IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Ingénierie des Systèmes d'Information
Vol. 30, No. 9, September, 2025, pp. 2375-2392
Journal homepage: http://iieta.org/journals/isi

A Systematic Review of Machine Learning Methods for Movie Genre Classification and Age-Appropriateness Prediction

Yaseen K. Abbas*, Ahmed Al-Azawei

College of Information Technology, University of Babylon, Babylon 51002, Iraq

Corresponding Author Email: yaseenkudhaira.sw@student.uobabylon.edu.iq

https://doi.org/10.18280/isi.300914

Received: 3 June 2025
Revised: 12 August 2025
Accepted: 19 August 2025
Available online: 30 September 2025

Keywords:
movie genre classification, age rating, machine learning, deep learning, systematic literature review, multimodal classification, PRISMA

ABSTRACT

Movies are one important source of entertainment, cultural dissemination, education, and evoking strong emotions, whether negative or positive. Therefore, accurate film classification has grown in importance with the evolution of film to ensure that its content is appropriate for the target audience and that its presentation is consistent with cultural and ethical standards. This research systematically reviews previous literature on movie classification methods for genre detection or rating their age-appropriateness, based on the PRISMA schema [Preferred reporting items for systematic reviews and meta-analyses]. Overall, 78 studies were selected according to specific inclusion and exclusion criteria. Different literature sources were searched to retrieve the relevant literature, including ScienceDirect, Springer, IEEE, MDPI, arXiv, and the Google search engine. Among the included papers, 76 studies addressed the topic of movie genre classification, whereas only two considered age-appropriate rating. The results indicate that various data types were utilized for this task, including textual, visual, audio, and hybrid data. Moreover, support vector machine (SVM) was the most commonly deployed machine learning method, and the MovieLens dataset was most frequently adopted. This research addresses the taxonomy to predict movie genre based on its sources. It also sheds light on the extant models and essential features for accurate classification. It also highlights the research gaps and makes recommendations for future research.

1. INTRODUCTION

In contemporary culture and society, cinema remains a popular and potent source of entertainment. Dating back to the 1880s, the film industry has evolved into a vast global influence beyond mere entertainment, with educational, emotional, and economic implications. As a result, movies can have a profound impact on culture and lifestyle and raise public awareness, as well as affect human emotions and thinking [1].

However, movies are categorized into various genres based on their unique features. The categorization of movie genres is a pillar of the film industry. Thus, through the categorization of films according to their narrative content, stylistic arrangements, and thematic considerations, viewers can select those that align with their interests. This helps filmmakers reach their targeted demographic. Common genres include action, comedy, drama, horror, science fiction, and other types [2]. Some genres are unsuitable for certain audiences, such as children and young adults, which requires their age-appropriateness to be determined.

Age ratings are a form of categorization applied by regulatory bodies when classifying movies according to their appropriateness for different age groups [3]. These ratings are meant to provide the audience, specifically parents, with an indication of the suitability of the content of a movie for children or teenagers. Age ratings are primarily used to protect children and young audiences from exposure to content that might be disturbing or have a negative influence. It therefore provides a layer of safety over inappropriate content. Conversely, this also gives filmmakers the freedom to cover a broad range of topics and material, with no need for blanket censorship [4]. Age ratings are therefore very important since they not only affect cinema audiences but also the wider society, helping people to make informed choices about the films they watch and protecting minors from exposure to unsuitable content. In this regard, various aspects of a movie are considered, such as the level of violence depicted, use of profanity, any sexual content and scenes of drug use.

Therefore, age rating balances creative expression against social responsibility [5]. However, different countries have their own classification boards and rating scales that reflect their prevailing cultural values and norms; for example, the Motion Picture Association of America (MPAA) in the US, the British Board of Film Classification (BBFC) in the UK, the Central Board of Film Certification (CBFC) in India, and the Australian Classification Board (ACB). Each classification board defines and applies age ratings to suit the target audience [6]. In the US, the MPAA classification consists of G for General Audience, PG for Parental Guidance, PG-13 for Parents Strongly Cautioned, NC-17 for Adults Only, and R for Restricted. Each rating is detailed according to the

appropriateness of the content for different age groups [7]. However, age ratings equally indicate to adult viewers the type of content they are likely to see if they watch a particular film so that they can select movies that correspond to their tastes.

The manual classification of movies is based on narrative content, stylistic elements, and audience expectations. Narrative content incorporates the plot and themes, which determine the movie genre; stylistic elements express the visual and aural modes, such as cinematography, editing, and sound; for instance, the use of low lighting and eerie music is suggestive of a horror movie [8]. The audience expectations refer to what one might anticipate from a movie of a specific genre where they can expect certain key elements in movies of the same genre, such as heroic characters in action movies [9]. Automated film classification can play an important role in addressing the limitations of manual classification systems, saving time and effort while enhancing efficiency and transparency. However, automated classification faces some challenges, such as hybrid films that combine features from two or more genres [10]. Machine learning and artificial intelligence (AI) could be used to address such limitations, with the possibility of determining types of movies [11] or classifying or predicting age-appropriate ratings. Accordingly, this study is considered the first attempt to bridge the identified research gap.

The remainder of this paper is structured as follows: Section 2 introduces the methodology adopted to identify relevant papers, Section 3 reports the research findings, and Section 4 discusses the research outcomes, highlighting the research gaps and suggesting possible future research directions. Finally, Section 5 concludes with a summary of the key findings of this study.

2. METHODOLOGY

A systematic literature represents an attempt to provide a clear understanding of a specific topic in response to research questions by exploring the previous literature published within a specific timeframe. It discusses the results of these selected studies to deliver new knowledge to researchers, thereby saving them the effort and time required to browse each of these studies individually. This current review was conducted according to the principles of the PRISMA statement [Preferred reporting items for systematic reviews and meta-analyses] [12]. The PRISMA statement consists of a checklist of items for researchers to include when reporting systematic reviews and meta-analyses. This protocol sets out the optimal steps for preparing a study, including a search strategy, determining the sources of information, identifying the literature, screening and selecting the studies, extracting the data, determining the eligibility criteria, synthesizing the results, and assessing the quality of the studies against the risk of bias. To ensure clarity in their reporting, researchers should carefully follow the PRISMA guidelines to present a complete and standardized format, thereby providing an understanding and evaluation of the results.

2.1 Search strategy

Two inter-related topics were explored in this study: movie genre classification and the prediction of movie age-appropriateness. Five electronic databases were consequently searched: ScienceDirect, IEEE, Springer, and arXiv, which belongs to the University of Cornell and Multidisciplinary Digital Publishing Institute (MDPI). In addition, Google Scholar was used as a search engine to retrieve relevant materials. Using the combination operator ‘OR’, different search terms were entered into the advanced search feature of these databases – the same search terms were used for the databases and search engine. Table 1 presents the number of search results and their relevance to the inclusion criteria. The search was conducted over the period of one week, ending 24 April 2024, using the three-term keywords namely, ‘movie genre classification’, ‘film type prediction’, and ‘movie MPAA age rating’.

2.2 Eligibility criteria

Studies related to movie genre classification and the prediction of movie age-appropriate rating were selected based on four inclusion criteria:

1. Research studies written in English.
2. Peer-reviewed journal and conference papers.
3. Fully accessible papers.
4. Papers published between 1 January 2010 and 31 December 2023, inclusive.

Meanwhile, research studies were excluded according to the following criteria:

5. Articles published before 1 January 2010 or after 31 December 2023.
6. Papers without empirical results.
7. Articles not downloaded in their full text.
8. Studies written in a language other than English.
9. Studies with no known publisher and not indexed in Science Direct or Clarivate.
10. Studies focused on specific movie genres, such as action movies only.

