The key information is extracted from the dataset metadata. When a song has more than one key signatures reported, we consider a single key signature extracted from the music21 key detection algorithm [15]. At the inference level of our model, we apply the latter method on the test score input. We do not apply any type of post-processing on the extracted melody. Also, we automatically merge all notes into a single MIDI part per piece, keeping the information of a note being a melody one as ground truth.

To facilitate a fair comparison of the models in our experiments, we use the same dataset subset and train-validation-test set split as in [13], which results to 865 songs. The steps above give us the amount of 1,408,056 notes, from which 19.8% are melody notes.

3. MODEL DETAILS

In this section, we present the details of our model architecture (3.1), the details about our training setup (3.2), as well as the metrics we have established (3.3). We have released the code of the model as open-source [1] .

3.1 Architecture

We implement a deep bidirectional LSTM architecture (biLSTM) [16] for our model, which is a type of Recurrent Neural Network (RNN). In our implementation, the biLSTM model has been setup after implementing a grid-search on the model hyper-parameters, using hyperopt [17]. After this process, the model results in having 6 layers, with hidden size of 140, followed by a forward layer with output size of 20. The Adam optimizer is used with a start learning rate of 0.001. We identified the set of hyper-parameters that produce the highest melody F-measure score (metric described in the following Section 3.3).

One characteristic of our dataset is the imbalance of the data labels (i.e. the two classification classes), meaning that non-melody notes outnumber the melody ones. To overcome this issue in our classification task, we adopt the focal loss [18] as the loss function for the model, which is defined as:

$$FL(p_t) = -a_t(1-p_t)^{\gamma} log(p_t), \tag{1}$$

where p_t denotes the model's estimated probability for an input to be classified to class t and $a_t \in [0, 1]$ is a weighting factor for the imbalanced classes which balances the importance of positive and negative examples. The term $(1-p_t)^{\gamma}$ acts as a modulating factor with γ controlling the rate at which over-weighted examples are down-weighted. We set $a_t = 0.25$ and $\gamma = 2$.

3.2 Training setup

The train-validation-test sets that we used from [13] contain a data percentage of 80-10-10%, respectively. The features have been scaled, given the data points in the train and the validation sets.

[1] https://github.com/bytedance/midi_melody_extraction

3.3 Metrics

The metrics that we have computed for this task are the following: `accuracy` indicating the overall accuracy of the predicted notes in percentage, as well as the `mel_P`, `mel_R` and `mel_F` for melody notes precision, recall and F measure values, respectively:

$$\texttt{mel\_P} = \frac{|\text{Correctly predicted melody notes}|}{|\text{Notes predicted as melody}|} \tag{2}$$

$$\texttt{mel\_R} = \frac{|\text{Correctly predicted melody notes}|}{|\text{Melody notes}|} \tag{3}$$

$$\texttt{mel\_F} = 2 * (\frac{\texttt{mel\_P} * \texttt{mel\_R}}{\texttt{mel\_P} + \texttt{mel\_R}}) \tag{4}$$

Also, we include the metric `Voice False Alarm` (VFA) from mir_eval [19], which is defined as the number of notes predicted as melody notes, although they are not, divided by the number of non-melody notes.

4. EXPERIMENTS

To evaluate the performance of our system, we have prepared a list of separate model setups, given the same train, validation and test sets. Also, we were interested in whether two types of data augmentation techniques would improve the accuracy of LStoM when applied in isolation as well as together. The first type is to shift the score key by altering the note pitches, where each piece has been transposed by n semitones, for n in $\{-6, -5, ..., 4, 5\}$ (the model is tagged as "LStoM PSaugm" standing for pitch shifting augmentation). The second type is to shift the melody notes one octave lower than the original one (the model is tagged as "LStoM MOaugm" standing for melody octave augmentation). Both techniques are common in the literature for the task of symbolic melody extraction [10].

4.1 Comparison with the state of the art

We investigate how other models that report state of the art results in various datasets perform in the case of the specific train, validation and test sets. To this end, we were able to train two models. One is following the default settings of [11] tailored to the task of the melody identification, using the augmentation techniques that are provided from this work and the pitch proximity method as a segment merging mode among the created note clusters (the model is tagged as "Hsiao-Su"). The other is following the default settings of [10], again using the augmentation techniques that are provided from this work (the model is tagged as "Lu-Su").

