Information Processing and Management 61 (2024) 103610

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

Predicting movies' eudaimonic and hedonic scores: A machine learning approach using metadata, audio and visual features

Elham Motamedi [a],*, Danial Khosh Kholgh [b], Sorush Saghari [c], Mehdi Elahi [d], Francesco Barile [e], Marko Tkalcic [a]

[a] *University of Primorska, Koper, Slovenia*
[b] *University of Oulu, Oulu, Finland*
[c] *K N Toosi University of Technology, Tehran, Iran*
[d] *University of Bergen, Bergen, Norway*
[e] *Maastricht University, Maastricht, Netherlands*

ARTICLE INFO

Keywords:
Eudaimonia
Hedonia
Machine learning approach
Movie recommender systems

ABSTRACT

In the task of modeling user preferences for movie recommender systems, recent research has demonstrated the benefits of describing movies with their eudaimonic and hedonic scores (E and H scores), which reflect the depth of their message and the level of fun experience they provide, respectively. So far, the labeling of movies with their E and H scores has been done manually using a dedicated instrument (a questionnaire), which is time-consuming. To address this issue, we propose an automatic approach for predicting E and H scores. Specifically, we collected E and H scores of 709 movies from 370 users (with a total of 3699 records), augmented this dataset with metadata, audio, and low-level and high-level visual features, and trained machine learning models for predicting the E and H scores of movies. This study investigates the use of machine learning models in predicting the E and H scores of movies using various feature sets, including audio, low-level and high-level visual features, and metadata. We compared the performance of predictive models using different combinations of features with the majority classifier as the baseline approach. The results demonstrate that our proposed machine learning-based models significantly outperform the baseline in predicting E and H scores, particularly when leveraging metadata features. Specifically, the random forest classifier achieved a 20% increase in ROC AUC compared to the baseline when predicting both the E score and the H score. These improvements were found to be statistically significant. Overall, our findings suggest that automated tools for predicting E and H scores in movies are promising alternatives to traditional questionnaire-based approaches.

1. Introduction

Recommender system algorithms have been shown to be effective in predicting the utility of an item for a target user. In particular, matrix factorization and recently deep learning algorithms are very efficient in extracting user preferences from past behavior (e.g., clicks, purchases, views, etc.). This information is usually encoded in the form of latent features that represent different aspects of an item and the potential preferences of the target user towards them. However, despite the tremendous ability to

* Corresponding author.
E-mail addresses: elham.motamedi@famnit.upr.si (E. Motamedi), danial.khoshkholgh@oulu.fi (D.K. Kholgh), s.saghary@email.kntu.ac.ir (S. Saghari), mehdi.elahi@polimi.it (M. Elahi), f.barile@maastrichtuniversity.nl (F. Barile), marko.tkalcic@famnit.upr.si (M. Tkalcic).

https://doi.org/10.1016/j.ipm.2023.103610
Received 8 May 2023; Received in revised form 16 November 2023; Accepted 7 December 2023
Available online 14 December 2023

generate accurate recommendations, what these approaches still lack is interpretable reasoning on why a certain item (e.g. a film or a music piece) is suitable for a user, mainly due to the difficulty of interpreting the latent features. This can result in scenarios where the recommender system cannot tell the quality of the user consumption experience, as illustrated by the Netflix example: Netflix reported that they cannot distinguish whether the time users spend on their platform is quality time or addiction (Hunt, 2014). This is why there is a need for utilizing not only the latent features but also additional features that describe the consumption experience of users in a human-understandable way. To this end, we have proposed to describe items with features, henceforth called *scores*, that research in positive psychology has shown to be effective in explaining the user experience: *hedonia* (the pure enjoyment aspect of the experience) and *eudaimonia* (the experience aspect that describes the deeper meaning, life purpose, and personal values) (Botella et al., 2012; Hrustanovic, Kavsek, & Tkalcic, 2021; Motamedi, Barile, & Tkalčič, 2022; Oliver & Raney, 2011).

In order to label items with eudaimonic and hedonic scores (E and H scores), there exist validated instruments, such as the Oliver and Raney's scale (Oliver & Raney, 2011). However, this process is time consuming and usually feasible only in small scale studies in research environments. In order to foster the usage of interpretable scores in large scale studies and in industrial environments, an automatic approach for labeling items with eudaimonic and hedonic scores is needed. In order to address this need, in this paper, we present a machine learning method for labeling movies with eudaimonic and hedonic scores based on a wide range of features describing the movies, including a novel set of *low-level* and *high-level* audio-visual features, in addition to the traditional metadata. Through a series of experiments, we demonstrated that movies can be classified into high/low eudaimonic/hedonic movies with a classification accuracy substantially higher than with the baseline predictor. This work represents an important step towards more explainable recommender systems using interpretable scores, such as eudaimonic and hedonic scores.

In a nutshell, the contributions of our paper are as follows:

1. To the best of our knowledge, this is the first study that proposes an automatic model to predict the eudaimonic and hedonic scores of a dataset of several hundreds of movies (precisely, 709), surpassing previous studies that used smaller datasets of several tens of movies.
2. We used a variety of features, including audio, metadata, and low-level and high-level visual features, to develop this model, which has not been done before.
3. For the first time, we report on the correlation between different movie genres and their E and H scores, providing insight into the predictive power of metadata features such as genres.

The rest of this paper is organized as follows. In the next section, the related works are summarized. In Section 3 the machine learning approach for predicting eudaimonic and hedonic scores as well as data acquisition are described. In Section 4 the data analysis output and prediction results are provided, followed by discussion and conclusion in Section 6.

2. Literature review

The work we present stands on the shoulders of research done in two separate domains: (a) eudaimonic and hedonic experience of movie consumption from positive psychology and (b) machine learning prediction models based on different movie features and metadata.

2.1. Eudaimonic and hedonic scores of movies

The ways to achieve happiness have been a subject of debate among Greek philosophers since ancient times. Aristippus, who lived in the third century BC, equated happiness with ease, plain pleasure and avoiding pain (Waterman, 1993). However, other philosophers later presented a different perspective on achieving happiness. Their view of happiness was based on eudaimonia, which was about pursuing more complex and meaningful goals. This debate found its way into positive psychology, where researchers have explored both eudaimonic and hedonic perspectives. Happiness was viewed from a hedonistic perspective by Kahneman, Diener, and Schwarz (1999) in their book titled *"The Foundations of Hedonic Psychology"*. However, later on, other researchers have seen well-being or happiness from another perspective named eudaimonia in which pursuing one's full potential can be a source of happiness. Positive psychology has since recognized both perspectives and acknowledges the distinction between the two perspectives, demonstrating that they are related but distinguishable. Ultimately, a person can achieve happiness by pursuing both motivations to a certain extent (Li et al., 2022).

Eudaimonic and hedonic motivations play a significant role in the selection of various means of entertainment. Oliver and Raney (2011) extended the scope of entertainment selection beyond pleasure-seeking (hedonic concerns) to include media selection as a means of truth-seeking (eudaimonic concerns). They developed an instrument to measure the eudaimonic and hedonic components of the user experience while watching movies. Recent studies have explored the application of these motivations in different means of entertainment such as movies and music. For instance, Hrustanovic et al. (2021) predicted the eudaimonic and hedonic scores of songs from their lyrics. In a similar work, Motamedi and Tkalcic (2021) predicted the eudaimonic and hedonic scores of movies from subtitles. De Ridder, Vandebosch, and Dhoest (2022) examined the eudaimonic and hedonic scores as indicators of motivation in stand-up comedy TV shows. They found that combining eudaimonic and hedonic entertainment experiences could lead to a more inclusive, accessible, and overall positive narrative. This is consistent with previous research showing that users typically seek entertainment items that offer both eudaimonic and hedonic motivations, rather than the ones with solely high eudaimonic or high hedonic scores (Motamedi et al., 2022). This shows that eudaimonic and hedonic scores are not mutually exclusive, and individuals can have varying degrees of orientation towards each.

The eudaimonic and hedonic scores can be used to describe both users and items. Tkalčič and Ferwerda (2018) showed that there is variance among users in their propensities (or orientations) for content that evokes eudaimonic and hedonic experiences. They proposed to model users with the eudaimonic and hedonic orientations for movie recommendations. To recommend suitable movies to a user with a certain eudaimonic and/or hedonic orientation, it is important to label movies with their potential to evoke such experiences, which we refer to as eudaimonic and hedonic scores of movies.

Annotating items with eudaimonic and hedonic scores is an intrusive and cumbersome task. Therefore, recent research has proposed automatic approaches to predict these characteristics using various features. Motamedi et al. (2022) developed an automatic model that predicts users' eudaimonic and hedonic orientations using user-related information and user–item interaction information. Their results indicated that a model using only user-related information outperformed the other model. In another study, Hrustanovic et al. (2021) developed a predictive model that uses song lyrics to predict their eudaimonic and hedonic scores. In a similar work, Motamedi and Tkalcic (2021) developed a model that predicts the eudaimonic and hedonic scores of movies using subtitles. While their study is similar to this work in that both generate models to predict eudaimonic and hedonic scores from movie features, there are important differences between them. This study uses a broader range of features, including metadata, audio, and low- and high-level visual features, rather than just subtitles. Additionally, their study was conducted on a small sample size of only 30 movies, leading to a large standard deviation in reported performance values. To our knowledge, there is no existing work predicting eudaimonic and hedonic scores of movies from audio-visual features and metadata, making this study particularly novel. Overall, this study's approach significantly advances our understanding of the predictive power of different movie features in predicting eudaimonic and hedonic scores.

2.2. Machine learning based on movie features

The task of modeling and analyzing movie content has been the focus of research in various fields such as machine learning (Atrey, Hossain, El Saddik, & Kankanhalli, 2010; Beheshti, Ghodratnama, Elahi, & Farhood, 2022; Gong & Xu, 2007). There are already comprehensive surveys on the state-of-the-art in movie content analysis and classification, that review different forms of movie features (e.g., visual, auditory, or metadata) (Brezeale & Cook, 2008; Hu, Xie, Li, Zeng, & Maybank, 2011). Such features can be automatically extracted and engineered to describe the content of movies and be incorporated for various purposes. While these features may substantially differ in their nature, they can still be classified into different categories, i.e., (i) *low-level* stylistic or aural features that capture the audio-visual properties of the movies, (ii) *high-level* syntactic features that capture objects and their interactions in the movies, and, (iii) *high-level* semantic features that capture conceptual modeling of movies (Wang, Xing, & Zhou, 2006). The latter type of features can also be referred to as *metadata*.