Table 1. Data sources and search terms used

Data Source	Search Term	Search Filter Type: Term1, Term2, Term3	Search Results	Relevant Results
Science Direct (Elsevier)	"Movie genre classification" OR "film type prediction” OR "movie MPAA age rating"	None	16	3
Springer	"Movie genre classification" OR "film type prediction” OR "movie MPAA age rating"	None	76	6
IEEE	"Movie genre classification" OR "film type prediction” OR "movie MPAA age rating"	Document title, Document title, Document title	21	12
MDPI	"Movie genre classification" OR "film type prediction” OR "movie MPAA age rating"	All fields, All fields, All fields	8	1
arXiv	“Movie genre classification” OR "film type prediction” OR "movie MPAA age rating"	All fields, Title, Title	18	8
Google Scholar	"Movie genre classification" OR "film type prediction” OR "movie MPAA age rating"	None	659	44

A total of 798 studies were sourced, of which 139 were found in the identified databases, while 659 were located by the Google Scholar search engine. The steps of the study selection procedure are illustrated in Figure 1, which depicts a PRISMA flow diagram for this review.

In the identification step, it was found that 136 studies were duplicated, out of a total of 798. As such, only 662 papers were left for further investigation. The screening step involved reviewing the title and abstract for each paper, whereupon the Rayyan.ai website [13] was accessed to remove the duplicate studies. As a result, 476 studies were excluded because they did not meet the inclusion criteria. The content of the remaining 186 studies was then carefully checked by two authors to identify the most relevant papers after applying the specified criteria. Out of a total of 186 studies, 133 were considered eligible for further content reading, out of which 53 articles were found to be inaccessible in their full texts. Moreover, 55 studies were deemed to be irrelevant to this study, according to the inclusion and exclusion metrics. Finally, just 78 studies were included after checking their titles, abstracts, and entire content.

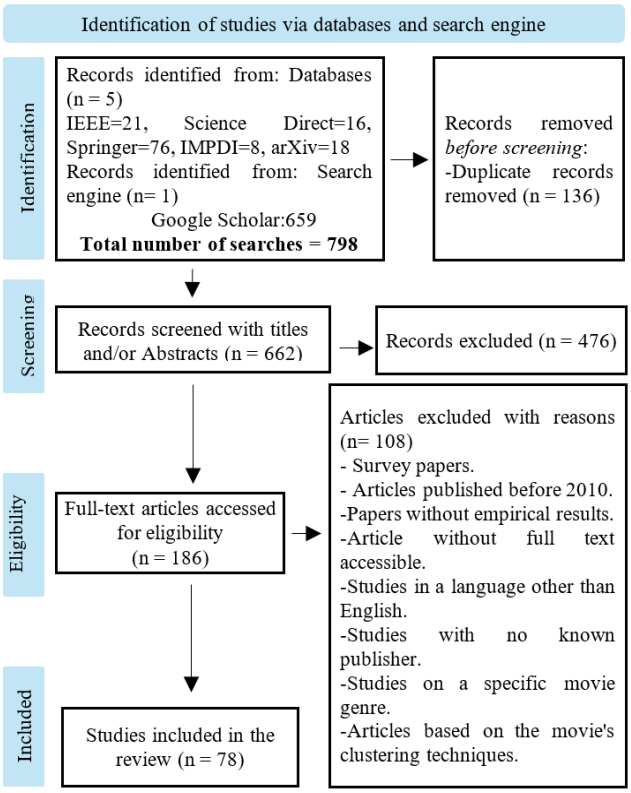

Figure 1. Study selection in a PRISMA statement

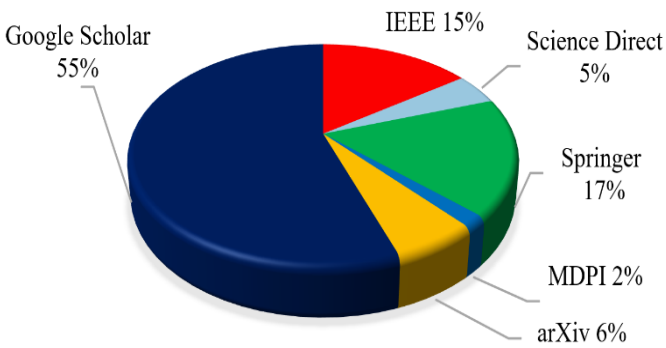

Figure 2. Rate of publication of the studies sourced

Results

The included studies were analyzed according to five categories: 1) Identifying the study topic: movie genre classification or the prediction of movie age rating, 2) Determining the prediction or classification method, 3) Exploring the datasets used, 4) Identifying the data type used, and 5) Determining the evaluation metrics. Based on these themes, this systematic analysis highlights the key aspects of the research topic, identifies the robustness of the previous literature, and sheds light on any unaddressed research gaps that could open the door to further research.

2.3 Description of the included studies

Table 2 shows the total number of studies that resulted from the search query, combining the results of the search engine records with those of the corresponding database source.

The search results indicated that the Springer database contained the largest number of publications from 2010 to 2023, comprising 17% of all the papers retrieved. From the IEEE database, a 15% publication rate was found, followed by arXiv at 6%, and Science Direct at 5%. Finally, the MDPI database yielded 2% of the publications sourced. Figure 2 presents the publication rate of all the studies retrieved by the Google Scholar search engine and the above-mentioned databases, demonstrating that around 55% of the papers appeared in the search engine and met the eligibility criteria.

Meanwhile, Figure 3 illustrates the number of published studies after removing duplicate papers and implementing the eligibility criteria. It indicates that studies on movie genre classification exceed the literature on age-appropriate movie rating. Moreover, Figure 4 depicts the number of publications per year in the five databases, indicating that the publication rate has increased since 2018, whereby the Springer and IEEE databases were found to have a higher publication rate than the MDPI database. Furthermore, Figure 5 demonstrates the publication rate for movie classification and age rating, revealing that the rate of publication increased from 2014 onwards.

Table 2. Sources of data used in the studies searched

Data Source	Type	No. of Publications	Publication Rate (%)
IEEE	Database	100	15%
Science Direct	Database	30	5%
Springer	Database	114	17%
MDPI	Database	11	2%
arXiv	Database	40	6%
Google Scholar	Search engine	367	55%

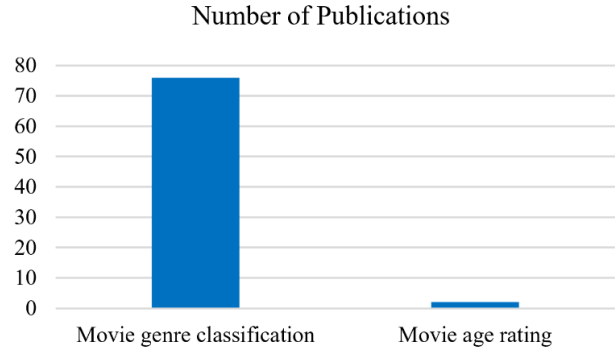

Figure 3. Rate of publication for both topics

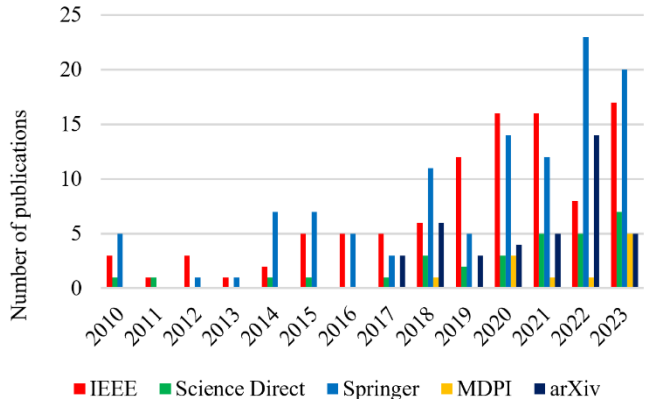

Figure 4. Studies per year in the databases

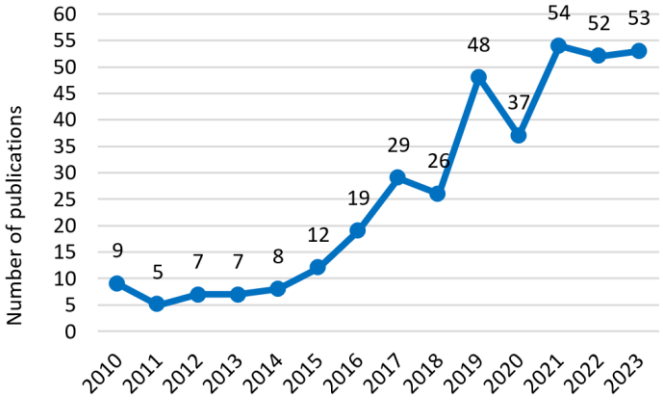

Figure 5. Publications retrieved from Google Scholar

Movie genre classification and the prediction of movie age-appropriateness can be determined through content analysis or collaborative filtering methods. Figure 6 presents a taxonomy for the available methods used in this area. In content analysis, different data types may be employed for this task, consisting of text, image, video, and audio content formats, as shown in Figure 7. Textual data can take the form of a plot summary, script, subtitle, or user review [14-34]. Conversely, visual data can be a movie poster [35-51] or movie trailer clip [52-65]. Meanwhile, audio files, such as movie soundtracks, have been used in other studies [66-68]. However, some researchers have combined different data types to enhance their models. These studies may be described as multimodal [69-90]. Collaborative filtering [15] methods offer another approach by leveraging users' viewing habits and preferences or exploring movie ratings to infer genre preferences. Table 3 summarizes the findings of the reviewed literature, identifying the research objectives, evaluation criteria, feature description, datasets, and data types.