The skyline algorithm [8], which is essentially picking the highest pitch at a given onset time, has been considered as our baseline. Also, we consider a skyline variation where we keep the duration of the note with the highest pitch, ignoring the lower-pitch notes that start after its onset and before its offset. Both original and variation skyline output are illustrated in Figure 1, using the input example as reported in [8].

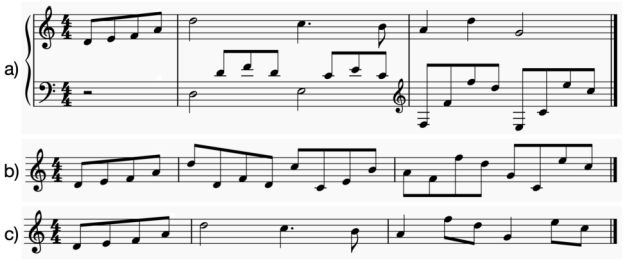

Figure 1. Illustration of our baselines. a) Input example, b) Output of skyline algorithm, c) Output of skyline variation algorithm.

Model	acc (%)	mel_F	mel_P	mel_R	VFA
LStoM*	92.3	0.816	0.778	0.872	0.065
LStoM* PSaugm	91.6	0.788	0.791	0.799	0.054
LStoM* MOaugm	92.3	0.800	0.815	0.811	0.066
LStoM* MO+PSaugm	91.9	0.798	0.797	0.811	0.054
MIDIBERT* [13]	**97.1**	**0.930**	**0.912**	**0.952**	**0.024**
LStoM	90.7	**0.774**	0.751	**0.814**	0.070
MIDIBERT [13]	**91.9**	0.772	**0.841**	0.727	**0.035**
skyline	63.1	0.499	0.353	0.881	0.435
skyline-variation	78.7	0.569	0.485	0.697	0.192
Lu-Su [10]	66.6	0.614	0.600	0.638	0.299
Hsiao-Su [11]	82.5	0.650	0.544	0.821	0.176

Table 2. Basic comparison results among testing models. An '*' is added to indicate that the MIDI files in the test set have been pre-processed.

Both LStoM and MIDIBERT [2] include pre-processing steps for note quantisation and alignment in downbeat level when preparing the MIDI files for training. At the testing stage, we considered as a fair comparison to first replicate the results of [13] where the MIDI files of the test set have been pre-processed, therefore our pre-processing steps were applied to the test set as well for LStoM and its augmentations (metrics reported at the top part of Table 2). The augmentations do not appear to improve the overall performance of LStoM. In the bottom part of Table 2, we report the metrics of the remaining comparison, where the notes of the test set have not been quantised nor aligned to the downbeat. Interestingly, MIDIBERT outperforms in the first scenario, while the results are mixed in the second one.

Lastly, LStoM is significantly lighter than MIDIBERT; the number of parameters are 875,462 compared to 111,298,052, respectively.

In the next two subsections, we are exploring LStoM in terms of how important the selected features for the training process are (Section 4.2) and we are discussing some observations from the models' outcome, respectively (Section 4.3).

[2] Function `align_midi_beats` in `https://github.com/wazenmai/MIDI-BERT/blob/CP/data_creation/preprocess_pop909/preprocess.py`, accessed 31 August 2022

4.2 Importance of features

We were interested to know how relevant the selected features are for our task and whether any of them are more important in a sense that they contain an amount of information that is crucial for such systems. Algorithmically it is feasible to examine and improve the interpretability of a predictive model to a degree, and to identify a type of ranking of importance for the input features. However, most recent feature ranking algorithms assume feature independence (for more examples and details we refer to [20]), an assumption which is not safe in our case.

Ranking	Feature	Ranking	Feature
1	`pitch`	4	`dur`
2	`pitch_distance_above`	5	`pos_in_bar`
3	`pitch_distance_below`	6	`pitch_in_scale`

Table 3. The ranking of feature importance obtained by RELIEF algorithm.

One algorithm that is not based on this assumption is RELIEF [21] which elaborates on a simple comparison idea whereby feature value differences between nearest neighbour instance pairs are identified. If a feature value difference is observed in a neighbouring instance pair with the same class, the feature ranking score decreases. However, the score increases when a feature value difference is observed in a neighbouring instance pair with different class values. We applied RELIEF to our dataset and the resulting feature ranking is reported in Table 3.