In the machine learning field, several models have been developed to leverage the above-described movie features and employ them in various tasks in research disciplines. Examples of such disciplines are computer vision, movie recommendation, and retrieval (Lew, Sebe, Djeraba, & Jain, 2006; Rasheed, Sheikh, & Shah, 2005) and different tasks are addressed by them, notably, defining a meaningful representation of movie content, and classification of movies based on different features. For example, Liu, Zhang, and Gulla (2020) used several fusion techniques for mutual association learning across modalities (textual and visual modalities) in the context of explainable recommendations. The majority of approaches based on these machine learning models often tend to exploit high-level features perhaps due to their closer alignment with human understanding of movies. High-level features are semantic descriptors that can express characteristics of items obtained from different structured or unstructured sources of metadata content. Examples of structured sources are databases, ontologies, and lexicons. Examples of unstructured sources are item descriptions, news articles, and social tags. On the other hand, low-level features usually capture *stylistic* properties of movies and can be directly extracted from movie files themselves (Deldjoo et al., 2016).

There have been less number of approaches focused on low-level features and the importance of these features in human perception of movie style has been an under-explored area in relevant research fields. This is while movie makers and directors extensively use human perception to convey emotions and feelings to the users (audience) during movie creation.

A number of studies have proved the effectiveness of exploiting low-level features for a wide range of prediction tasks (Allamanche, Hellmuth, Fröba, Kastner, & Cremer, 2001; Cerf, Harel, Einhäuser, & Koch, 2008; Montalvo-Lezama, Montalvo-Lezama, & Fuentes-Pineda, 2022). Other studies exploited these features in combination with high-level features to enhance the accuracy of their predictive model (Nguyen, Scholer, Miele, Edwards, & Fujita, 2022). Yang et al. (2007) is one of the early studies proposing a machine learning model based on both high-level and low-level features (i.e., textual, and audio-visual) extracted from movies to improve the click-through rate scores. In a related study (Zhou, Hermans, Karandikar, & Rehg, 2010), authors proposed a framework for the automatic classification of movie genres by exploiting temporally structured features based on the intermediate level of movie representation. In Rasheed et al. (2005) a practical movie genre classification scheme was proposed leveraging a set of computable visual cues, while in Rasheed and Shah (2003) audio features were exploited to further enrich these visual cues. In a different study, Zhao et al. (2011) proposed a model to compute ranking lists based on various movie features including visual features. The ranking lists for a user were computed to reflect her interests in movies based on the movies previously accessed by that user. The ranking lists were ultimately integrated to form recommendations of movies for the user. Xi, Xu, Chen, Zhou, and Yang (2021) proposed a deep neural network integrating multimodal information to predict whether viewers will send Danmaku comments. Their study examined the impact of the interaction among textual, audio, and visual features on this behavior.

In a more recent study (Moghaddam, Elahi, Hosseini, Trattner, & Tkalcic, 2019), authors presented an innovative approach to leverage visual features from movie trailers to predict the popularity and the average rating of movies. These visual features can be extracted and engineered automatically without any human involvement and can be highly effective in representing the visual attractiveness of movies. To validate this approach, a set of experiments was conducted using a large dataset comprising over 13,000 movie trailers. The results demonstrate the effectiveness of the approach. Another study is the work of Rimaz et al. (2019) where the potential of incorporating visual features into movie recommendations was investigated. A set of experiments was conducted including exploratory analysis to gain initial insights into the data and model development and evaluation to assess the quality of visually-aware movie recommendations. The results of the experiments demonstrated the promising potential of such features in representing movies' superior quality of recommendations based on them in comparison to baselines.

Our paper differs from the above-described works from different perspectives. First, we focus on a specific task of predicting the eudaimonic and hedonic scores of movies. This is a novel line of research, and, to the best of our knowledge, no or very limited prior works investigated this. In addition to that, we leverage audio-visual features together with traditional metadata to perform the task at hand. This is another novelty that has been less explored in this particular area of research.

The work we present here is the application of features extracted from video, audio, and metadata to predict target variables that have not had predictive models so far, the eudaimonic and hedonic scores. The concrete research questions we address are:

- **RQ1**. Can we develop an automatic model that predicts movies' eudaimonic and hedonic scores?
- **RQ2**. Which individual features (i.e., audio, metadata, and low-level and high-level visual features) and combinations of features result in the most accurate predictions of movies' eudaimonic and hedonic scores?

To validate the proposed approach, we collected a dataset of movies and their associated eudaimonic and hedonic scores, used features gathered from the video and audio of the movies, collected movies' metadata, trained several prediction algorithms, and compared the performances of these algorithms using different sets of features.

3. Methods and materials

This study examines the use of metadata, audio, and visual features in predicting the eudaimonic and hedonic scores of movies. In the domain of movie consumption, the eudaimonic (E) score represents how much a movie conveys a meaningful message to the viewer, while the hedonic (H) score reflects the level of entertainment the movie provides for users. Although intuitively one would say that the eudaimonic and hedonic scores are mutually exclusive (i.e. a high eudaimonic score implies a low hedonic score and vice versa), our data show the two dimensions are quite independent (as seen in Fig. 3). While some movies are high in one score and low in the other (e.g. *Airplane* is high in hedonic and low in eudaimonic, *The Pianist* is high on eudaimonic and low on hedonic) there are several movies that are high on both scores (e.g. *Glory*) or low on both (e.g. *Strangers on a Train*).

To determine the extent to which predictive models can predict the movies' eudaimonic and hedonic scores, we have developed machine learning models that incorporate various data sources including metadata such as movie genre, released year, popularity, and critical ratings, as well as audio and visual features of the movies. The process of eudaimonic and hedonic score prediction is demonstrated in Fig. 1 which involves: (i), data acquisition, (ii), data pre-processing, (iii), building, training, and testing the predictive models for the classification task, and (iv), evaluating the model performance. The findings of this study will provide insights into the relationship between different movies' features and their eudaimonic and hedonic scores.

3.1. Data acquisition

We collected the dataset in three steps. Initially, we obtained ratings from 370 users for 709 movies, yielding E and H scores. Next, we collected metadata for the same movies using the IMDb API. Finally, we augmented our dataset by incorporating low-level and high-level descriptors from video using the dataset collected by Moghaddam et al. (2019) and Elahi et al. (2021). We used the dataset collected by Rimaz, Hosseini, Elahi, and Moghaddam (2021) for audio descriptors. Further information is presented in the subsequent paragraphs.

In this study, we leveraged various movie features, including low-level and high-level visual features, audio features, and metadata, to predict movies' eudaimonic and hedonic scores. To collect eudaimonic and hedonic scores of movies, we conducted a user study with several steps. One part of the user study involved users completing a questionnaire designed by Oliver and Raney (2011) to evaluate their perception of the movies. We used the responses collected from this questionnaire to assign eudaimonic and hedonic scores to the movies. The movies shown to the users were selected from a pool of 1000 popular movies across different years, whose popularity was estimated based on the number of votes they received from MovieLens users. We developed a web application to acquire movies' E and H scores. When designing the data collection web app, particular attention was paid to avoiding data sparsity. We implemented a web interface where the participants filled out the questionnaires on personal characteristics and provided ratings and labels for at least ten movies out of 50. These sets of 50 movies, different for each participant, were selected through a sparsity-minimization mechanism. In this mechanism, we took a subpool of 200 movies from the larger, 1000-movies pool of popular movies from the Movielens dataset. We randomly pulled out 50 movies from the sub-pool for each participant to label/rate. These 50 movies were compiled from sets of 5, ensuring that the initial set included movies curated by one user, followed by two in the next set, and so forth, until reaching ten. After a movie reached a certain number of ratings/labels (10 in our case), we removed it from the sub-pool and replaced it with another from the larger pool. This ensured that we did not have a long tail problem (with few movies getting many labels/ratings and the majority having few) but the rated movies in our dataset had a

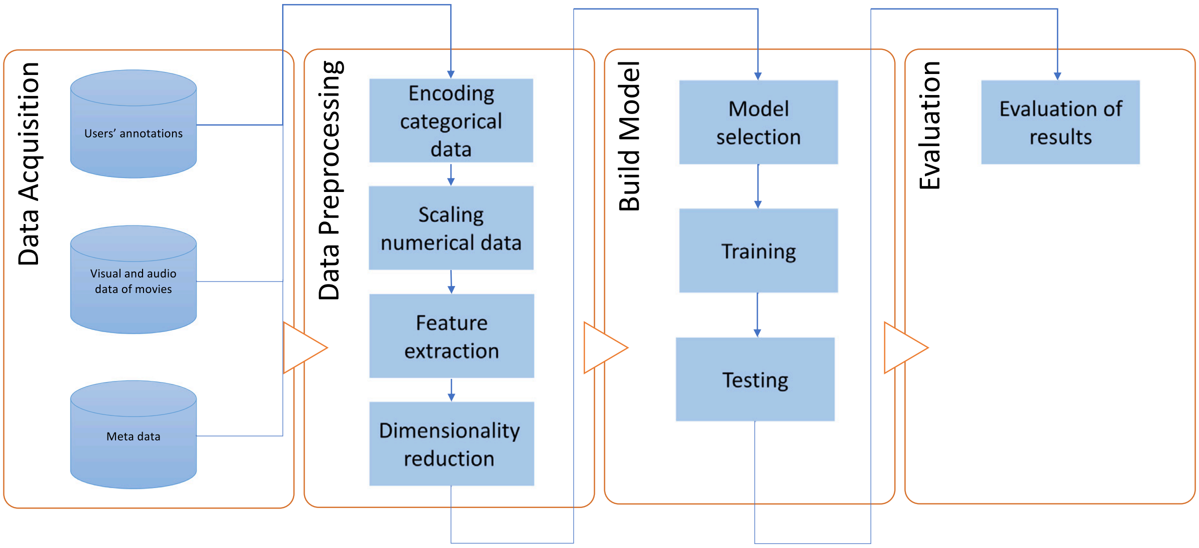


Fig. 1. The process of eudaimonic and hedonic scores prediction using machine learning models.

more flat distribution of labels/ratings. We chose a threshold of ten ratings to strike a balance between gathering sufficient data for each movie and obtaining information for a wide range of movies. When the sub-pool of 200 movies was used up, another batch of 200 movies was introduced. In this study, the average E and H experiences annotated by users are referred to as the E and H scores, respectively. In total, we collected E and H scores for 709 movies released from 1930 to 2015. To formulate the classification problem, we used the median split method to define low/high E score and low/high H score classes.