Table 3 provides an organized overview of the key characteristics of earlier research. It identifies each work along with its primary goals, either for genre classification (G) or age rating prediction (A). It also determines the evaluation measures used to assess models' performance as well as the computational methods used across studies. Moreover, the source and scale of the utilized datasets and the type of the used data are highlighted. This could include text (T), poster (P), audio (Au) and/or video (V). Hence, Table 3 offers a comparative view of research trends, highlighting dominant modalities and algorithms in this field. From the results summarized in this Table 3, it is clear that various types of data were used in the reviewed studies, such as movie posters, plot summaries, and trailers. These were deployed individually or in combination, for example, including a poster with audio data. This is known as a multimodal approach. Table 4 lists the studies that adopted a multimodal classification approach, which integrates multimedia information and harnesses the features of each type of media used. Such algorithms and multimodels were used to enhance classification performance and make trustworthy predictions. It is clear that visual data could include posters, trailers or videos, while textual data could encompass multiple sources, such as scripts, plot summaries, synopses, subtitles, metadata, and descriptions or overviews. Moreover, the auditory features extracted from movies or trailers were incorporated with other data types.

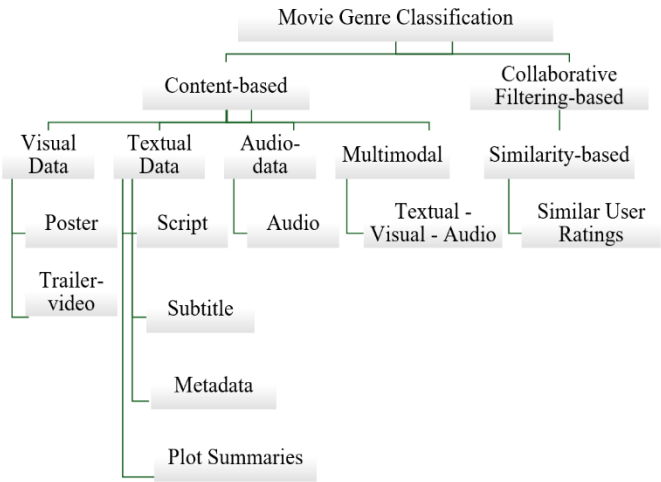

Figure 6. A proposed taxonomy of approaches in movie genre classification and movie age rating

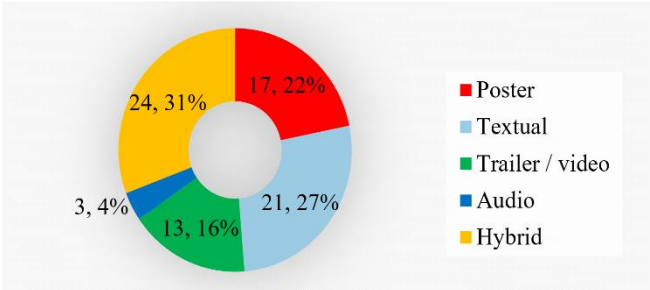

Figure 7. Ratio of data types used in the included studies

2.4 Datasets

A movie genre dataset typically contains information about various films and their associated genre. Such a dataset is essential for various applications in film studies, data analysis, machine learning, and recommendation systems. The earlier literature included numerous datasets, some containing textual data such as plot summaries, script, subtitles, synopses, and metadata. Examples of these datasets comprise the Internet Movie Database (IMDb) [52, 78], Movie Database (TMDB) [48, 81], Multi- language Movie Review Dataset (MLMRD) [34], and the Movie Summary Corpus (CMU) [22, 30].

Table 3. Characteristics of the reviewed studies

Study	Objective		Evaluation	Algorithm	Dataset	Data			
	G	A				T	P	Au	V
[52]	√		Accuracy	K-mean	IMDB, Apple Trailers.				√
[68]	√		Accuracy	SVM	MediaEval.			√	
[79]	√		Precision, Recall, F1-measure	KNN, Naïve Bayes NB	Collection of videos.			√	√
[87]	√		Accuracy	SVM	Collection of videos.			√	√
[88]	√		F1-score, AUC-PR	Multilayer Perceptron (BR-MLP), BR-SVM, BR-DT, LSTM, Multi-label k-Nearest Neighbors (ML-kNN)	Created dataset with 152,622 movies.	√	√	√	√
[69]	√		Precision, Recall, F1-score, correct detection (CD).	K-NN, SVM, Linear Discriminant Analysis (LDA).	Collection of videos.			√	√
[78]	√		Accuracy	SVM, Self-adaptive Harmony Search (SAHS).	Apple Trailers, IMDb.			√	√
[53]	√		Accuracy, Matthew's correlation coefficient (MCC).	WKLR	1000 frames downloaded from YouTube.				√
[66]	√		F-measure	DT, SVM, Sequential Minimal Optimization. (SMO), PART, k-Nearest Neighbors, ZeroR.	Collection of 769 audio tracks downloaded from YouTube.			√	
[48]	√		------	Distance Ranking (DR), NB, RAkEL.	TMDB.		√		
[51]	√		F1-score, Recall, Precision	RAkEL, ML-kNN, NB.	TMDB.		√		
[77]	√		F1-score, Recall, Precision	SVM, Hierarchical Clustering Relevance Feedback (HCRF).	MediaEval.		√	√	
[64]	√		Accuracy, mean Average Precision (mAP)	Used Optimum-path Forest (OPF).	MediaEval.				√
[81]	√		Accuracy	SVM + Vector Space Model (VSM).	TMDB.	√	√		
[90]	√		------	SVM, ANN.	Collection of videos.		√		√
[47]	√		Accuracy	CNN-MoTion.	LMTD-4.				√
[67]	√		F1-score	KNN, SVM, RF, Canonical Correlation Analysis (CCA)	BBC documentaries, RAI TV broadcasts.			√	
[16]	√		Precision, Recall, F1-score.	RF, SVM, NB, AdaBoost, C4.5.	500,000 subtitles	√			
[20]	√		Precision, Recall, F1-score.	SVM	MEG, MovieLens	√			
[44]			Recall, F1-score	RAkEL Ensemble (NB, C4.5 DT, k-NN).	TMDB.		√		
[50]	√		Accuracy, Recall, Precision, F1-score.	Binary relevance, RAkEL, NB.	TMDB.	√			
[86]	√		$\overline{\mathrm{AU(PRC)}}$, $AU\overline{(\mathrm{PRC})}$ $AU\overline{(\mathrm{PRC})}\mathrm{w}$, Ranking Loss.	CTT-MMC-(A/B/C), (CTT-MMC-S) for audio-based evaluation, ConvNets.	LMTD-9			√	√
[19]	√		F1-score	Topological data analysis (TDA).	IMDB	√			
[34]	√		Accuracy	Hybrid-model (LSTM, FCNN)	MLMRD	√			
[26]	√		Accuracy, Recall, Precision, Hamming Loss.	MLB, KNN	Large Movie Review Dataset v1.0, IMDb.	√			
[29]	√		Jaccard Index, F1-score.	NB, Word2Vec + XGBoost, RNN.	IMDB.	√			
[31]	√		Macro Precision, micro-Recall, macro F1-score, micro F1-score.	Bi-LSTM, LR, RNN	MovieLens, OMDb API2	√			
[32]	√		F1-scores, Precision, Recall.	RF, MLP, DT, Extra Trees Classifier.	TMDB	√			
[55]	√		Recall, Precision	Content-based filtering (CBF), Hybridized CBF.	MovieLens				√
[15]	√		F1-score, hit rate	Parametric Adaptive Rank Cut, Neighbors, WCN (Weighted Common Neighbors).	MovieLens	√			
[14]	√		Accuracy, Recall, F1-score.	HAN Architecture, Bidirectional LSTM	Wikipedia Movie Plots	√			
[17]	√		F1-scores.	SVM	Collection of videos	√			
[21]	√		F1 micro, Hamming Loss (HL), Exact Match (EM).	RF, NB, SVM	Opensubtitles, IMDb	√			
[24]	√		$\overline{Au(PRC)}$, $\overline{\mathrm{Au(PRC)}}$ $AU\overline{(\mathrm{PRC})}\mathrm{w}$.	SAS-MC-v2	LMTD	√			
[25]	√		Precision, F1-score, Recall.	SVM	IMSDb	√			
[30]	√		Accuracy, Recall, Precision	Naïve Bayes	CMU	√			
[36]	√		Recall, Precision, F1-score.	ResNet34, ML kNN.	MovieLens		√		
[39]	√		Accuracy, Precision, Recall.	ANN, SVM, RESNET-152.	MovieLens, LMTD.		√		