Not by surprise, the pitch has been ranked as the most important feature, followed by the one that computes how much is the distance from the pitch above. The third feature in the ranking list is the similar one, computing the pitch distance from the pitch below. The note duration comes to the fourth place, followed by the metrical-related feature of computing the position of the note within the score bar, and finally whether the note pitch is in scale.

In order to evaluate the incremental contribution of each of the features studied, we have considered feature subsets by incrementally adding features to the training set one at a time, starting with the most important one, i.e. the note pitch, and continuing to add features according to their rank order. The results obtained are reported in Table 4.

Model	acc	mel_F	mel_P	mel_R	VFA
LStoM1	86.5%	0.672	0.652	0.708	0.094
LStoM1-2	88.4%	0.718	0.697	0.752	0.082
LStoM1-3	89.4%	0.738	0.723	0.768	0.073
LStoM1-4	90.3%	0.742	0.780	0.724	0.052
LStoM1-5	92.0%	0.797	0.803	0.806	0.051
LStoM1-6	92.3%	0.816	0.778	0.872	0.065

Table 4. Basic comparison results among the variations of LStoM model, where LStoM*x-x* indicates the range of ranked features used at the training process.

It is worth noticing that the highest ranked feature contains on its own substantial predictive power: the melody F-measure score for the model induced with only pitch information alone is 0.672, which is slighter higher than the

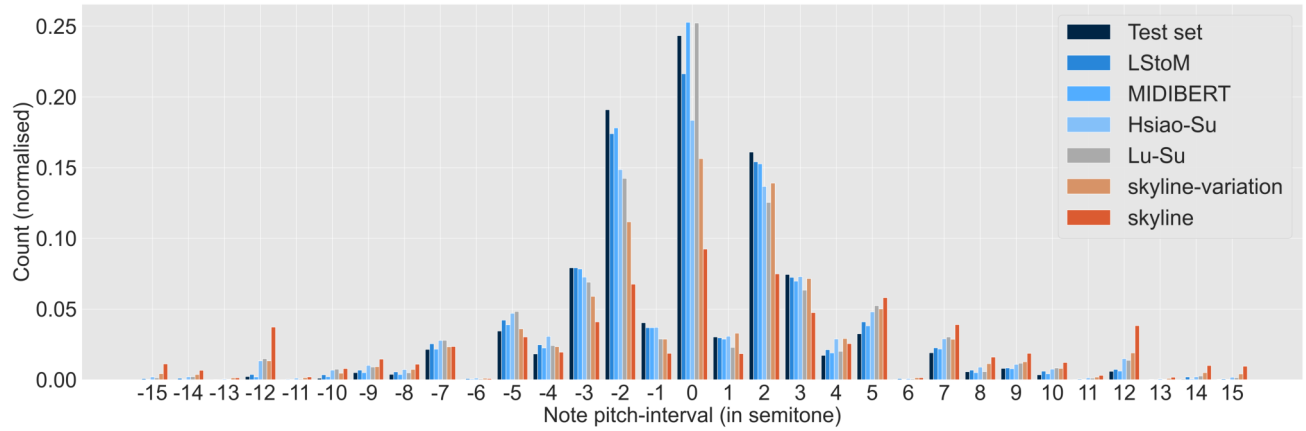

Figure 2. Pitch interval distribution among melodies predicted by LStoM, MIDIBERT, "Hsiao-Su", "Lu-Su", skyline and skyline-variation compared to the melodies from the test set.