To determine the reliability of the categories of high or low E and high or low H scores assigned by the users, we calculate Cronbach's alpha values for E and H classes. This approach allowed us to assess the consistency among users in the given labels and establish their validity as reliable ground truth, given that the E and H scores in this study represent the average scores reported by users. Our analysis yielded Cronbach's alpha values of 0.76 and 0.84 (90% confidence) for the E and H scores, respectively, after excluding movies with fewer than ten annotations. These results indicate good consistency among the annotations, which supports the use of the labels as a reliable measure of E and H scores.

The metadata for the movie dataset was acquired using the IMDb API, including several features. The genre of each movie, as well as its Metacritic rating, popularity measured by the number of ratings in the IMDb dataset, the year of release, the countries of production, the duration of the movie in minutes, and the primary spoken language(s) information, exists in the metadata dataset. In total, the metadata was available for 709 movies.

The low-level visual features of the movie trailers in this study were obtained from the MA14KD dataset collected by Moghaddam et al. (2019). This dataset contains ten low-level visual features that measure the attraction of each movie trailer frame, including sharpness, variations in sharpness, contrast, RGB contrast, saturation, variations in saturation, brightness, colorfulness, entropy, and naturalness. They provided an aggregated version of the dataset we used in this study. According to their paper, after splitting the movie shots i.e., sequences of consecutive movie frames captured without interruption of the camera, keyframes were identified as representative of every individual shot. Then keyframes were thoroughly analyzed to extract visual features. Finally, visual features were aggregated over the whole movie to create an individual feature vector representative of every movie.

In addition to the low-level visual features, the study incorporated high-level visual features from the MVT9KD dataset collected by Elahi et al. (2021). This dataset contains visual tags, i.e., tags that were automatically annotated by deeply analyzing the content of the movies and detecting faces, objects, and even celebrities within the movies (e.g., Titanic 1997 annotated with #KateWinslet tag). The dataset contains visual tags for over 9000 movie trailers that were automatically generated using the AWS Rekognition service[1] along with a confidence level. This service uses pre-trained models that provide labels based on video content along with the confidence level for each label. In this study, we included the data with a confidence level of 95% or higher that contains the features of faces that appeared in the movie frames. The face tags included the predicated age range of the faces in the trailers and the strength of emotions such as happiness, sadness, anger, confusion, disgust, surprise, calmness, and fear, with values ranging from 0 to 100. The merged dataset, including low-visual features and eudaimonic/hedonic scores, consisted of 394 movies, whereas the merged datasets, including high-visual features and eudaimonic/hedonic scores, consisted of 303 movies.

For audio aspects of movies, we used the dataset generated by Rimaz et al. (2021), which includes features such as danceability, energy, liveness, loudness, tempo and valence, extracted from over 9000 movies. We used the 617 movies for which we had eudaimonic and hedonic scores.

[1] A cloud-based computer vision platform offered by Amazon for image recognition and video analysis: https://aws.amazon.com/rekognition/.

3.2. Data pre-processing

Several steps were done as data pre-processing steps: (i) encoding categorical data, (ii) scaling numerical data, (iii) feature extraction, and (iv) dimensionality reduction.

To prepare the data for the next steps, we transformed the categorical features (i.e., genre, countries, and languages) into numerical representations through one-hot encoding, as these variables were not ordinal. The remaining numerical data, such as the Metacritic rating and run time, were scaled before passing to machine-learning models.

We created several datasets by incorporating different sets of features including audio, metadata, high-level, and low-level visual features that were collected from available datasets (explained more in detail in Section 3.1). Specifically for high-level visual features, we used the dataset MVT9KD collected by Elahi et al. (2021), and we aggregated the high-level visual features, by calculating various statistical measures, including the minimum, maximum, mean, standard deviation, median, first quartile, and third quartile, for each movie's features.

Since the datasets we used in this study encompass many features, it naturally leads to a high dimensionality challenge. We opted for Principal Component Analysis (PCA) as a dimensionality reduction technique across all datasets to address this. This approach significantly reduced the overall dimensionality of the data, as indicated in Tables 2 and 3. While feature selection is a common practice for reducing dimensionality, we would like to provide insight into why we selected PCA over traditional feature selection methods for addressing this concern. Our decision was informed by an extensive exploration of the dataset and its characteristics. We initiated our machine learning pipeline by thoroughly analyzing feature correlations and assessing the individual importance of features in predicting eudaimonic and hedonic scores. Our correlation analysis did not unveil any linear relationships between features and target values. We also conducted feature selection with varying numbers of features (k) as a hyperparameter in a nested cross-fold validation technique using the different machine learning models used in our study. Notably, the results from feature selection exhibited better performance than the baseline but were noticeably inferior to the outcomes achieved through feature reduction. We speculate that the reason can be attributed to the potential information loss inherent in feature selection. Considering that our dataset consists of preprocessed aggregated features from larger datasets, the inherent probability of encountering noise within the features was low. Consequently, the potential risk of information loss through feature selection seemed a valid concern. Employing PCA enabled us to reduce feature dimensions while still preserving a substantial variance in the original features. We attribute the success of PCA to its ability to capture complex relationships across different feature types without sacrificing predictive power.

In order to find a trade-off between the number of dimensions (as low as possible) and variance retained (as high as possible), we performed the PCA analysis on individual and combined feature subsets (as reported in Tables 2 and 3). The inspection of the PCA results showed that retaining 90% of the variance reduced the number of features to an order of magnitude less than the number of data points. This is in line with the general rule of thumb in machine learning that requires roughly 10 data points per feature. We believe this approach strikes a balance between dimensionality reduction and the preservation of variance in data, contributing to the enhanced performance of the models.

3.3. Machine learning pipeline

To predict movies' eudaimonic and hedonic scores, we used various movie features such as low-level and high-level visual features, audio features, and metadata. We approached this prediction task as two binary classification problems: (i) via median splitting of the eudaimonic score and (ii) via median splitting of the hedonic score.

The inputs to the models are preprocessed data, including metadata, audio, and visual features. The machine learning models include linear and non-linear models to investigate which models can more effectively predict E and H scores. The models were trained and tested using the k-fold nested cross-validation repeated five times with different folds. One single iteration of the nested cross-validation approach is shown in Fig. 2. We used ten folds for the outer cross-validation (for evaluating the model's performance) and five for the inner cross-validation (for tuning the model's parameters).

In outer cross-validation, we split the data into 10 folds, keeping one for testing. The model trains on 9 folds and is tested on the left-out test fold. This procedure is repeated within an outer loop, with different folds serving as the test fold in each iteration. The final assessment involves computing average metrics across the 10 distinct test folds. For each test fold, from the 9 training folds, we create 5 subsets for inner cross-validation. One subset becomes the validation set, while the others form the training set for inner hyperparameter tuning. This inner cross-validation trains the model on 4 folds and is evaluated on the left-out validation fold. In the inner cross-validation, we experiment with different parameters to find the best setup for the validation set. We then use this chosen setup on the left-out test set. The model is fine-tuned using various parameters on the validation set. The best parameters are used to assess the model on the reserved test part for the test. We used a search algorithm called Bayesian search to create diverse parameter combinations.

Many candidate machine learning algorithms could perform well on this task. However, using them in our study would imply consuming many resources. So, we decided to make a triage of models to include in the final thorough optimizations and evaluations. In the triage process, we used AutoML tools, did a one-time train–test splitting (80% train set) and ranked several models according to their accuracy of the predicted variables (E and H scores). This exploratory analysis enabled us to create ranked lists of algorithms for each dataset, predicting both E and H scores. Our approach involved compiling the best-performing algorithms from these ranked lists. Notably, we took the top three algorithms for each dataset and prediction task. While evaluating the outputs, we observed that the top-3 algorithms that were above the dummy classifier in the ranked list displayed competitive performance, exhibiting

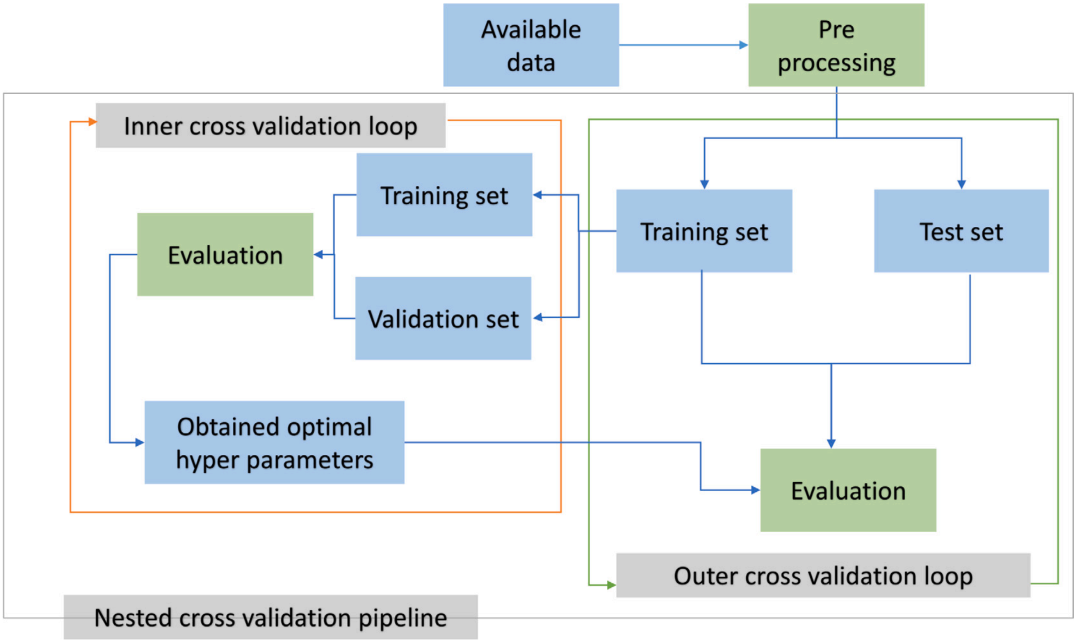

Available data
Pre processing
Inner cross validation loop
Training set
Evaluation
Validation set
Obtained optimal hyper parameters
Training set
Test set
Evaluation
Outer cross validation loop
Nested cross validation pipeline

Fig. 2. The block diagram of a single iteration of the K-fold nested cross-validation machine learning pipeline.