Table 3. Characteristics of the reviewed studies (Continued)

Study	Objective		Evaluation	Algorithm	Dataset	Data			
	G	A				T	P	Au	V
[54]	√		RMSE, MAE, Precision.	Content-based (CB)recommender system.	MovieLens.				√
[65]	√		Precision, Recall, Specificity.	Measure of existence (ME), relevance measure, cosine similarity.	SUN dataset				√
[73]	√		mAP, micro–Average Precision (uAP), sample Average Precision (sAP)	RF, LSTM, CRNN, Video-audio-poster-plot Text-metadata (VAPTM), VGG16, CRNN, TempConv	Moviescope	√			√
[3]		√	Weighted F1-score	RNN-based, SVM, CNN, LSTM with Attention	Creation of a dataset containing around 7000 movies.	√			
[22]	√		Precision, Recall, Accuracy, F1-score.	LR, SVM, ANN	CMU	√			
[28]	√		Accuracy	KNN, SVM, RNN, LSTM, CNN.	TMDB	√			
[38]	√		Accuracy, Precision, Recall, F1-score	ResNet with Gram layer CNN.	Collected dataset based on IMDb.		√		
[49]	√		-----	Inception v3	New dataset based on IMDB.		√		
[57]	√		Accuracy	3D CNN	New dataset downloaded from YouTube.				√
[58]	√		Accuracy	ILDNet	EmoGDB dataset, LMTD-9, MMTF14, ML-25M.				√
[59]	√		Recall, Precision	HMM, NBC	New dataset downloaded from BBC and Trecviddata.				√
[63]	√		AUPRC	Video representation fusion network (VRFN).	LMTD-9				√
[70]	√		Accuracy, Sensitivity, Specificity, Precision, F1 score	Deep convolutional neural network (DCNN), Deer Hunting Optimization (DHO).	LMTD				√
[71]	√		Accuracy	MobileNet, ResNet50, Inception, LSTM (256), Universal Sentence Encoder.	New dataset	√			√
[84]	√		F1-score, Precision, Recall	NB, DT, ML-KNN, RAKEL, VGG16.	A created dataset	√	√		
[23]	√		Precision, Recall, F1-score.	Parameter Optimized Hybrid Classification (POHC).	YIFY	√			
[27]	√		Accuracy, F1-score, Recall, Precision.	LR, NB	IMDB	√			
[33]	√		Accuracy, Recall, Precision	HANN, BiLSTM, LSTM, SVM, RF.	IMDB	√			
[41]	√		Accuracy, F1-score Precision, Hamming Loss.	Lenet, Alexnet, VGG-16, VGG-19, and Resnet-50, Proposed CNN-based.	IMDB	√			
[45]	√		Recall, F1-score	AlexNet, YOLO v3, NB	A created dataset, MovieLens.	√			
[46]	√		Accuracy	MobileNet	A created dataset		√		
[60]	√		Precision, Recall, F1-score, AU (PRC)	TFAnet	EMTD				√
[61]	√		mAP	Attention-based Spatiotemporal Sequential (ASTS).	New dataset				√
[80]	√		Micro (uAP), mAP, sAP.	Multi-GMU	Moviescope.	√	√	√	√
[89]	√		mAP, sAP	GloVe 42B, VGG 16	Moviescope.	√	√		
[91]		√	Weighted F1-score	RNN, CNN, LSTM, Gated Multimodal Unit (GMU), SVM, CNN-LSTM, BERT, DeepMoji.	Multimodal movie trailer rating (MM-Trailer).	√		√	√
[18]	√		F1-score	SVM, LR, NB.	IMDb	√			
[40]	√		Accuracy	CNN architecture with the Federal Learning Approach.	IMDB		√		
[56]	√		Accuracy	PlacesNet-LSTM	New dataset downloaded from YouTube.				√
[75]	√		uAP, mAP, wAP, sAP	Dual Image and Video Transformer Architecture (DIViTA).	ImageNet, Kinetics, a new dataset trailer.		√		√
[82]	√		Macro mAP, Micro mAP	Movie-CLIP + MLP	MovieNet, Condensed Movies.	√		√	
[83]	√		Accuracy, Precision, Recall, F1-score	CNN, T-Conv2D, Discrete Fourier Transform (DFT), Discrete Cosine	MovieLens			√	√

			Transform (DCT), Gated Recurrent Unit (GRU), ResNet, VGGNet, EfficientNetB7, MGT + Inception.						

Table 3. Characteristics of the reviewed studies (Continued)

Study	Objective		Evaluation	Algorithm	Dataset	Data			
	G	A				T	P	Au	V
[35]	√		Precision, Recall	Residual Dense Transformer (RDT), Ensembled Residual Dense Transformers (ERDT).	IMDb		√		
[37]	√		Accuracy, Precision, Hamming Loss	VGG16, ResNet, DenseNet, Inception, MobileNet, and ConvNeXt.	IMDB		√		
[72]	√		Hamming Loss, F-score, micro-AUC, weighted AUC.	GRU, 1D Convolution Model, SVM, CNN, LSTM	Multi-label Trailer Database, LMTD			√	√
[74]	√		Ranking Loss, AU(PRC)	Hierarchical Transformer Frame with Audio (HT-FA).	LMTD			√	√
[76]	√		Micro Precision, macro-Precision, weighted Precision, samples Precision	Self-supervised Attention, Knowledge Graph Feature Formation, incorporating Domain Knowledge Graph (IDKG)	MM-IMDb	√			√
[85]	√		F1-score	Weighted Average Ensemble Model, CNN, VGG-16, LSTM.	IMDB	√	√		
[42]	√		Accuracy	ResNet101, VGG19, AlexNet , KNN, NB.	IMDB		√		
[43]	√		Jaccard score	MNB, SVM, RF + SEMPD, MLkNN, RAndom k-labELsets (RAkELd)	MovieLens		√		

G: genre classification.
A: age suitability classification.
T: Textual data.
P: Poster data.
Au: Audio data.
V: video data.