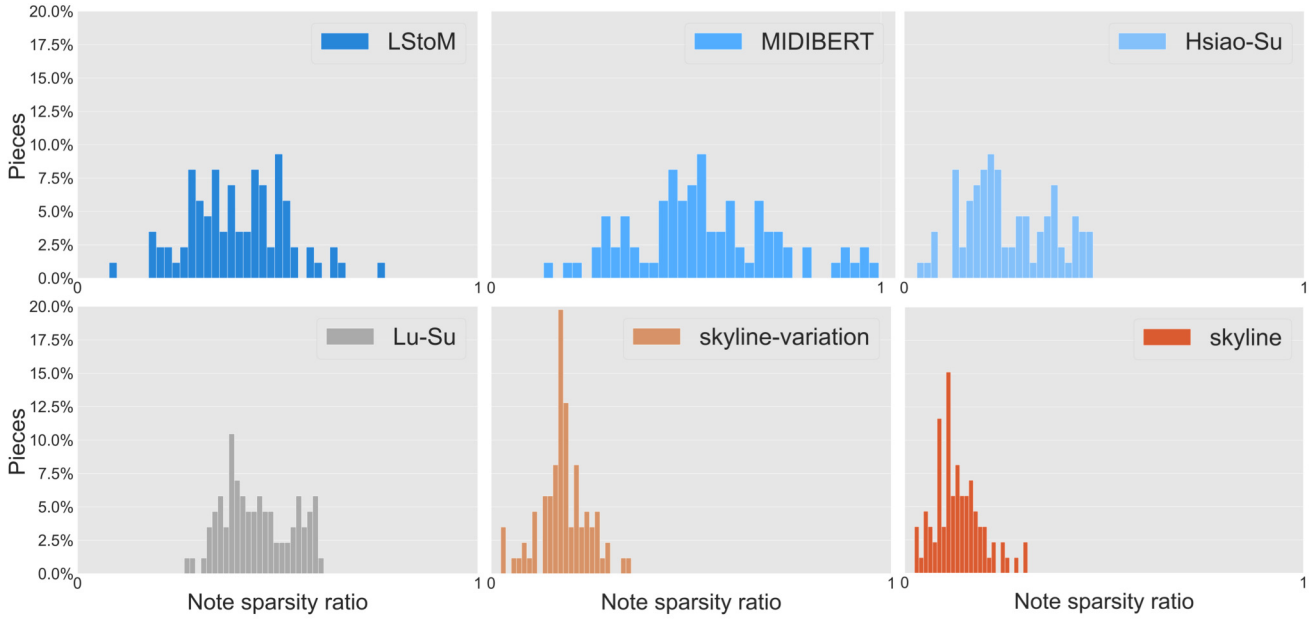

Figure 3. Sparsity ratio among melodies predicted by LStoM, MIDIBERT, "Hsiao-Su", "Lu-Su", skyline and skyline-variation.

one obtained by our trainings on "Hsiao-Su" and "Lu-Su" systems. Interestingly, the precision is reduced by a very small margin and the voice alarm error value is slightly increased, when adding the feature `pitch_in_scale`, however the accuracy is increased at the same time.

4.3 Observations

Having a closer look to the predicted melodies by LStoM, MIDIBERT, "Hsiao-Su", "Lu-Su" and the baselines, we can highlight two separate statistics. One is the distribution of the note intervals among the consecutive melody note pairs (Figure 2 – the x axis has been limited to the range of around two and a half octaves for paper fitting purposes), where overall the melodies predicted by the models tended to reflect characteristics from the original melody contour. Another statistic is the percentage of predicted melodies with various degrees of sparsity in them. By sparsity, we mean the ratio of the total duration of the silent parts over the total duration of the notes in a score. Figure 3 shows the

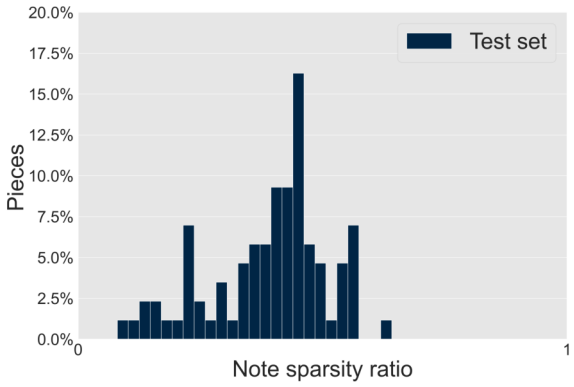

Figure 4. Sparsity ratio among melodies in the test set.

sparsity distribution among the predicted melodies, while Figure 4 highlights the ground truth from the test set. A small amount of melodies from MIDIBERT tends to be more sparse than the scores from the test set.

The most challenging melody notes to classify are those that are not the highest note; i.e., those that the skyline method would certainly miss. We have used a test set where around 8% of the melody notes are such challenging cases; of these, the LStoM system classifies correctly around 82%. What pattern in the feature data did the model use to achieve this high performance?