Table 1
ML models and hyperparameters. The optimal parameter will be selected based on the largest accuracy obtained during the inner cross-validation loop. Scikit-learn was used to implement all models (https://scikit-learn.org/stable/).

Model	Parameters	Tested values
LogisticRegression	solver max_iter	newton-cg, lbfgs, sag, saga 1000
RidgeClassifier	alpha	20 equally spaced samples from 1 to 10
SVC	Kernel C gamma	linear , rbf, sigmoid 0.001, 0.01, 0.1, 1.0, 10 20 equally spaced samples from 2^{-15} to 2^{3} (log-scale)
KNeighborsClassifier	n_neighbors	5, 7
DecisionTreeClassifier	min_samples_leaf max_features	4, 5, 6, 7, 8 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9
RandomForestClassifier	min_samples_leaf max_features min_bootstrap	4, 5, 6, 7, 8 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9 True, False

marginal differences in key metrics such as accuracy, F1 score, and ROC AUC. With the knowledge that there was not a single model standing out as a significantly superior choice, we opted to proceed with a selection of foundational and well-known machine learning models encompassing distinct underlying algorithms. Interestingly, among the consistently high-performing models, the decision tree classifier was frequently the best algorithm across various datasets and prediction tasks. Given this consistency, we intentionally included it in our final selection. Our aim was to present a varied set of classification algorithms while also focusing on the model that consistently performed well across different datasets and prediction situations in our initial exploratory analysis.

We employed logistic regression, K-nearest neighbors, Ridge classifier, SVC, decision tree and random forest to predict the E and H scores (Hastie, Tibshirani, Friedman, & Friedman, 2009). We explored different models that incorporate both linear and non-linear approaches. Moreover, we selected a range of hyperparameter values based on the feature sets' dimensions, and findings from previous studies (Probst, Boulesteix, & Bischl, 2019). A detailed list of the algorithms we used and their hyperparameters is available in Table 1.

Furthermore, we also evaluated the performance of deep learning models. In particular, we performed experiments with lightweight architectures of neural networks to avoid overfitting. The performance of the neural network models was inferior to the performance of the several traditional models, although in certain cases, they exhibited the potential to outperform some of the used algorithms. Nonetheless, we acknowledge that the margin of improvement was not substantial. With this observation, we conjectured that the inherent limitations of our dataset size have been a key factor contributing to the relatively modest performance exhibited by deep learning models in our study. In light of these observations, we shifted our attention from deep learning models to traditional machine learning methods for this study.

3.4. Model evaluation

We evaluated the classification prediction performance of our selected models against the majority class classifier (Zangerle & Bauer, 2023). It is important to note that when the positive class is the majority, the base algorithm's performance will always result in a recall of 1 since there will be no false negative cases. Consequently, we used other metrics such as mean accuracy, precision, F1 score, and ROC AUC to better compare the performance of the other algorithms against the baseline. We calculated these measures along with the standard deviation of all folds to be able to interpret the results better.

We repeated the whole k-fold cross-validation five times, each with different folds, and averaged the performance scores. However, we ensured that in each repetition, all models used the same folds to guarantee that the samples were drawn from the same subjects. To choose the right statistical test, we conducted the Shapiro–Wilk test to test if the distribution of dependent variables (i.e., accuracy, precision, F1 score, and ROC AUC) was normal. We found that none of the performance measures followed a normal distribution. Therefore, we used a non-parametric test of the Wilcoxon signed-rank test to calculate the statistical significance between the two models for all dependent variables. We applied the Wilcoxon signed-rank test for the 5 * k performance scores for different models and the base model and determined if the performance scores of the model and the base model differed significantly. Asterisk notation is used in Tables 4 and 5 to visualize statistical significance (*: $p < 0.05$, and **: $p < 0.01$).

4. Results and analysis

This section presents the findings from an exploratory analysis of the movie dataset, the results of the predictive models, and feature importance. We begin with presenting the results of the exploratory analysis, which aims to understand the data and its characteristics better. Next, we report the outcomes of the predictive models used to predict E and H scores compared to the baseline. These models were developed using various datasets, including metadata, audio, and high and low-level visual features. Finally, we investigated which features were more important in the best predictive models.

4.1. Exploratory analysis

We used the dataset of 709 movies collected from 370 users. However, before developing the predictive model, first, we performed a descriptive analysis of the data to gain insight into our data. To explore how movies were distributed across the two dimensions of eudaimonic and hedonic scores we created a scatter heatmap plot as seen in Fig. 3. By investigating the heatmap of movies we observed that most movies have mid-range values for both eudaimonic and hedonic scores.

To visualize the position of movies in E and H score dimensions, we selected a subset and plotted their titles on the graph representing their E and H scores. In Fig. 3, several movies are situated in the top right corner, indicating that they have a high eudaimonic score and a high hedonic score. One such movie is the *1976 comedy-drama* movie, *Network*. The movie is related to the television industry and the media's influence on society but also deals with a comedy theme that justifies its high E and H scores. Another well-known movie is *The Pianist* released in *2002*, an autobiography and the story of a man's survival in Warsaw from 1939 to 1945, has a high E score of 6, but a low H score of around 4, indicating the movie evokes more sadness and thought-provoking moments than a fun experience. Another movie, *The Cable Guy*, released in *1996*, is a *comedy-drama* movie featuring *Jim Carrey*. The movie conveys deep messages despite its fun content, which justifies its genre. However, the movie is generally cheerful, explaining its high H and low E scores as perceived by the participants. Nonetheless, as seen in Fig. 3, the plotted movies align more or less with their content.

4.2. Results of the predictive models

In Section 3, we outlined a machine learning pipeline that we developed to predict the classification of high/low E scores and high/low H scores. To define the E and H classes, we performed a median split on E and H scores. To predict E and H classes of movies we used various datasets including metadata, audio features, low-level visual features, and high-level visual features. Table 2 presents the number of movies in the intersection of the EH dataset with the other datasets and the number of available features for each dataset. Prior to feeding the features into our machine learning models, we conducted principal component analysis (PCA) to reduce the dimensionality of features and avoid the risk of overfitting. This allowed us to keep only the most informative components that accounted for at least 90% of the variance in our data. Table 2 includes the number of components that have 90% of the variance for each dataset.

To evaluate the efficiency of combining datasets for predicting EH scores, we aggregated the datasets listed in Table 3, each of which is represented by a letter corresponding to its modality. More specifically, "A" represents audio, "M" represents metadata, "L" represents low-level visual features, and "H" represents high-level visual features. For example, the combination of low-level visual features, high-level visual features, audio, and metadata features is denoted as "LHAM", while the combination of low-level visual features and audio features is denoted as "LA". In Table 3, we provide a detailed overview of the combined datasets, including the number of available movies, features, and the number of components that keep 90% of variance for each dataset. We explored all possible combinations of two, three, and four datasets to evaluate the predictive power of different feature combinations.

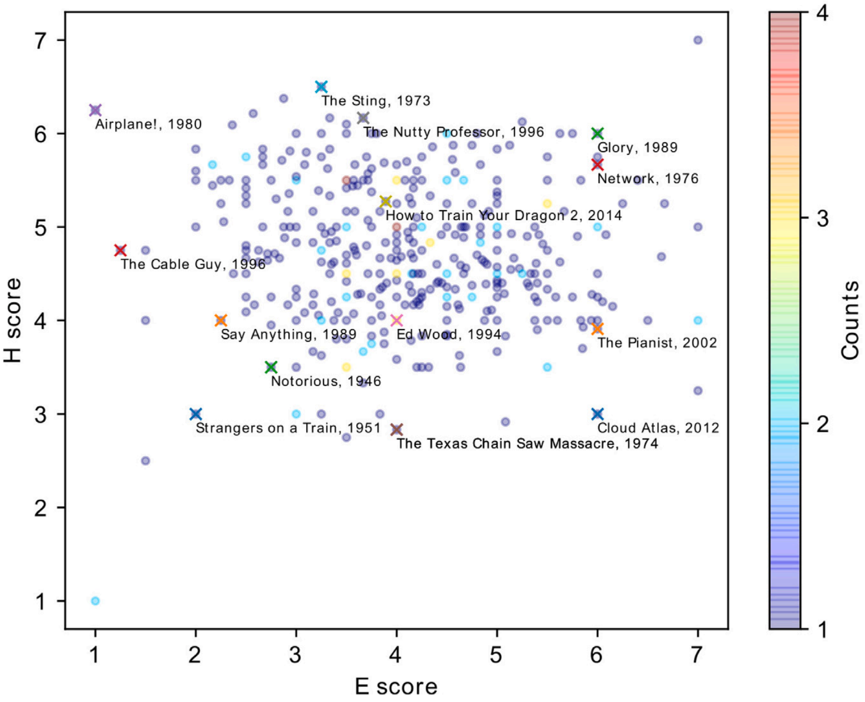


Fig. 3. Scatter plot of eudaimonic (E) and hedonic (H) scores for the dataset of 709 movies. The E score is shown on the x-axis and the H score on the y-axis. Some of the data points are labeled in the lower and right positions with the title of the movie and its release year.

Table 2
Number of movies, features, and the number of principal components retained in PCA analysis to capture 90% of the variance in each dataset.

