Abstract: Cover song identification is a field of music information retrieval where the task is to determine whether two different audio tracks represent different versions of the same underlying song. Since covers might differ in tempo, key, instrumentation and other characteristics, many clever features have been developed over the years. We perform a rigorous analysis of 32 features used in related works while distinguishing between exact and scalable features. The former are based on a harmonic descriptor time series (typically chroma vectors) and offer better performance at the cost of computation time. The latter have a small constant size and only capture global phenomena in the track, making them fast to compute and suitable for use with large datasets. We then select 7 scalable and 3 exact features to build our own two-level system, with the scalable features used on the first level to prune the dataset and the exact on the second level to refine the results. Two distinct machine learning models are used to combine the scalable resp. exact features.

We perform the analysis and the evaluation of our system on the Million Song Dataset. The experiments show the exact features being outperformed by the scalable ones, which lead us to a decision to only use the 7 scalable features in our system. The performance of this model is comparable with other state-of-the-art methods tested on the same dataset. The surprisingly poor performance of exact features is discussed and we conclude that the main culprit is probably the inferior quality of the descriptors used in the dataset and that the scalable features manifest remarkable robustness to the lower quality of the data.