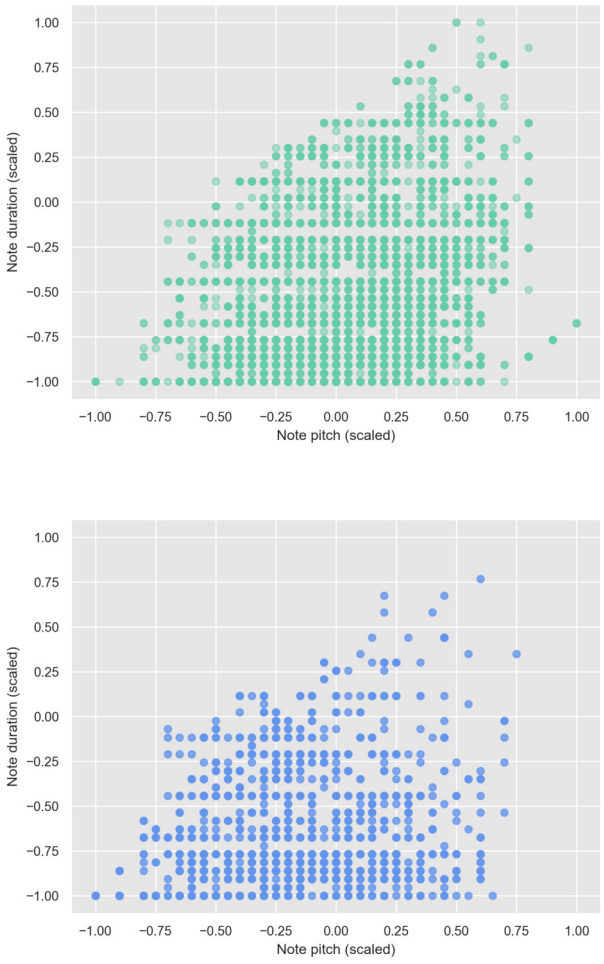

Figure 5. Scaled representation for the pitch-duration feature pair, for the melody notes that are not the highest note and have been either predicted correctly–top– or not (missed)–bottom.

It is not easy to examine the reason why sometimes such notes are not predicted as melody notes, but to investigate, we scaled the feature values of these true melody notes, and compared distributions of features between the correctly labelled (i.e., retrieved) and incorrectly labelled (i.e., missed) notes. One distribution that stood out is shown in Figure 5, a scatter plot of the pitch and duration of the melody notes. One may observe that although the distributions are similar, there are many more correctly labelled notes with high pitch and long duration. Such findings indicate that disentangling such points is not trivial, if at all feasible with our current feature set alone.

5. CONCLUSIONS AND PERSPECTIVES

In this paper we have presented a novel deep learning model of identifying the melody line from a polyphonic symbolic music. The key characteristics of the model is the input data representation as a set of features from the relevant task of multiple voice separation as well as the bidirectional nature of the architecture which provides information of future observations during the prediction process. The features have been examined regarding their ability to contain the information that the model needs to more accurately predict a melody note. Also, two data augmentation techniques are examined in isolation. Results indicate that the model trained and tested on a set of pop songs is capable of achieving a higher accuracy than other state of the art models. Observations on the results reveal the ability of the model to reflect the concepts of score sparsity or the melodic pitch intervals from the test set. LStoM also performs well in situations where the melody note to be identified is not in the highest pitch of the score. One direction to explore is expanding the set of input features. Note velocity information or metrical representation extracted from the dataset could accommodate additional features for our model. Post-processing techniques have not been explored thoroughly, however they could help improving the results.

Setting research experiments for the task of melody extraction in the symbolic domain is challenging, in terms of data gathering and manipulating separate model architectures to fit the needs of a comparison task. Accumulating the available datasets and the existing systems to a common space for easy approach, such as in a dedicated MIREX page is a promising step forwards. To this end, examining how our model performs using different datasets in the training or inference process, is a step towards a more comprehensive review.

A direction that has been little explored is combining information from the symbolic domain to the prediction of melodies in the audio domain. Most of the existing literature for audio melody transcription relies only on information from the audio signal in isolation from other musical context explicitly. However, in [22], given a polyphonic music audio signal, the model is able to convert it to a piano arrangement; as part of this style transferring process, MIDI scores have been used as ground truth. Also, in [23], symbolic data representation has been used to pre-train a model which then operates to the audio domain. Tasks such as de-noising could be revisited following this perspective.

6. ACKNOWLEDGEMENTS

We thank Jordan B. L. Smith and Janne Spijkervet for their extensive paper review and support before submission.