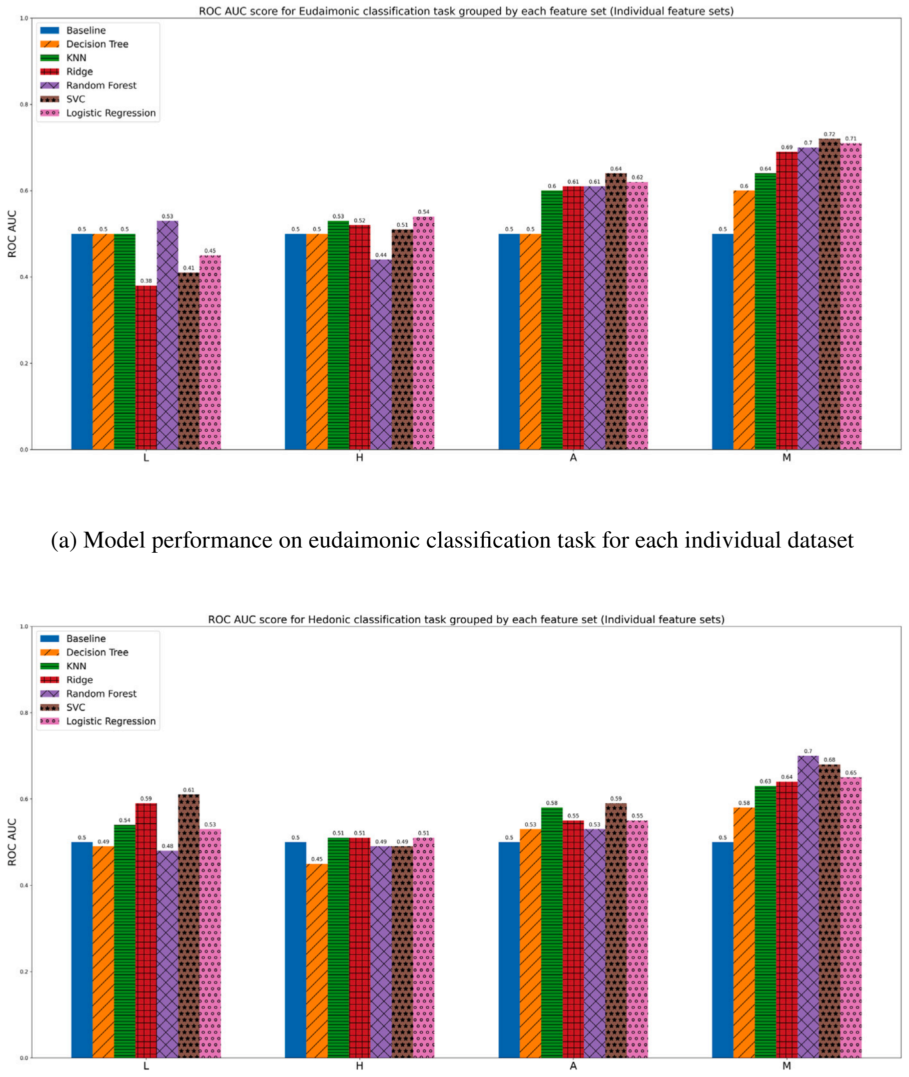
Dataset	Description	Number of movies	Number of features	Number of components (PCA)
L	Low-level visual	400	167	16
H	High-level visual	307	77	15
A	Audio	617	151	26
M	Metadata	709	168	95

Table 3
Number of movies, features, and the number of principal components retained in PCA analysis to capture 90% of the variance in combined datasets.

Dataset	Number of movies	Number of features	Number of components (PCA)
LA	332	318	40
LH	307	244	30
AM	611	319	109
LM	394	335	80
HM	303	245	67
HA	245	228	34
LHA	245	395	46
HAM	243	396	70
LHM	303	412	74
LAM	329	486	89
LHAM	243	243	77

We evaluated the performance of multiple machine-learning models using the datasets listed in Tables 2 and 3. Our evaluation was based on key metrics such as mean accuracy, precision, F1 score, and ROC AUC, which were computed for all folds alongside their corresponding standard deviations. To prevent overfitting and ensure consistency across datasets, we applied PCA to identify the most informative components that accounted for at least 90% of the variance in our data before feeding them into the machine learning pipeline. Additional details on the data and models used in this study can be found in Section 3.

Fig. 4 presents the performance comparison of various models with the base algorithm based on ROC AUC for different datasets, including metadata (M), audio (A), low-level visual features (L), and high-level visual features (H). The base algorithm achieved a ROC AUC of 0.5.



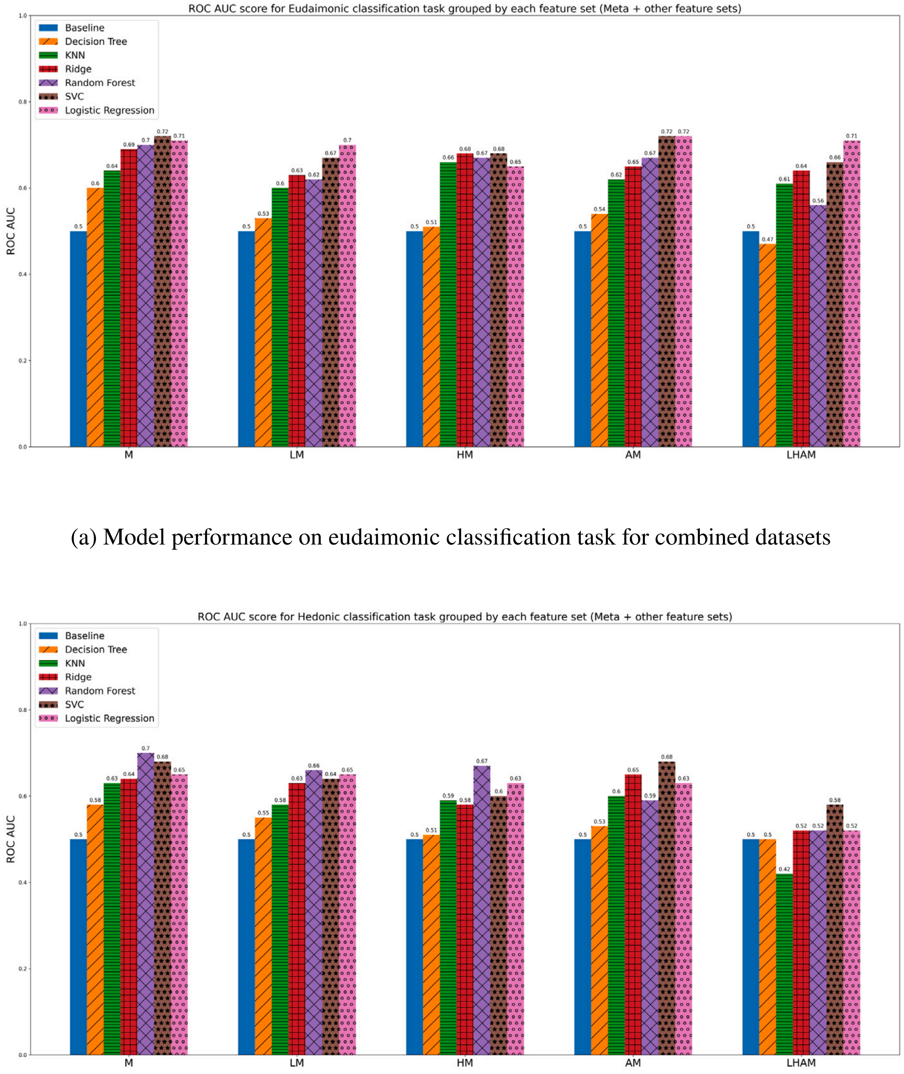
(a) Model performance on eudaimonic classification task for each individual dataset

(b) Model performance on hedonic classification task for each individual dataset

Fig. 4. Model performance on eudaimonic and hedonic classification task for each individual dataset.

Among the different datasets, using metadata resulted in considerable performance improvements for all machine learning models, with at least an 8% improvement in both eudaimonic and hedonic classification problems. In the eudaimonic classification problem, the support vector classifier (SVC) outperformed the baseline the most, with a ROC AUC of 0.72, representing a 22% improvement. Similarly, in the hedonic classification problem, the ROC AUC achieved by SVC was 0.68, which is 18% more than the baseline. In the case of using the metadata dataset, the Support Vector Classification (SVC) model achieved the highest ROC AUC of 0.72 (22% improvement) for the eudaimonic classification problem, while the random forest algorithm achieved the maximum ROC AUC of 0.70 (20% improvement) for the hedonic classification problem, outperforming all other models evaluated.

The audio dataset showed that SVC achieved the best performance among different models for both eudaimonic and hedonic classification problems, with ROC AUC values of 0.64 and 0.59, respectively. The high-level visual features dataset resulted in a performance close to the baseline compared to other datasets. The logistic regression model outperformed other models in this dataset, although with only a slight improvement. Similarly, the dataset of low-level visual features showed worse or slightly better performance for most of the models. These results suggest that the visual features used in this study may not be sufficient to predict the eudaimonic and hedonic classes with a significant improvement compared to the baseline. However, we further explored if combining these features with other features could enhance performance.



(a) Model performance on eudaimonic classification task for combined datasets

(b) Model performance on hedonic classification task for combined datasets

Fig. 5. Model Performance on eudaimonic and hedonic classification task for merged datasets.

We conducted a thorough investigation of all possible combinations of dataset features and identified the most effective combined datasets, including metadata, as demonstrated in Fig. 5. For the sake of simplicity, we only present the results of the best-performing combinations. While combining other datasets with metadata can improve performance compared to using them individually, the metadata dataset alone yielded superior results to other datasets. Additional metrics and the results of the Wilcoxon signed-rank test on the performance of various machine learning models with the metadata dataset compared to the baseline are presented in Tables 4 and 5.

The results in Table 4 show that the best model in predicting eudaimonic class using metadata outperforms the baseline with a 21% increase in accuracy, 30% increase in precision, 16% increase in F1 score, and 22% increase in ROC AUC. All of these improvements are statistically significant, with p-values lower than at least 0.05. Similarly, the results presented in Table 5 demonstrated that the models' performances in predicting the hedonic class using metadata were significantly better than the baseline. Specifically, the best models achieved a 15% increase in accuracy, a 32% increase in precision, a 17% increase in F1-score, and a 20% increase in ROC AUC. All of these performance values were confirmed to be statistically significantly higher than the baseline, indicating the effectiveness of the proposed models in accurately predicting the hedonic class.

The findings demonstrate that the developed models outperformed the base algorithm in predicting the E and H classes across various datasets, such as metadata, audio, high and low-level visual features. Notably, the metadata dataset exhibited superior

Table 4
Prediction results for eudaimonic class using metadata. Results obtained from nested cross-validation with 10 outer and 5 inner splits, repeated five times. The classification problem involves two classes: (a) High_E (high eudaimonic class), and (b) Low_E (low eudaimonic class). Mean values of five repetitions on 10 outer splits are presented with standard deviation values in parentheses.

ML algorithm	Accuracy	Precision	F1 score	ROC AUC
Base	0.47 (0.05)	0.39 (0.20)	0.52 (0.26)	0.50 (0.04)
Logistic regression	0.66 (0.05)**	0.67 (0.08)**	0.65 (0.06)*	0.71 (0.04)**
Ridge	0.68 (0.05)**	0.69 (0.06)**	0.68 (0.04)*	0.69 (0.05)**
SVC	0.67 (0.05)**	0.69 (0.09)**	0.67 (0.07)**	0.72 (0.04)**
KNN	0.60 (0.05)**	0.63 (0.11)**	0.58 (0.07)*	0.64 (0.06)**
Decision tree	0.58 (0.04)	0.60 (0.06)	0.57 (0.05)	0.60 (0.05)*
Random forest	0.65 (0.02)**	0.65 (0.04)**	0.65 (0.03)	0.70 (0.05)**

* Notation is used to visualize statistical significance $p < 0.05$.