Table 4. Studies on multiple data types

Study	Visual		Textual					Audio
	Poster	Trailer, Video.	Script	Plot Summary	Synopsis	Subtitle	Metadata	Description, Overview.	
[79]		√.							√
[87]		√.							√
[88]	√	√.			√	√			√
[78]		√.							√
[77]		√.							√
[81]	√				√				
[86]		√.							√
[73]		√		√					√
[3]			√						
[71]	√							√	
[84]	√							√.	
[80]	√	√.							√
[89]	√	√.		√			√		
[91]		√.				√			√
[75]	√	√.							
[82]		√.							√
[83]		√.							√
[72]		√.							√
[74]		√.							√
[76]	√			√			√		
[85]	√				√		√		
[90]		√.							√

Other datasets contain visual data, divided into poster, trailer, or video files, with examples including open movie datasets (OMDb) [31], MovieLens [15, 31, 39, 55], the Labelled Movie Trailer Dataset (LMTD) [24, 39, 86], and MediaEval [64, 68, 77]. In contrast, other datasets contain multiple data types such as posters, plot summaries, audio content, and movie trailers, for example, trailers in Moviescope [73, 80, 89] and multimodal movie trailer datasets (MM-Trailer) [91]. Conversely, some researchers have created their own datasets with a variety of data types to address data quality and reliability in relation to the problem of imbalance labels [52, 66, 68, 84, 90].

Figure 8 illustrates the employed datasets and the number of studies that used each one, along with the using ratio across the included studies. The IMDb dataset [18, 19, 26, 27, 29, 33, 35, 37, 40-42, 76, 85, 92] is considered to be the most significant and is widely accessed for information on films, TV shows, and the entertainment industry in general. It is especially useful for movie data analysis, machine learning models, and developing recommendation systems, since it contains basic information about films such as the title, genre, year of release, cast, director, and screenwriter. This dataset

has garnered a great deal of attention, having been sourced in 14 published papers (15%), as illustrated in Figure 8. Meanwhile, the MovieLens dataset [15, 20, 24, 31, 36, 43, 45, 46, 54, 55, 58, 60, 63, 70, 74, 83, 86] has been effectively deployed for movie genre classification, with 17 research studies using it for this purpose (18%). Meanwhile, the Labeled Movie Trailer Data (LMTD) dataset [24, 39, 46, 58, 60, 63, 70, 72, 74, 86] contains around 500 trailer files and was included in 10 studies (10%). The Movie Database (TMDb) [16, 28, 32, 48, 50, 51, 81] represents another important dataset, frequently utilized in the development and testing of various machine learning models in movie research. It contains comprehensive information about films and TV shows and was referenced in seven of the selected studies (7%). Alternatively, the MediaEval dataset [68] is a compilation of multimedia datasets, de-signed for benchmarking and evaluating diverse multi-media processing and analysis tasks. The above-mentioned dataset, which encompasses a variety of media types such as video, audio, text, and images, serves as a valuable resource for multimodal analysis and was found to have been employed in three of the research papers reviewed (3%).

In contrast, the Movie Summary Corpus (CMU) was sourced in just two of the included papers (2%). The CMU contains detailed plot summaries for a large number of movies, along with metadata such as genre information [22, 30]. Meanwhile, the Moviescope dataset contains rich metadata and annotated content, for example, title and overview, with a movie summary included in three of the studies (3%) [73, 80, 89]. Finally, other datasets were only mentioned once in the literature, such as the Kaggle Movie Dataset [36], SUBTIEL Corpus [14], Opensubtitles [21], MovieNet [82], Mul-ti-modal IMDb (MM-IMDb) [76], MovieLens Mul-ti-relational Dataset (MLMRD) [34], Internet Movie Script Database (IMSDb) [25], Open Movie Database (OMDb) [31], and Scene Understanding (SUN) [65]. Moreover, 31 datasets were created in several different studies to gather the distribution of different data types (visual and textual) in data records of varying size (32%).

2.5 Machine learning models applied

Different methods are utilized for predicting the label or class of trained datasets. The selection of a model is based on various metrics such as the type of data, complexity of the problem, effort required in terms of time and resource consumption, and level of accuracy. Traditional machine learning and artificial neural networks (ANNs) represent two different approaches in the machine learning domain. Traditional machine learning techniques involve a set of algorithms that learn patterns from data to make predictions or decisions. These methods often require careful feature engineering and domain knowledge to improve their performance.

The most common techniques in traditional machine learning include logistic regression (LR), decision trees (DT), random forest (RF), SVM, K-nearest neighbors (k-NN), Naïve Bayes (NB), and gradient boosting machines (GBM). These models have some advantages since they are typically considered faster to train and evaluate with small datasets. However, they can struggle with very large datasets or high-dimensional data, and their performance depends heavily on the quality and relevance of the features. A brief description of these techniques is presented in this subsection.

Artificial neural networks are computational models that are inspired by the human brain. Hence, they consist of layers of interconnected neurons that can learn complex patterns. This category includes several different techniques such as feedforward neural networks, recurrent neural networks (RNNs), convolutional neural networks (CNNs), and Transformers. Feedforward neural networks are especially powerful in deep learning architecture, whereas CNNs are specialized neural networks, designed to process grid-like data such as images. Convolutional neural networks are widely used in image classification tasks and many models are constructed upon them, for example, ResNet, AlexNet, MobileNet, Inception, and VGG networks. In contrast, RNNs are designed for sequential data, such as time series or natural language processing, wherein they capture temporal dependencies. Numerous models are designed around their structure, including Long Short-term Memory (LSTM), Bidirectional Long Short-term Memory (BiLSTM), and the Gated Recurrent Unit (GRU).

Alternatively, Transformers constitute an advanced neural network architecture that is designed to deal with sequential data based on an Attention Network Model. These models have become state-of-the-art in the field of natural language processing since they can handle large datasets, complex high-dimensional data, and learning features, directly from raw data and by reducing the need for manual feature engineering. They can be applied to a wide range of tasks, including image, text, audio, and time-series analysis.