7. REFERENCES

[1] J. Salamon, E. Gómez, D. P. Ellis, and G. Richard, "Melody extraction from polyphonic music signals: Approaches, applications, and challenges," *IEEE Signal Processing Magazine*, vol. 31, no. 2, pp. 118–134, 2014.

[2] N. Cook, P. Cook, and A. C. of Learned Societies, *Music, Imagination, and Culture*, ser. ACLS Humanities E-Book. Clarendon Press, 1990. [Online]. Available: https://books.google.co.uk/books?id=5DKiUOw519oC

[3] E. Cambouropoulos, "Voice and stream: Perceptual and computational modeling of voice separation," *Music Perception*, vol. 26, no. 1, pp. 75–94, 09 2008.

[4] R. de Valk and T. Weyde, "Deep neural networks with voice entry estimation heuristics for voice separation in symbolic music representations," in *Proceedings of the 19th International Society for Music Information Retrieval Conference*, 2018.

[5] R. Martín, R. A. Mollineda, and V. García, "Melodic track identification in midi files considering the imbalanced context," in *Pattern Recognition and Image Analysis*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 489–496.

[6] Z. Jiang and R. B. Dannenberg, "Melody identification in standard midi files," in *Proceedings of the 16th Sound and Music Computing Conference (SMC)*, 2019, pp. 65–71. [Online]. Available: https://www.cs.cmu.edu/~rbd/papers/melody-identification-midi-smc2019.pdf

[7] S. Li, S. Jang, and Y. Sung, "Melody extraction and encoding method for generating healthcare music automatically," *Electronics*, vol. 8, no. 11, 2019. [Online]. Available: https://www.mdpi.com/2079-9292/8/11/1250

[8] A. L. Uitdenbogerd and J. Zobel, "Melodic matching techniques for large music databases," in *Proceedings of the 7th ACM international conference on Multimedia (Part 1)*, 1999, pp. 57–66.

[9] F. Simonetta, C. E. Cancino-Chacón, S. Ntalampiras, and G. Widmer, "A convolutional approach to melody line identification in symbolic scores," *Proceedings of the 20th International Society for Music Information Retrieval Conference*, vol. abs/1906.10547, 2019.

[10] W.-T. Lu and L. Su, "Deep learning models for melody perception: An investigation on symbolic music data," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2018, pp. 1620–1625.

[11] Y.-W. Hsiao and L. Su, "Learning note-to-note affinity for voice segregation and melody line identification of symbolic music data," *Proceedings of the 22nd International Society for Music Information Retrieval Conference*, pp. 285–292, 2021.

[12] H. Zhao and Z. Qin, "Tunerank model for main melody extraction from multi-part musical scores," in *Proceedings of the 2014 Sixth International Conference on Intelligent Human-Machine Systems and Cybernetics - Volume 02*, ser. IHMSC '14. USA: IEEE Computer Society, 2014, p. 176–180. [Online]. Available: https://doi.org/10.1109/IHMSC.2014.145

[13] Y.-H. Chou, I.-C. Chen, C.-J. Chang, J. Ching, and Y.-H. Yang, "MidiBERT-Piano: Large-scale pre-training for symbolic music understanding," *arXiv preprint arXiv:2107.05223*, 2021.

[14] Z. Wang, K. Chen, J. Jiang, Y. Zhang, M. Xu, S. Dai, X. Gu, and G. Xia, "Pop909: A pop-song dataset for music arrangement generation," in *Proceeding of the 21st International Society for Music Information Retrieval (ISMIR)*, 2020.

[15] M. S. Cuthbert and C. Ariza, "music21: A toolkit for computer-aided musicology and symbolic music data," 2010.

[16] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm and other neural network architectures," *Neural Networks*, vol. 18, no. 5, pp. 602–610, 2005, iJCNN 2005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0893608005001206

[17] J. Bergstra, D. Yamins, and D. D. Cox, "Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures," in *30th International Conference on Machine Learning, ICML 2013*, vol. 28, 2013, pp. 115–123.

[18] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[19] C. Raffel, B. Mcfee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, D. P. W. Ellis, C. C. Raffel, B. Mcfee, and E. J. Humphrey, "mir_eval: a transparent implementation of common mir metrics," in *In Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR*, 2014.