** Notation is used to visualize statistical significance $p < 0.01$.

Table 5
Prediction results for hedonic class using metadata. Results obtained from nested cross-validation with 10 outer and 5 inner splits, repeated five times. The classification problem involves two classes: (a) High_H (high hedonic class), and (b) Low_H (low hedonic class). Mean values of five repetitions on 10 outer splits are presented with standard deviation values in parentheses.

ML algorithm	Accuracy	Precision	F1 score	ROC AUC
Base	0.48 (0.03)	0.34 (0.23)	0.46 (0.30)	0.50 (0.00)
Logistic regression	0.60 (0.03)**	0.61 (0.06)**	0.59 (0.06)*	0.65 (0.04)**
Ridge	0.59 (0.05)**	0.59 (0.09)**	0.59 (0.06)	0.64 (0.06)**
SVC	0.63 (0.08)**	0.65 (0.11)**	0.60 (0.09)	0.68 (0.07)**
KNN	0.60 (0.05)**	0.61 (0.09)**	0.59 (0.06)	0.63 (0.05)**
Decision tree	0.56 (0.06)**	0.57 (0.08)**	0.53 (0.09)	0.58 (0.04)*
Random forest	0.64 (0.06)**	0.66 (0.10)**	0.63 (0.08)*	0.70 (0.08)*

* Notation is used to visualize statistical significance $p < 0.05$.

** Notation is used to visualize statistical significance $p < 0.01$.

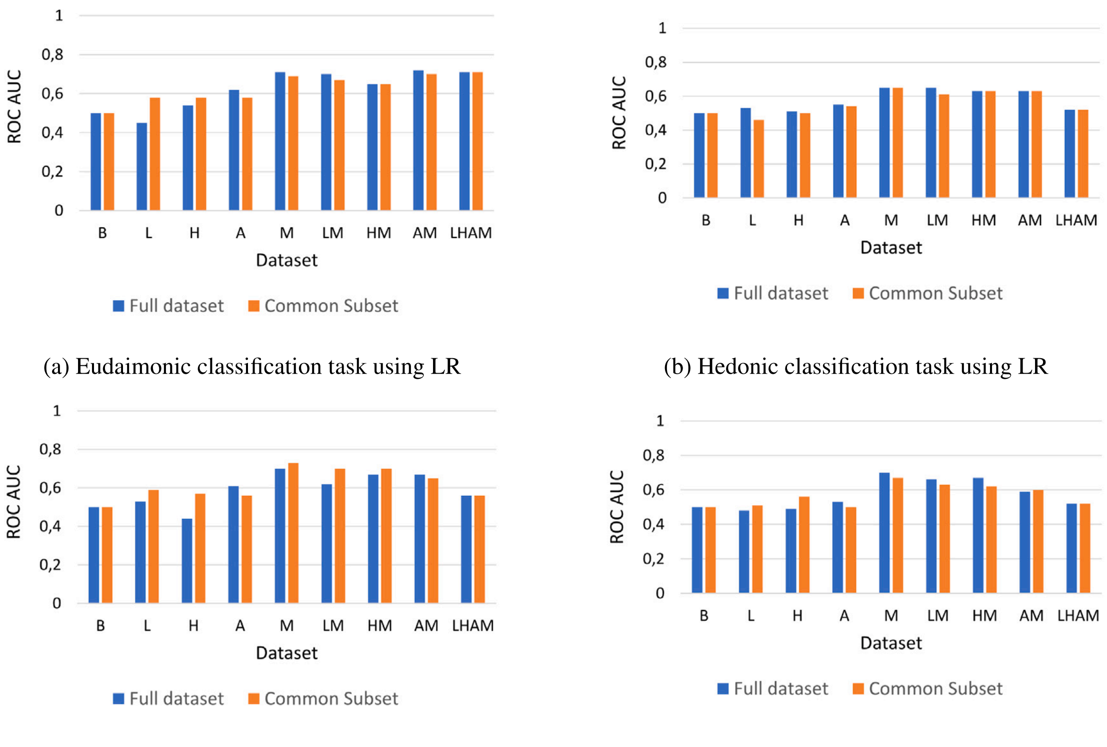
predictive performance compared to the other datasets for the E and H classes. These results suggest that incorporating metadata features into the models could be a promising approach to improve their performance in predicting the E and H classes.

Thus far, we have observed the top-performing models across various dataset combinations in comparison to the baseline. As indicated in Tables 2 and 3, there is variability in the number of movies across datasets that makes it difficult to conclusively attribute performance differences among different feature sets solely to feature variations or potentially inadequate data points. To address this concern and assess the contribution of the feature sets to the models' performance, we conducted additional experiments to ensure consistency. By using the same set of 243 movies from the smallest dataset (i.e. LHAM) across all datasets, we aimed to mitigate size-related concerns. Subsequently, we trained and evaluated the models following the same pipeline used for larger datasets.

After thoroughly examining results across diverse algorithms and dataset combinations, we found no substantial differences in the models' performance. These variances had a negligible impact on the overall performance of the models compared to the larger datasets. While we explored all the algorithms discussed in the paper across various dataset combinations, we opted for a focused presentation in line with the paper structure to maintain clarity and avoid information overload in a single figure. We presented the results for logistic regression and random forest, given these algorithms performed superior in most cases. In Fig. 6, the eudaimonic and hedonic classification performance is shown for different feature sets using random forest (RF) and logistic regression (LR). The blue bar chart represents the model performance (measured by ROC AUC) using all available movies for each dataset, while the orange bar chart represents the model performance for the smallest common subset of movies in all datasets. Consistent with the approach used for larger datasets, we applied the Wilcoxon signed-rank test to assess the performance scores of various models and the base model when using common subsets of movies across all datasets. This aimed to determine if the performance scores of the models significantly differ from those of the base model (Details of the statistical test can be found in Section 3.4). Using the subset of common movies instead of the dataset with all movies maintained a consistent outcome in the statistical test, indicating that the values that were initially found to be significantly different remained so. While some metrics saw an increase in p-values from $p < 0.01$ to $p < 0.05$, none exceeded 0.05, underscoring the overall stability and consistency in the outcomes of the statistical tests.

4.3. Feature importance

After analyzing the results presented in the previous section, we found that metadata can provide predictor features for eudaimonic and hedonic scores. We employed a permutation feature importance method to identify the most effective metadata variables in predicting the E and H scores. The list of metadata features is provided in 3.1. We used permutation feature importance, which involves shuffling a single feature value to assess the decrease in model score. The greater the change in score, the more predictive the feature is. We used a random forest model as it was one of the best-performing algorithms in predicting E and H



(a) Eudaimonic classification task using LR

(b) Hedonic classification task using LR

(c) Eudaimonic classification task using RF

(d) Hedonic classification task using RF

Fig. 6. Comparison of model performance in eudaimonic and hedonic classification tasks: Logistic regression (LR) and random forest (RF) models on full datasets (i.e. datasets with all movies) versus common subset datasets (i.e. datasets with common subset of movies). The leftmost bar in the charts is the baseline model represented by the letter B in the figures. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

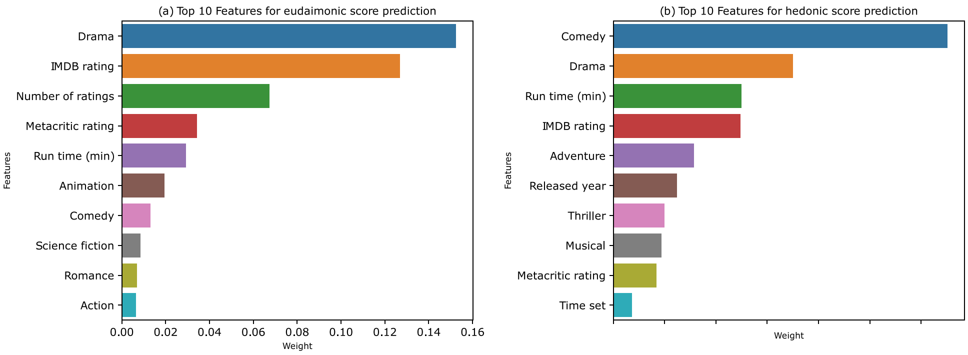


Fig. 7. Permutation importance scores of the top 10 features for predicting eudaimonic and hedonic scores. The left plot (a) shows the importance scores for predicting eudaimonic scores, while the right plot (b) shows the importance scores for predicting hedonic scores. For calculating the permutation importance scores, a random forest regression model was trained on 80% of the randomly sampled data.

scores in our study. Then, we evaluated the importance of features in predicting eudaimonic and hedonic scores. As shown in Fig. 7, comedy and drama genres are among the most important features for predicting E and H scores in line with expectations. We expected that drama genres should receive a higher E score, while comedy movies should receive a higher H score, as drama movies are associated with serious subjects and comedy movies are more cheerful and fun. Adventure, thriller, and musical genres also appeared in the top ten predictive features for hedonic scores. For eudaimonic scores, the other important genres, apart from comedy and drama, were animation, science fiction, romance, and action. Additionally, we found that the Metacritic rating was more predictive of the eudaimonic score, while the IMDB rating was more predictive of the hedonic score, which is an interesting finding.