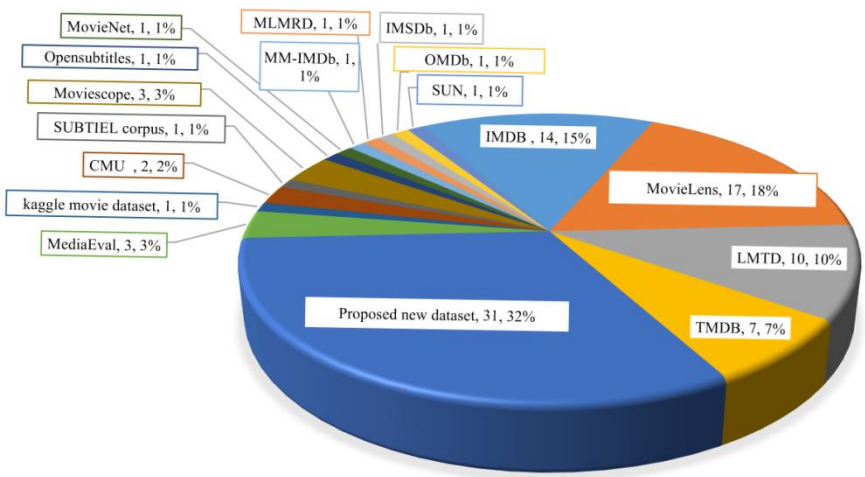

Figure 8. Datasets used in previous studies

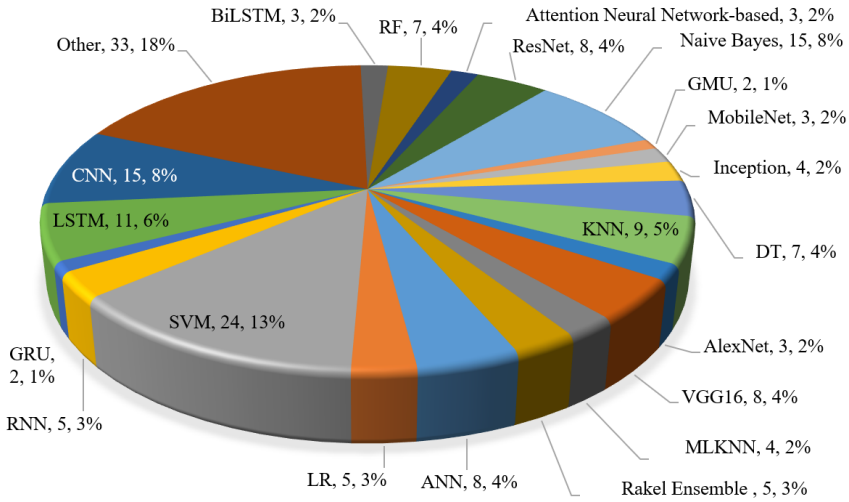

Figure 9. Machine learning models used in the reviewed studies

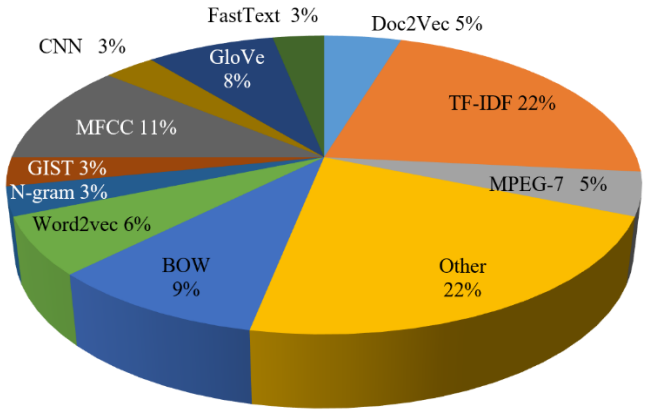

Figure 10. Feature extractors used in the reviewed literature

However, they require significant computational resources and training time to achieve high performance, especially in the case of deep networks with large amounts of labeled data. Nevertheless, to enhance models' performance, multiple models may be combined to produce reliable final output predictions, potentially outperforming the results of individual models.

These are known as ensemble methods, two popular examples being Random k-Labelsets (RAkEL), and Gated Multimodal Units (GMU).

Some models, such as Binary Relevance (BR) can only handle binary classification. Thus, in order to address multi-class or multi-label classification, techniques such as one-vs-rest (OvR) or one-vs-one (OvO) are required, using multiple binary classification. Moreover, other models, like Multinomial Logistic Regression or Multi-label K-nearest Neighbors (ML-kNN) have been extended to handle multi-class and multi-label classification directly using sigmoid activation or optimizing a softmax function.

Figure 9 illustrates the distribution of methods and models used across the reviewed studies. Each model type is shown with the frequency used and the corresponding percentage of adoption in the included research. The researchers deployed SVM [3, 16-18, 20-22, 25, 28, 33, 34, 39, 43, 66-69, 72, 78, 81, 87, 88, 90, 91] in 24 studies (13%); 15 studies (8%) included Naïve Bayes [16, 18, 21, 27, 29, 30, 42-45, 50, 59, 79, 84] and CNN [3, 28, 34, 38, 40, 41, 46, 57, 70, 72, 73, 83, 85, 86, 91]; 11 studies (6%) contained LSTM [3, 28, 31, 33, 34, 56, 72, 73, 85, 88, 91]; nine studies (5%) employed kNN [26, 28, 36, 42, 48, 67, 69, 88, 93]; VGG [37, 41, 42, 73, 83-85, 89] was used in seven studies (4%); eight studies (4%) utilized ResNet [36-39, 41, 42, 71, 83] and ANN [22, 26, 32, 33, 39, 82, 88, 90]; seven publications (4%) featured DT [16,32,66,77,84,88] and RF [21, 32, 33, 43, 67, 73, 92]; five studies (3%) employed the RAkEL ensemble approach [44, 48, 50, 51, 84]; five studies (3%) utilized LR [18, 22, 27, 31, 53] and RNN [3, 28, 29, 31, 91]; Inception [49, 71, 83] and ML-kNN [36, 43, 48, 84] were applied in four studies (2%); BiLSTM [14, 31, 33], Attention Neural Network [13, 33, 76], MobileNet [37, 47, 71], and AlexNet [41, 42, 45] were used in three studies (2%), and GRU [72, 83] and GMU [80, 91] were deployed in just two studies (1%).

Finally, some methods were used in just one study, with a combined 18% rate of inclusion. These methods included CTT-MMC [86], Movie-CLIP [82], the Vector Space Model (VSM) [81], the Hidden Markov Model (HMM) [59], content-based filtering (CBF) [55], distance ranking (DR) [48], YOLO v3 [45] SEMPD [43], GloVe 42B [89], PART [66], Lenet, canonical correlation analysis (CCA) [67], HAN architecture [14], EfficientNetB7 [83], XGBoost [29], AdaBoost [16], ConvNeXt [37], the Universal Sentence Encoder [71], DenseNet [37], SAS-MC-v2 [24], CRNN [73], the Residual Dense Transformer (RDT) [35], BERT [91], Topological Data Analysis (TDA) [19], Linear Discriminant Analysis (LDA)[69], Knowledge Graph Feature Formation [76], the Content-based Recommender System (CB) [54], ILDNet [58], TFAnet [60], the Feature Fusion Network (FFN) [63], Dual Image and Video Transformer Architecture (DIViTA) [75],

and hierarchical clustering relevance feedback (HCRF) [77].

2.6 Feature selection techniques

Machine learning models require feature extraction techniques during the important step of converting raw data into a more suitable form for analysis and processing. The choice of feature extractor will depend on the type of data and task. Various types of feature extractor are used with different data types such as textual, visual, and/or audio-data. The most frequently used feature extractors in textual data include Bag of Words (BoW), Term Frequency-inverse Document Frequency (TF-IDF), and N-gram. Bag of Words converts text into fixed-length vectors by counting the occurrence of each word in a document. Meanwhile, TF-IDF weighs the importance of words by considering their frequency in a document and their rarity across all documents. On the other hand, N-gram is used to transform text into numerical features based on contiguous sequences of 'N' items. Other techniques involve word-embedding and are exemplified by Word2Vec, GloVe, fasttext, and Doc2Vec [94]. These techniques map words to dense vectors in a continuous vector space where semantically similar words are close together. Moreover, contextual embedding, as performed by BERT [95], provides dynamic word representations based on the context in which words appear. Conversely, the LSTM model is based on RNN architecture and can be used to capture long-term dependencies in textual data.

There are a large number of feature descriptors specified for visual data. These are distributed via traditional methods such as SIFT, YOLO, and others, which collect the key points, edges, and textures from images or CNNs that deploy convolutional layers, in order to learn hierarchical feature representations from raw pixel data via automated means [96]. Some models are pre-trained; for instance, VGG, ResNet, and Inception use pre-trained models on large datasets (like ImageNet) to extract features from images. Furthermore, MPEG-7 refers to Multimedia Content Description Interface, which is used to describe multimedia content data such as audio and image data, and GIST is a descriptor that can capture the spatial structure in image data. Meanwhile, for audio-data, some feature extractors like Mel-Frequency Cepstral Coefficients (MFCCs) can capture the power spectrum of audio-signals and represent them in a way that closely resembles human auditory perception or spectrograms. This consists of visualizing the spectrum of frequencies in a signal as it fluctuates over time.

Figure 10 depicts the variety of feature extractors employed in the literature. Each one is presented alongside its usage frequency. The TF-IDF [17-21, 23, 26, 31, 32, 44, 82, 84, 88, 90, 91] appearing in 14 studies (22%), representing a relatively high rate. Meanwhile, BOW [23, 27, 29, 31, 81, 88], MFCC [66, 72, 79, 83, 86, 97], Glove [3, 24, 32, 84, 89], Word2Vec [14, 29, 32, 84], and Doc2Vec [22, 25, 28] were involved in six (9%), seven (11%), five (8%), four (6%), and three (5%) of the studies, respectively. Conversely, FastText [32, 73], CNN [83, 88], GIST [50, 52], and N-gram [30, 88] were used in 3% of the studies, while other feature descriptors were used only once, such as DNN [55], Local Binary Pattern (LBP) [88], Wavelet features [90], VGG16 [84], LSTM [88], Inception v3 [88], resNet50 [84], fastVideo [73], Wang2vec [32], Discrete Fourier Transform (DFT) [83], and Discrete Cosine Transform (DCT) [83].