[20] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

[21] R. J. Urbanowicz, R. S. Olson, P. Schmitt, M. Meeker, and J. H. Moore, "Benchmarking relief-based feature selection methods," *CoRR*, vol. abs/1711.08477, 2017. [Online]. Available: http://arxiv.org/abs/1711.08477

[22] Z. Wang, D. Xu, G. Xia, and Y. Shan, "Audio-to-symbolic arrangement via cross-modal music representation learning," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 181–185.

[23] W. T. Lu, L. Su *et al.*, "Vocal melody extraction with semantic segmentation and audio-symbolic domain transfer learning." in *ISMIR*, 2018, pp. 521–528.

A REPRODUCIBILITY STUDY ON USER-CENTRIC MIR RESEARCH AND WHY IT IS IMPORTANT

Peter Knees[1,3] **Bruce Ferwerda**[2] **Andreas Rauber**[1]
Sebastian Strumbelj[1] **Annabel Resch**[1] **Laurenz Tomandl**[1] **Valentin Bauer**[1]
Fung Yee Tang[1] **Josip Bobinac**[1] **Amila Ceranic**[1] **Riad Dizdar**[1]
[1] Faculty of Informatics, TU Wien, Austria
[2] Department of Computer Science and Informatics, Jönköping University, Sweden
[3] School of Music, Georgia Institute of Technology, USA
peter.knees@tuwien.ac.at

ABSTRACT

Reproducibility of results is a central pillar of scientific work. In music information retrieval research, this is widely acknowledged and practiced by the community by re-implementing algorithms and re-validating machine learning experiments. In this paper, we argue for an increased need to also reproduce the results and findings of user studies, including qualitative work, especially since these often lay the foundations and serve as justification for choices taken in algorithmic design and optimization criteria. As an example, we attempt to reproduce the study by Kim et al. [1] presented in the RecSys (2020) paper "Do Channels Matter? Illuminating Interpersonal Influence on Music Recommendations". By repeating this study on how interpersonal relationships can affect a user's assessment of music recommendations on a new sample of $n = 142$ participants, we can largely confirm and support the validity of the original results. At the same time, we extend the analysis and also observe differences with regards to adoption rates between different channels as well as different factors that influences the adoption rate. From this specific reproducibility study, we conclude that potential cultural differences should be accounted for more explicitly in future studies and that systems development should be more explicitly connected to its intended target audience.

1. INTRODUCTION AND CONTEXT

Ensuring the demonstrable reproducibility of experiments is an essential part of the scientific method to acquire knowledge. Reproducibility in music information retrieval research traditionally has a strong focus on the repeatability of experiments that demonstrate the performance of a system. To this end, the formalization and complete documentation of MIR workflows and processes is a top priority (e.g., [2]). This includes proper documentation and provision of access to the data involved, to avoid contributing to the so-called reproducibility crisis found throughout virtually all scientific disciplines. In the MIR community the aspect of data access has been a topic of discussion since the beginning, as dataset sharing is a particularly challenging and sensitive matter, foremost due to copyright issues, cf. [3,4]. Also for evaluation, emphasis is given to applying transparent, open, and sustainable methods, to objectively quantify the performance of systems and consistently compare systems [5]. In short, many efforts of the community are devoted to sharing resources, re-implementing algorithms, and re-validating machine learning experiments to ensure reliable and valid scientific results.

In order to avoid potentially contributing to the reproducibility crisis on another front, we argue that there is urgent need in MIR research to also reproduce the results and findings of user-centric studies, including qualitative work. The system-centric view in MIR has been repeatedly subject to criticism and raised calls for more user-centric approaches (e.g., [6, 7]) as the objectives of the systems developed are ultimately rooted in users' needs. In practice this has led to user-centric work taking the essential roles of requirement engineering (e.g. [8–10]) and/or of a system evaluation vehicle, either by generating ground truth, rating the outcomes of systems, or reflecting on their use (e.g., [11–13]). As their outcomes have (or are supposed to have) a fundamental impact and serve as justification for choices taken in algorithmic design and deployed optimization criteria [14, 15], their validity is by no means of lesser importance than that of system-based experiments. Clearly, reproducibility of experiments involving human subjects is hard, as witnessed in the social sciences [16], and has led to the development of the current evaluation strategies as a workaround [17]. Nonetheless, when it comes to the justification of systems objectives, further attention should be paid to the scientific validity of findings by reproducing and corroborating the findings of studies.