Several genres, including comedy, drama, adventure, thriller, musical, animation, action, science fiction (Sci-Fi), and romance, were found to be important features in predicting E and H scores. To further investigate the correlation of these genres with E

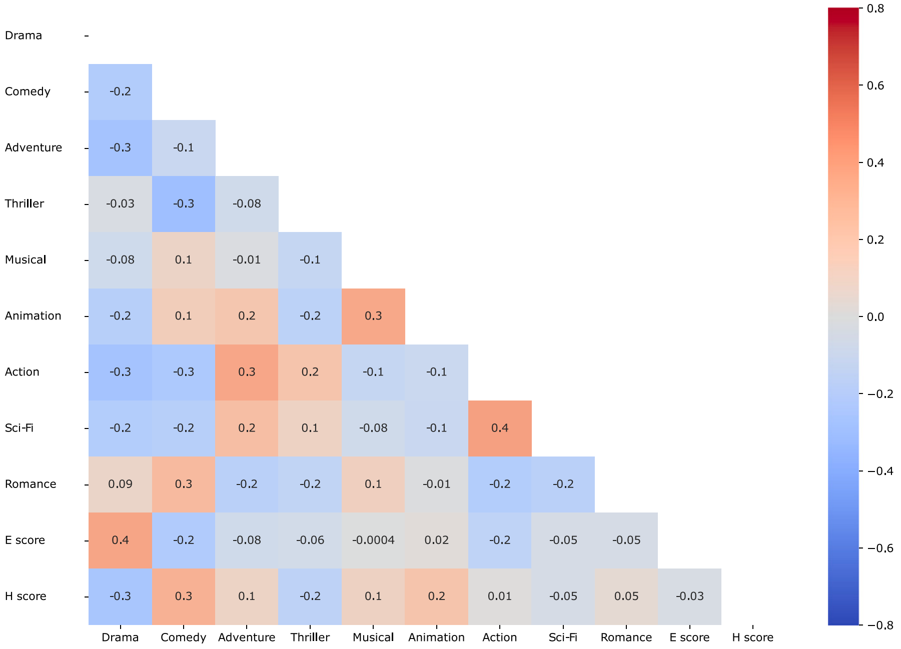


Fig. 8. Heatmap displaying the correlation between selected movie genres and their corresponding E score and H scores. The selected movie genres are the genres listed as the top 10 features sorted by permutation importance score in Fig. 7.

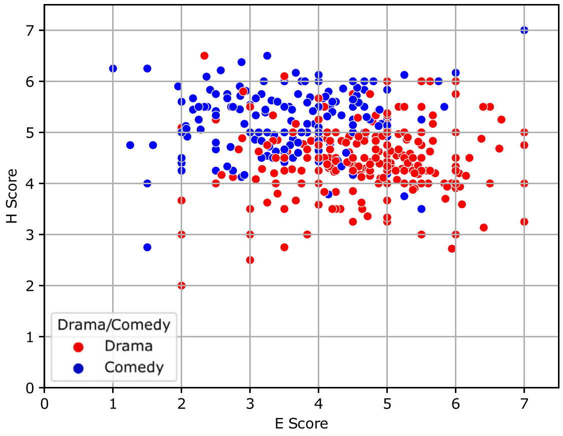


Fig. 9. Scatter plot of eudaimonic (E) and hedonic (H) scores for the movies of genre Comedy or Drama. The E score is shown on the x-axis and the H score on the y-axis.

and H scores, we created a heatmap in Fig. 8 that depicts the correlation of all features. Our findings showed that only drama was positively correlated with E scores, while comedy and action were negatively correlated with E scores. Additionally, comedy, adventure, musical, and animation genres were positively correlated with H scores, whereas drama and thriller were negatively correlated with H scores. These results offer valuable insights into the relationship between different genres and E or H scores.

The importance of drama and comedy in predicting both E and H scores is highlighted by their high permutation importance scores. To further explore their relationship with the two dimensions, we plotted their distributions across E and H in Fig. 9. The plot clearly shows a visible separation between drama and comedy along the E and H axes.

5. Discussion and implications

The results of this study reveal that our proposed models outperform the baseline, as confirmed by various evaluation metrics. Specifically, the findings demonstrate that incorporating metadata features improves the prediction accuracy of the models. This success highlights the potential value of using machine learning-based models for automatically labeling movies' E and H scores.

Given the scale of the movie dataset and the number of features used in this study, the concern of overfitting naturally arises. It is important to highlight that we were aware of this potential challenge and took proactive measures to construct our machine-learning pipeline thoughtfully, aiming to reduce the risk of overfitting. Among these precautions, lies our careful selection of machine learning models. To reduce the overfitting risk, we adopted the following strategies:

- *Regularization techniques*: Regularization is employed in several machine learning models, such as logistic regression, ridge classifier, and SVC. In our logistic regression model, we implemented "L2" regularization, which adds a penalty term proportional to the square of the model's parameters. This regularization approach leads to a simpler model less prone to overfitting. For the Ridge classifier, "alpha" parameter determines the regularization strength in the case of assigning a positive value for it similar to parameter "c" of "SVC" as a regularization term. Table 1 illustrates the different positive values of the alpha parameter used for the Ridge Classifier and "c" parameter for "SVC". The hyperparameter tuning incorporated in a cross-fold validation method attempts to use the best level of regularization for the models.
- *Dimensionality reduction*: Given the limited scale of movies in our dataset compared to the number of features, we employed PCA to reduce the feature dimensionality. This choice enabled us to effectively reduce data dimensions while retaining a substantial portion of its variance (i.e. 90%). The 90% variant in data leads to roughly one-tenth of the original features in terms of size (see Tables 2 and 3), underscoring our dedication to enhancing model performance and alleviating overfitting risks.
- *Nested cross-validation technique*: We employed a cross-validation technique to evaluate the model's performance on different subsets of data. Through different repetitions, we observed variations in the performance of different folds. While we did not explicitly report the performance of the model on validation sets during hyperparameter tuning, we ensured that our model's performance on the training set was not significantly better than on the test set. Since no such scenario was observed, we assumed the overfitting likelihood was low.
- *Model selection*: In addition to models incorporating regularization, we implemented the k-nearest neighbor classifier with carefully chosen values of k as a hyperparameter. Likewise, when it comes to decision trees, the cautious selection of parameters tied to pruning the tree can significantly decrease the likelihood of overfitting. For example, higher values of "min_samples_leaf", which determines the minimum number of features in a leaf for a split to occur, prevent the model from capturing noise and outliers, yielding simpler models. Additionally, we used the power of the random forest model, an ensemble technique that combines predictions from multiple models, effectively reducing overfitting risks.

Our work showed that the features we used contain information that is useful for predicting the E and H scores. However, there is still room for improvement. The obvious thing is attempting to achieve a higher accuracy by increasing the data and complexity of the predictors. Given enough data, it is always possible to label more movies, use more modalities for feature extraction (e.g. subtitles) and spend more resources on more complex algorithms.

Our results are likely to be improved if we use the full-length versions of the movies instead of the trailers for audio and visual features. There are two potential reasons for that. The first is the mere shortness of the trailers compared to the full-length versions, which results in less data available for feature extraction. The second reason is that trailers are mini-movies in themselves that are meant to attract the candidate audience and do not disclose the full storyline, making them less representative of the full-length version. Due to these factors, trailers tend to be more similar to each other than their full-length counterparts. However, legal and computational limitations prevented us from working on full-length movies. Further research could explore the use of full-length movies to enhance the accuracy of feature extraction and the overall performance of machine learning models for predicting E and H scores.

Finally, the prediction accuracy in itself is just a research metric for the prediction problem we addressed. As stated earlier, the goal of this line of research is to do user modeling for recommending movies. Hence, the final verdict of the practical usefulness of the proposed method is the improvement of the recommender system that uses E and H scores as features. We leave this to our future work, as outlined in Section 6.

The proposed work has several theoretical and practical implications. One such implication is the potential for recommender systems to incorporate E and H concepts to improve recommendations. Specifically, these concepts can be used to develop explainable recommender systems that draw from extensive psychological research on eudaimonic and hedonic experiences. Given the psychological roots of these concepts, they are easily understood by users, which enables meaningful explanations to be provided to end-users. This could enhance transparency and trust in the recommendations and ultimately improve user satisfaction (Hadash, Willemsen, Snijders, & Ijsselsteijn, 2022; Musto et al., 2022). By establishing an automated model capable of predicting these attributes for movies, we not only address an immediate need but also pave the way for future investigations within the realm of explainable recommender systems.

The application of machine learning-based predictor enabled us to uncover previously undiscovered patterns and correlations between different movies' characteristics (i.e. visual characteristics, audio, and metadata) and their eudaimonic and hedonic scores. These insights have important implications for explainable recommender systems, particularly in using the diverse modalities inherent to movies. The successful adaptation of machine learning models to our domain has broader implications beyond the immediate scope of our domain. It complements the existing work in predicting various user factors from behavioral and content data. Furthermore, the extensive research surrounding the modeling of emotions and personality traits can synergize with the outcomes of our study. The parallel between emotions and the eudaimonic and hedonic scores lies in their shared contribution to characterizing movies based on users' perceptions. This connection opens the door for future research that combines these ideas, creating an interesting path to explore.

6. Conclusions and future works

There is a growing need to provide interpretable reasoning in movie recommender systems and contribute to explainable recommender systems. Specifically, eudaimonic and hedonic qualities have proven to be easily understandable by users and used for modeling users and items. However, achieving such explainable movie recommender systems requires E and H labels for movie items. As acquiring these scores through the existing tools is time-consuming, we explored if we could create a predictive model to predict the E and H scores of movies automatically. We developed an automatic machine-learning method for labeling movies with eudaimonic and hedonic scores using audio features, metadata, and low-level and high-level visual features.

We found that various combinations of movie features could outperform the baseline in predicting eudaimonic and hedonic scores, with metadata features proving to be the most effective. Among the metadata features, genre was identified as one of the most important predictors for E and H scores. The fact that the most informative metadata features are semantically rich highlights a compelling challenge: the need to disentangle these features into sub-dimensions and identify which sub-dimensions contribute most to the overall variance. Addressing this challenge is not a straightforward task, as it necessitates additional steps such as further data collection to establish ground truth for disentangled features and experimental exploration with predictive models.

The contributions of our study are as follows: (i) To the best of our knowledge, this is the first work to propose an automatic model to predict the eudaimonic and hedonic scores of movies on a scale of 709 movies, surpassing previous works that explored this for fewer movies (i.e. 30 movies). (ii) This is the first work that leverages various features (i.e., audio, low-level and high-level visual features and metadata) to propose an automatic model to predict the eudaimonic and hedonic scores of movies. (iii) We present the first study that reports correlations between several genres and movies' E and H scores. Additionally, we explored the predictive power of several metadata features, including genres, in predicting E and H scores through exploratory data analysis. In conclusion, this study has demonstrated the potential benefits of incorporating several movie features in predicting movies' E and H scores. However, this study has some limitations, which are discussed in more detail in Section 5 and need to be addressed in future research. One limitation is that audio and visual features were extracted from trailers of movies rather than full-length movies. Therefore, future studies could benefit from using full-length movies to extract features for more accurate prediction models.

