, September 2018.

[3] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial Intelligence*, vol. 89, no. 1, pp. 31–71, January 1997.

[4] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, "CNN architectures for large-scale audio classification," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017, pp. 131–135.

[5] J. Pons and X. Serra, "musicnn: Pre-trained convolutional neural networks for music audio tagging," *arXiv preprint arXiv:1909.06654*, September 2019.

[6] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio Spectrogram Transformer," *arXiv preprint arXiv:2104.01778*, July 2021.

[7] A. Pandey and D. Wang, "TCNN: Temporal Convolutional Neural Network for Real-time Speech Enhancement in the Time Domain," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 6875–6879.

[8] D. Stoller, S. Ewert, and S. Dixon, "Wave-U-Net: A Multi-Scale Neural Network for End-to-End Audio Source Separation," *arXiv preprint arXiv:1806.03185*, June 2018.

[9] Z. Yang, R. Salakhutdinov, and W. W. Cohen, "Transfer Learning for Sequence Tagging with Hierarchical Recurrent Networks," *arXiv preprint arXiv:1703.06345*, March 2017.

[10] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" *arXiv preprint arXiv:1411.1792*, November 2014.

[11] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *arXiv preprint arXiv:1311.2524*, October 2014.

[12] M. Long, Y. Cao, J. Wang, and M. I. Jordan, "Learning Transferable Features with Deep Adaptation Networks," *arXiv preprint arXiv:1502.02791*, May 2015.

[13] A. K. Sharma, G. Aggarwal, S. Bhardwaj, P. Chakrabarti, T. Chakrabarti, J. H. Abawajy, S. Bhattacharyya, R. Mishra, A. Das, and H. Mahdin, "Classification of Indian Classical Music With Time-Series Matching Deep Learning Approach," *IEEE Access*, pp. 102 041–102 052, 2021.

[14] E. Demirel, B. Bozkurt, and X. Serra, "Automatic makam recognition using chroma features," in *Proceedings of the 8th International Workshop on Folk Music Analysis; Thessaloniki, Greece, p. 19-24*, 2018.

[15] K. K. Ganguli, S. Şentürk, and C. Guedes, "Critiquing task-versus goal-oriented approaches: A case for makam recognition," in *Proceedings of the 23rd Int. Society for Music Information Retrieval Conf., Bengaluru, India*, December 2022.

[16] K. Choi, G. Fazekas, and M. Sandler, "Automatic tagging using deep convolutional neural networks," *arXiv preprint arXiv:1606.00298*, June 2016.

[17] J. Pons, O. Nieto, M. Prockup, E. Schmidt, A. Ehmann, and X. Serra, "End-to-end learning for music audio tagging at scale," *arXiv preprint arXiv:1711.02520*, June 2018.

[18] E. Law, K. West, M. Mandel, M. Bay, and J. S. Downie, "Evaluation of algorithms using games: The case of music tagging," in *10th International Society for Music Information Retrieval Conference, ISMIR 2009*, 2009, pp. 387–392.

[19] M. Defferrard, K. Benzi, P. Vandergheynst, and X. Bresson, "FMA: A Dataset For Music Analysis," *arXiv preprint arXiv:1612.01840*, September 2017.

[20] C. Papaioannou, I. Valiantzas, T. Giannakopoulos, M. Kaliakatsos-Papakostas, and A. Potamianos, "A Dataset for Greek Traditional and Folk Music: Lyra," in *Proceedings of the 23rd Int. Society for Music Information Retrieval Conf., Bengaluru, India*, December 2022.

[21] B. Uyar, H. S. Atli, S. Şentürk, B. Bozkurt, and X. Serra, "A corpus for computational research of turkish makam music," in *Proceedings of the 1st International Workshop on Digital Libraries for Musicology*, 2014, pp. 1–7.

[22] A. Srinivasamurthy, G. K. Koduri, S. Gulati, V. Ishwar, and X. Serra, "Corpora for music information research in indian art music," in *Proceedings of the 2014 International Computer Music Conference, ICMC/SMC; 2014 Sept 14-20; Athens, Greece*, 2014.

[23] S. Şentürk, "Computational analysis of audio recordings and music scores for the description and discovery of ottoman-turkish makam music," Ph.D. dissertation, Universitat Pompeu Fabra, 2016.

[24] X. Serra, "Creating research corpora for the computational study of music: the case of the compmusic project," in *Audio engineering society conference: 53rd international conference: Semantic audio*, 2014.

[25] A. Porter, M. Sordo, and X. Serra, "Dunya: A system for browsing audio music collections exploiting cultural context," in *Proceedings of the 14th Int. Society for Music Information Retrieval Conf., Curitiba, Brazil*, 2013.

[26] S. Dieleman and B. Schrauwen, "Multiscale approaches to music audio feature learning," in *14th International Society for Music Information Retrieval Conference (ISMIR-2013)*, 2013, pp. 116–121.

[27] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv preprint arXiv:1409.1556*, April 2015.

[28] M. Won, A. Ferraro, D. Bogdanov, and X. Serra, "Evaluation of CNN-based Automatic Music Tagging Models," *arXiv preprint arXiv:2006.00751*, June 2020.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Kaiser, and I. Polosukhin, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, 2017.

[30] S. J. Pan and Q. Yang, "A Survey on Transfer Learning," *IEEE Transactions on Knowledge and Data Engineering*, no. 10, pp. 1345–1359, October 2010.

[31] A. van den Oord, S. Dieleman, and B. Schrauwen, "Transfer learning by supervised pre-training for audio-based music classification," in *Conference of the International Society for Music Information Retrieval, Proceedings*, 2014.

[32] P. Hamel, M. E. P. Davies, K. Yoshii, and M. Goto, "Transfer Learning In MIR: Sharing Learned Latent Representations For Music Audio Classification And Similarity," in *14th International Conference on Music Information Retrieval (ISMIR '13)*, 2013.

[33] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A Comprehensive Survey on Transfer Learning," *arXiv preprint arXiv:1911.02685*, June 2020.

[34] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko, "Simultaneous Deep Transfer Across Domains and Tasks," *arXiv preprint arXiv:1510.02192*, October 2015.

[35] K. Saenko, B. Kulis, M. Fritz, and T. Darrell, "Adapting Visual Category Models to New Domains," in *Computer Vision – ECCV 2010*. Springer, 2010, pp. 213–226.

[36] 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*, vol. 8, 2015, pp. 18–25.

[37] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv preprint arXiv:1412.6980*, January 2017.

[38] J. Davis and M. Goadrich, "The relationship between precision-recall and roc curves," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 233–240.