As an example, in this paper, we attempt to reproduce the study by Kim et al. [1], presented in the ACM Rec-

Sys 2020 paper "Do Channels Matter? Illuminating Interpersonal Influence on Music Recommendations". We have chosen this particular user study for several reasons. First, the study subject is highly relevant for the development and understanding of music recommender systems, therefore clarifying the reproducibility of the experiments is important for future research. Second, the study is well documented with a clearly laid out methodology. Third, as the original data was collected via a web-based survey, to obtain a new set of responses of comparable quality, we can follow the same strategy, with the main difference that recruited participants have a different cultural, i.e. predominantly European, background.

In addition to repeating the original study to investigate the differences of music recommendations from music recommender systems (non-interpersonal) and friends and acquaintances (interpersonal), we carry out further analyses to gain insights regarding gender differences. We further make use of the PRIMAD framework for reproducibility of research in e-Science for positioning our work [18]. In the terminology of the PRIMAD framework, in our reproducibility study, we foremost prime the factor of data and consequently investigate the generalization capabilities of the analysis wrt. another sample of users. Through the additional analyses, we also extend the methodology used.

The remainder of this paper is structured as follows. Section 2 briefly motivates and summarizes the original study by Kim et al. [1]. In Section 3, we describe the methods used to reproduce the study. Section 4 reports on the results and findings of the reproduced study, which are further discussed in relation to the original study in Section 5. In Section 6, we use the PRIMAD model as a tool to systematically categorize the dimensions of reproducibility in this work and suggest its use for future reproducibility studies (not only) of user-centric studies in the MIR research community. Finally, in Section 7 we draw conclusions from this reproducibility study and point out aspects to be considered for future work.

2. ORIGINAL WORK TO BE REPRODUCED

We aim at reproducing the findings of Kim et al. [1], who have conducted a web survey to investigate the perceptions, evaluations, and differences of music recommendations received through two different channels: *interpersonal*, i.e., music recommendations received from friends or acquaintances, and *non-interpersonal*, i.e. music recommendations received from recommendation systems and through updates on artists followed. The rationale behind that study is to gain a deeper understanding of the characteristics of interpersonal music recommendations and how the interpersonal relationship affects the evaluation of the recommendation.

To this end they recruited 175 Korean-speaking participants (56% female; all above age 18, with 91% of participants between 18 and 32). [1] Furthermore, they collected information that 85% of the participants were active users of a music subscription service for more than one year and that 62% of participants were actively sharing music with others more than once per week (see Table 1).

Study	**Kim et al. [1]**	**Current study**
Number of participants (n)	175	142
Female participants	56%	45%
Age between 18 and 32	91%	86%
Use music subscription service	85%	69%
Sharing music with others	62%	16%

Table 1. Demographics and characteristics of survey participants in the original study and the reproducibility study. Values for the current study refer to percentages in recorded responses.

The study shows an interesting trend, namely that music recommendations obtained from a system are typically considered to be more relevant for a user's own taste (which is the primary objective of recommender systems), more convenient, and more frequently used and adopted than interpersonal recommendations. On the other hand, aspects of diversity, novelty, and serendipity were evaluated higher when recommendations were obtained through interpersonal channels. More detailed results of the study are also included in this paper for comparison in Section 4.

From these results, the authors conclude that indeed interpersonal and non-interpersonal channels have different characteristics that should be accounted for when designing music recommendation platforms, particular to take advantage of their individual strengths (e.g., relevance vs. diversity and novelty). They further highlight a limitation of their study and point to a possible next step: "Since the study was conducted in Korea, Korean cultural background was reflected in the results. Future studies with participants from other cultures (e.g., with different norms, technologies available, etc.) will enable us to explore how cultural differences can influence users' perceptions and behaviors towards music recommendations from different channels." In this paper, we aim at gaining such insights by conducting the same study again with a fresh set of participants, with a different background (i.e., predominantly European), as described next.

3. METHODS

Following the method of Kim et al. [1], we used a web-based survey to investigate interpersonal and non-interpersonal music channels in the context of music playlist creation and consumption. A similar survey was designed with a section regarding interpersonal channels (i.e., from friends or acquaintances) and a section regarding non-interpersonal channels (i.e., from recommendation systems) and sections with the respective questions were presented in a random order to the participants (see Table 2 for the survey questions). The survey was conducted in English with all participants. The questions for each

[1] Despite highlighting that all participants spoke Korean fluently, it is not explicitly stated whether the survey was conducted in Korean or English.