Additionally, collecting more data would allow for the investigation of more complex models. In the future, we plan to explore the impact of additional features, such as shot types and the presence of audio signals over time, as well as more metadata, such as the director of the movie, on machine learning models' performance. We also intend to investigate how textual information, such as subtitles, Metacritic reviews, and IMDB reviews, may predict E and H scores of movies. Further investigation could provide a more comprehensive understanding of the factors that influence the performance of the predictive models and investigate the practical usefulness of the proposed method in recommender systems.

CRediT authorship contribution statement

Elham Motamedi: Conceptualization, Data curation, Formal analysis, Methodology, Software, Visualization, Writing – original draft. **Danial Khosh Kholgh:** Software, Validation, Visualization. **Sorush Saghari:** Software, Visualization. **Mehdi Elahi:** Conceptualization, Writing – review & editing. **Francesco Barile:** Writing – review & editing. **Marko Tkalcic:** Conceptualization, Data curation, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgment

This work was supported by the CogniCom grant (ID SM13103) funded by the University of Primorska.

References

Allamanche, E., Hellmuth, O., Fröba, B., Kastner, T., & Cremer, M. (2001). Content-based identification of audio material using MPEG-7 low level description. *System*, *8*, 197–204, http://ismir2001.ismir.net/pdf/allamanche.pdf.

Atrey, P. K., Hossain, M. A., El Saddik, A., & Kankanhalli, M. S. (2010). Multimodal fusion for multimedia analysis: A survey. 16(6), 345–379. http://dx.doi.org/10.1007/s00530-010-0182-0.

Beheshti, A., Ghodratnama, S., Elahi, M., & Farhood, H. (2022). *Social data analytics*. CRC Press.

Botella, C., Riva, G., Gaggioli, A., Wiederhold, B. K., Alcaniz, M., Baños, R. M., et al. (2012). The present and future of positive technologies. *Cyberpsychology, Behavior, and Social Networking*, *15*(2), 78–84. http://dx.doi.org/10.1089/cyber.2011.0140.

Brezeale, D., & Cook, D. J. (2008). Automatic video classification: A survey of the literature. *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, *38*(3), 416–430. http://dx.doi.org/10.1109/TSMCC.2008.919173.

Cerf, M., Harel, J., Einhäuser, W., & Koch, C. (2008). Predicting human gaze using low-level saliency combined with face detection. In *Advances in neural information processing systems 20 - Proceedings of the 2007 conference* (pp. 1–8).

De Ridder, A., Vandebosch, P. D. H., & Dhoest, P. D. A. (2022). Examining the hedonic and eudaimonic entertainment experiences of the combination of stand-up comedy and human-interest. *Poetics*, *90*(December 2020), Article 101601. http://dx.doi.org/10.1016/j.poetic.2021.101601.

Deldjoo, Y., Elahi, M., Cremonesi, P., Garzotto, F., Piazzolla, P., & Quadrana, M. (2016). Content-based video recommendation system based on stylistic visual features. *Journal on Data Semantics*, *5*(2), 99–113. http://dx.doi.org/10.1007/s13740-016-0060-9.

Elahi, M., Bakhshandegan Moghaddam, F., Hosseini, R., Rimaz, M. H., El Ioini, N., Tkalcic, M., et al. 2021. Recommending videos in cold start with automatic visual tags. (pp. 54–60). http://dx.doi.org/10.1145/3450614.3461687.

Gong, Y., & Xu, W. (2007). *Machine learning for multimedia content analysis, vol. 30*. Springer Science & Business Media.

Hadash, S., Willemsen, M. C., Snijders, C., & Ijsselsteijn, W. A. (2022). Improving understandability of feature contributions in model-agnostic explainable AI tools. In *Conference on human factors in computing systems - proceedings*. http://dx.doi.org/10.1145/3491102.3517650.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction, vol. 2*. Springer.

Hrustanovic, S., Kavsek, B., & Tkalcic, M. (2021). Recognition of eudaimonic and hedonic qualities from song lyrics. In *CEUR workshop proceedings, vol. 3054* (pp. 45–53).

Hu, W., Xie, N., Li, L., Zeng, X., & Maybank, S. (2011). A survey on visual content-based video indexing and retrieval. *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, *41*(6), 797–819. http://dx.doi.org/10.1109/TSMCC.2011.2109710.

Hunt, N. (2014). Quantifying the value of better recommendations. https://recsys.acm.org/recsys14/keynotes/.

Kahneman, D., Diener, E., & Schwarz, N. (1999). *Well-being: Foundations of hedonic psychology*. Russell Sage Foundation.

Lew, M. S., Sebe, N., Djeraba, C., & Jain, R. (2006). Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions on Multimedia Computing, Communications and Applications*, *2*(1), 1–19. http://dx.doi.org/10.1145/1126004.1126005.

Li, W., Zhang, L., Li, C., Zhu, N., Zhao, J., & Kong, F. (2022). Pursuing pleasure or meaning: A cross-lagged analysis of happiness motives and well-being in adolescents. *Journal of Happiness Studies*, *23*(8), 3981–3999. http://dx.doi.org/10.1007/s10902-022-00576-5.

Liu, P., Zhang, L., & Gulla, J. A. (2020). Dynamic attention-based explainable recommendation with textual and visual fusion. *Information Processing and Management*, *57*(6), Article 102099. http://dx.doi.org/10.1016/j.ipm.2019.102099.

Moghaddam, F. B., Elahi, M., Hosseini, R., Trattner, C., & Tkalcic, M. (2019). Predicting movie popularity and ratings with visual features. http://dx.doi.org/10.1109/SMAP.2019.8864912.

Montalvo-Lezama, R., Montalvo-Lezama, B., & Fuentes-Pineda, G. (2022). Trailers12k: Improving transfer learning with a dual image and video transformer for multi-label movie trailer genre classification. *SSRN Electronic Journal*, *60*(3), Article 103343. http://dx.doi.org/10.2139/ssrn.4253487.

Motamedi, E., Barile, F., & Tkalčič, M. (2022). Prediction of eudaimonic and hedonic orientation of movie watchers. *Applied Sciences (Switzerland)*, *12*(19), http://dx.doi.org/10.3390/app12199500.

Motamedi, E., & Tkalcic, M. (2021). Prediction of eudaimonic and hedonic movie characteristics from subtitles. 3054, 54–61.

Musto, C., Delic, A., Inel, O., Polignano, M., Rapp, A., Semeraro, G., et al. (2022). Workshop on explainable user models and personalised systems (ExUM). In *UMAP2022 - Adjunct proceedings of the 30th ACM conference on user modeling, adaptation and personalization* (July), (pp. 160–162). http://dx.doi.org/10.1145/3511047.3536350.

Nguyen, T., Scholer, A. A., Miele, D. B., Edwards, M. C., & Fujita, K. (2022). Predicting academic performance with an assessment of students' knowledge of the benefits of high-level and low-level construal. *Social Psychological and Personality Science*, http://dx.doi.org/10.1177/19485506221090051.

Oliver, M. B., & Raney, A. A. (2011). Entertainment as pleasurable and meaningful: Identifying hedonic and eudaimonic motivations for entertainment consumption. *Journal of Communication*, *61*(5), 984–1004. http://dx.doi.org/10.1111/j.1460-2466.2011.01585.x.

Probst, P., Boulesteix, A. L., & Bischl, B. (2019). Tunability: Importance of hyperparameters of machine learning algorithms. *Journal of Machine Learning Research*, *20*, 1–32.

Rasheed, Z., & Shah, M. (2003). *Video categorization using semantics and semiotics* (pp. 185–217). Boston, MA: Springer US, http://dx.doi.org/10.1007/978-1-4757-6928-9_7.

Rasheed, Z., Sheikh, Y., & Shah, M. (2005). On the use of computable features for film classification. *IEEE Transactions on Circuits and Systems for Video Technology*, *15*(1), 52–63. http://dx.doi.org/10.1109/TCSVT.2004.839993.

Rimaz, M. H., Elahi, M., Moghadam, F. B., Trattner, C., Hosseini, R., & Tkalčič, M. (2019). Exploring the power of visual features for the recommendation of movies. (pp. 303–308). http://dx.doi.org/10.1145/3320435.3320470.

Rimaz, M. H., Hosseini, R., Elahi, M., & Moghaddam, F. B. (2021). AudioLens: Audio-aware video recommendation for mitigating new item problem. 12632 LNCS, 365–378. http://dx.doi.org/10.1007/978-3-030-76352-7_35.

Tkalčič, M., & Ferwerda, B. (2018). Eudaimonic modeling of moviegoers. In *Proceedings of the 26th conference on user modeling, adaptation and personalization* (pp. 163–167). Singapore Singapore: ACM, http://dx.doi.org/10.1145/3209219.3209249.

Wang, Y., Xing, C., & Zhou, L. (2006). Video semantic models: Survey and evaluation. *International Journal of Computer Science and Network Security (IJCSNS)*, *6*(2), 10–20.

Waterman, A. S. (1993). Two conceptions of happiness: Contrasts of personal expressiveness (eudaimonia) and hedonic enjoyment. *Journal of Personality and Social Psychology*, *64*(4), 678–691. http://dx.doi.org/10.1037/0022-3514.64.4.678.

Xi, D., Xu, W., Chen, R., Zhou, Y., & Yang, Z. (2021). Sending or not? A multimodal framework for Danmaku comment prediction. *Information Processing and Management*, *58*(6), Article 102687. http://dx.doi.org/10.1016/j.ipm.2021.102687.

Yang, B., Mei, T., Hua, X. S., Yang, L., Yang, S. Q., & Li, M. (2017). Online video recommendation based on multimodal fusion and relevance feedback. (pp. 73–80). http://dx.doi.org/10.1145/1282280.1282290.

Zangerle, E., & Bauer, C. (2023). Evaluating recommender systems: Survey and framework. *ACM Computing Surveys*, *55*(8), 1–38. http://dx.doi.org/10.1145/3556536.

Zhao, X., Li, G., Wang, M., Yuan, J., Zha, Z. J., Li, Z., et al. (2011). Integrating rich information for video recommendation with multi-task rank aggregation. Number November, (pp. 1521–1524). http://dx.doi.org/10.1145/2072298.2072055.

Zhou, H., Hermans, T., Karandikar, A. V., & Rehg, J. M. (2010). Movie genre classification via scene categorization. In *Proceedings of the 18th ACM international conference on multimedia* (pp. 747–750).