2.7 Assessment criteria

Evaluation metrics are deployed to measure the performance of classification models. According to the metric used, different insights may be gained into the level of a model's performance. There are three types of classification: binary, multi-class, and multi-label. Accordingly, different metrics are designed for this purpose. In binary classification, accuracy metrics [98] are noted as the most commonly used evaluation measure for the proportion of correctly predicted instances out of a total number of instances. Precision [98] is considered as another important metric, if measuring the proportion of true positive predictions out of all positive predictions. Recall (sensitivity, true positive, or hit rate) measures [98] the proportion of true positive predictions out of all actual positives (number of relevant items selected). The F1-Score measures [98] the harmonic mean of Precision and Recall, providing a balance between the two. Specificity (true negative rate) measures [98] the proportion of true negative predictions out of all actual negatives. The ROC-AUC (Receiver Operating Characteristic - Area Under Curve) [99] is used to measure the trade-off between true and false positive rates across different thresholds, whereby a higher AUC indicates better model performance. Meanwhile, the Loss metric measures the performance of a classification model if the prediction input is a probability value of between 0 and 1, with lower log loss indicating a better model.

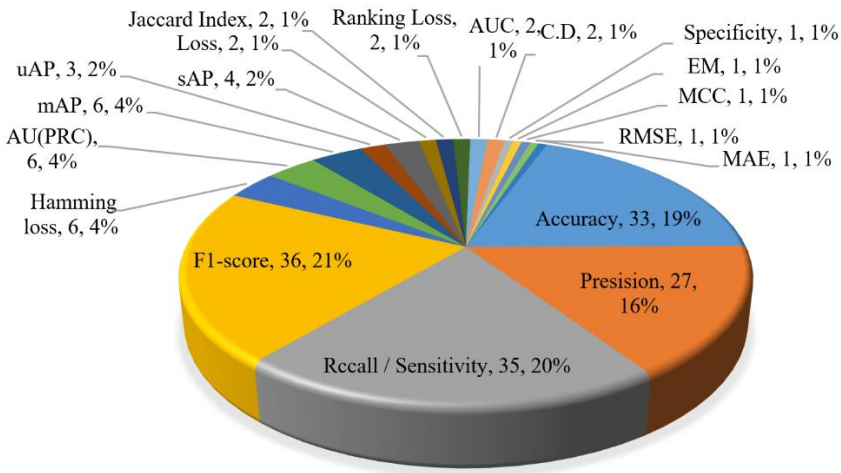

Figure 11. Metrics used in the studies

The Jaccard Index [100] and the Matthews Correlation Coefficient (MCC) [99] are metrics used to evaluate the performance of binary classifiers. In multi-class classification, the same accuracy metric is used but considers all classes, whereby Precision, Recall, and F1-score [98] can be calculated for each class separately. Alternatively, if in macro averaged format, the metric can be averaged across all classes. However, when micro-averaged [98], the metric is calculated globally across all instances and classes. Sample Average Precision (sAP) [98] is an evaluation metric that is primarily used in the context of object detection and image classification tasks, where it measures the average precision of the detection results over a set of samples. Unlike traditional average precision (AP) which considers Precision-Recall curves across an entire dataset, sAP calculates Precision for each sample and then averages the Precision. The calculation of sAP [98] involves two steps: first, Precision is calculated for each sample. Average Precision is calculated from the Precision values across all the samples.

The Area Under the Precision-Recall Curve AU (PRC) or AUC-PR measure [101] is an evaluation metric used to assess the performance of binary classification models, especially when dealing with imbalanced datasets where one class is significantly more frequent than the other. It focuses on the trade-off between Precision and Recall, which are often more informative than the ROC curve in such scenarios.

Another metric is the Confusion Matrix, this being a table that presents the number of correct or incorrect predictions for each class, thereby helping to clarify where the model is making errors. A further evaluation metric, known as Ranking Loss, indicates how frequently false prediction labels are given a higher ranking than true prediction labels.

In multi-label classification, the Hamming Loss measure [98] is considered to be one of the most important metrics to measure the percentage of wrongly predicted labels across all samples in a dataset.

Some studies included more than one metric for evaluation, for instance [63, 74, 79, 87], whereas others [52, 68, 81] included just one metric. Figure 11 provides an overview of the evaluation metrics utilized in the included studies. It identifies not only the total occurrences of each metric, but also their relative distributions within the literature. This provides insight into measures that are most frequently used for evaluating models' performance. The F1-score [69, 72, 84] was included in 36 of the studies (21%); Recall measurement [33, 45, 83] was employed in 35 studies (20%); the Accuracy metric [14, 19, 34, 37] occurred in 33 (19%) of the studies, and Precision [21, 22, 24, 25, 34] was mentioned in 27 of the reviewed publications (16%).

The other metrics, mentioned in Table 3, had a lower rate of inclusion, starting with Hamming Loss [21, 41], in six publications (4%); mAP [73-75, 80] in six publications (4%); AU(PRC) [24, 61, 63] in six publications (4%); sAP [73-75, 80] in four publications (2%), and uAP [73-75, 80] in three publications (2%). However, some metrics were used in only two (1%) studies, namely, Loss [37], Jaccard Index [29, 43], Ranking Loss [74], AUC [37, 72] and CD [69]. Meanwhile, other metrics appeared just once (1%), these being Specificity [65], EM [21], MCC [53], RMSE [54], and MAE [54].

3. DISCUSSION

This study aimed to systematically highlight the research directions for classifying genre and rating the age-appropriateness of movies, applying the PRISMA statement as a research protocol. The content of the literature retrieved and included was analyzed according to five main themes. However, out of 798 studies identified in the relevant literature, only 78 articles were included based on the inclusion and exclusion criteria.

Implementing an automatic model to classify movies into genres and age-appropriate ratings can save time and effort because a manual approach would otherwise require a specialist team to examine the content of movies, such as analyzing plot summaries and looking at the characters and their roles to make a final decision. Nevertheless, it should be clarified that comparing the findings of such studies is difficult, due to the use of different datasets, methods, evaluation metrics, feature extractors, and data types. Regarding the types of data used, hybrid data represents a combination of data types, and this approach is widely used to improve the accuracy of models. Additionally, textual data such as subtitles, scripts, plot summaries, and synopses were also adopted in many of the studies sourced, whereas poster data and movie trailers were utilized in a few of the studies because the processing of visual data requires powerful hardware with a high computational cost. Meanwhile, audio-data tends to receive less attention in studies, although it requires fewer computational resources. Nevertheless, many different audio-tracks must be analyzed before the genre of a movie can be identified.

With regard to the datasets used, the MovieLens dataset has attracted a great deal of attention from researchers because it contains a variety of data types, such as visual and textual data. This means that previous studies have employed a multimodal approach more frequently. Conversely, numerous researchers have created new datasets [15, 49, 88] to address the issue of class imbalance [81].

To evaluate the performance of the suggested classification models, various evaluation metrics have been employed in the literature. Some metrics are designed for single-label classification, such as accuracy and precision, while others are used for multiple and multi-label classification. In single-label classification, F1-scores comprised the most frequently used metric in the reviewed publications, with regard to multiple and multi-label classifications, followed by Accuracy and Recall. The most popular metrics appeared to be Hamming Loss and Average micro with Average macro. It was noted that there was no general standard metric used across all the studies. Thus, the comparison of findings from different studies may be complicated.

The most common methods applied in the classification of textual movie data were identified as SVM and Naïve Bayes. However, these methods can lead to inaccurate results when dealing with data imbalance, large datasets, or high-dimensional feature spaces. Many other studies have used deep learning models such as LSTM and VGG. One of the reviewed studies involved models that rely on Transformer architecture and self-Attention mechanisms such as BERT [91]. These methods can be used with large datasets to produce excellent results, but they need to be able to determine effective features accurately. Such methods are state-of-the-art in this field [3, 14]. By leveraging information from different sources, multimodal [80] methods are expected to achieve better results, compared to using single models. A multimodal method is based on the fusion concept, whether early or late fusion. Early fusion combines raw data from different

modalities at input level before feeding it into a model. In contrast, late fusion processes each modality separately and combines the output. Nevertheless, although such methods can produce better outcomes than a single model, they incur higher computational cost.

Classification methods rely on a crucial step known as feature extraction, whereby the feature extractor converts raw data into a set of features that can be used by a machine learning model. The most commonly employed feature descriptor is illustrated in Figure 10, followed by the BOW method. Nevertheless, although these methods are simple, they ignore the contextual meaning behind the text. In contrast, Glove and Word2Vec are the most frequently used methods that depend on word-embedding techniques, which maintain the contextual form by acquiring computational resources.

In visual and audio data, respectively, CNN and MFCC were identified as the models that are most frequently used as feature descriptors. For complex visual features, CNN is considered to be excellent and is robust to variations in input data. However, it requires large datasets and high computational resources to train models effectively. Some studies highlight the problem of multi-label classification [19, 30-33], while others depend on single-label classification [16]. The problem of multi-label classification has been addressed in a variety of ways in the literature, such as problem transformation. This includes binary relevance, classifier chains [43, 50, 88], and label powerset. The application of binary relevance offers simplicity but can fail to capture label correlation. However, it is suitable for movies with no dependencies between genres. Classifier chains address the matter of label dependencies, but the process requires intensive computation, which is sensitive to chain order. Conversely, label powerset can handle label dependencies, but struggles when the class label distribution is imbalanced. Other algorithms, like ML-kNN, are based on adaptation methods that can be used to handle multi-label classification issues [36], although this is considered to be computationally expensive, especially with large datasets. Moreover, there is a high reliance on the choice of distance metric, which can lead to poor classification if the wrong distance metric is chosen. This is particularly evident in the case of large datasets, because it requires calculating the distance between the test instance and all instances in the training set. Ensemble methods where a collection of classifiers is used, such as random k-labelsets (RAkEL), have outperformed other techniques identified in the reviewed literature [44, 84].

Multi-label classification relates to the problem of imbalanced datasets where some classes (labels) are significantly underrepresented, compared to others. Several studies have addressed this problem in different ways, since some researchers have created their own balanced datasets [49, 81, 84], which can help overcome the overfitting issue by reducing bias. In another study, the OvO approach was adopted since it can handle imbalances between classes [17]. However, this can also increase computational complexity, with difficulty in managing large numbers of classifiers.

In relation to evaluation techniques, the earlier literature deployed several different measurements, including the use of macro-averaging, micro-averaging, AU (PRC) [60, 63], and weighted F1-score [91], which can reflect model performance across classes or within individual classes. Moreover, some studies have implemented resampling techniques, such as multi-label synthetic methods, including the Minority Over-Sampling Technique (ML-SMOTE) and Multi-Label Tomek Link (MLTL) [88]. However, these techniques can lead to overfitting, especially if there are insufficient samples. Generally, the best movie genre classification result for Accuracy in textual synopses data was 90%, obtained using Doc2Vec and ANN [22]. It indicates that the use of a word-embedding feature extraction technique will outperform a traditional feature descriptor method, which cannot capture semantic meaning or context and is only suitable for small datasets. Meanwhile, the best result achieved for visual posters was 90.58%, based on Accuracy using DensNet [37], and the best Accuracy result for video data reached 95.23% using the DCNN algorithm with a set of feature descriptors [70]. For audio [68], the F1-score was 99% based on random forest with 50 high-dimensional audio features [67]. When using a multimodal technique, the best Accuracy result was 91.9%, using audio- and video-data with a wide range of feature extractors and SVM methods [78]. In the findings for the prediction of movie age-appropriateness, [91] the outcome of the F1-score was 86.06%, applying GMU as a classification model based on textual data, with tweets that included emojis and an MFCC audio feature descriptor, and the video feature extracted according to CNN and LSTM.

In the case of movie genre classification and age rating prediction, certain algorithms, such as support vector machines (SVM), are dominant. This is due to their potential ability for handling large-dimensional feature spaces, such as textual embedding and visual descriptors. This also shows their potency for multi-label classification problems. Despite the current popularity of deep learning techniques, SVM remains a strong choice in situations where the available data is small or the computing resources are limited. An insightful comparison of multimodal and unimodal techniques reveals that multimodal techniques are widely adopted, such as combining textual summaries, poster images, and emotion features. This is because such models benefit from richer representation and superior predictive outcomes, although their computational cost is high. On the other hand, unimodal techniques are efficient in their computational cost and can be simply interpreted. However, they are unable to capture the semantic variability of films' content.

Nevertheless, despite the developments in this topic and the use of advanced techniques, several gaps remain that require further solutions. For instance, this review revealed that no one dataset can ensure data quality or include all data types. Thus, researchers are invited to develop a more general dataset to address this gap. Movie genre classification has been extensively studied in 76 studies in this systematic review, whereas only two papers dealt with age rating prediction. The first research study was published in 2019. It states that no previous literature was conducted to predict the MPAA rating. This should invite further studies to bridge this research gap. Moreover, using state-of-the-art deep learning models, such as the Transformers and attention-based architectures, is still rare in this field. This is despite their demonstrated success in capturing complex sequential and contextual relationships in textual, visual, and multimodal information. Leveraging such models can improve the handling of long-range contextual dependencies in scripts or multimodal features. This may lead to improving the overall accuracy and generalizability of classification systems. Furthermore, regarding the cultural role in movies' classification, genre definitions and rating systems are not universal. They are actually determined by regionally and culturally defined environments and customs. For instance, the US-based MPAA and the UK-based BBFC classify movies

based on several different criteria. Such variations affect label allocation for movies as well as genre interpretation. Thus, future research should consider multi-regional datasets and shift to cross-cultural design. This may encourage broader use of genre classification systems. most studies have failed to consider cultural and regional differences in terms of genre preferences and definitions. However, different movie rating systems reflect the culture of their respective countries based on a number of factors, such as violence and sexual content. This means that the same film may have a different rating, according to where it is screened. For example, Deadpool (2016) was awarded a rating of '15' (suitable for 15 years and over) in the UK by the BBFC but was rated 'R' (restricted) in the US by the MPAA. Another example is Toni Erdmann (2016), which was rated '12' (age 12 or above) in Germany by Freiwillige Selbstkontrolle der Filmwirtschaft (FSK) [Voluntary Self-regulation of the Film Industry – the German motion picture rating organization] but rated 'R' in the US by the MPAA. This demonstrates that films can be rated differently around the world, corresponding to the national identity and values.

With regard to evaluation metrics, there is a lack of standardized evaluation measures across studies, which can make it difficult to compare research findings between earlier studies. Thus, no clear judgment can be made on the outcomes of the previous literature. However, this could be resolved by standardized datasets, containing a high number of diverse movies that are rendered accessible to researchers, for example, Hamming Loss, F1-score, Precision, and/or Accuracy to ensure a clear result for the models used.

4. CONCLUSION

This study represents a systematic review of the literature on movie classification, both in relation to movie genre and the prediction of age-appropriate rating. These areas of movie classification are considered important in research on information retrieval and multimedia use, due to their significant applications in recommendation systems, content organization, and digital libraries. Various data types were used in the studies sourced for this review, classifiable into textual data (for example, plot summaries, synopses, and scripts) or visual data (for example, posters and movie trailers). However, even though audio data was considered to be an important source, this type of data has received less attention in earlier studies. Thus, a combination of diverse data sources could have attained greater accuracy in the results.

In the reviewed studies, various data types were fed into different classifier models that were distributed across traditional machine learning and deep neural networks. The authors of the previous studies explored and created different datasets, as well as using an array of evaluation metrics to evaluate the performance of their models, handle multi-label classification problems, and deal with data imbalance. Therefore, due to the heterogeneity of the studies, a direct comparison between their outcomes is difficult.

Although this review makes significant contributions in this direction, it is not without its limitations. Therefore, this review invites further systematic reviews to consider the following recommendations for future research. First, creating benchmark datasets and standard evaluation metrics could help establish a foundation for comparing different algorithms and proving their performance. Second, most studies have limited their attention to English language movies, without considering movies in other languages. Therefore, it is important to address the challenge of genre classification across different languages and cultures. Third, advanced deep learning techniques based on Transformer models, such as BERT, should be explored to enhance the capture of complex relationships in the data. Forth, conducting cross-cultural studies is essential for testing the generalizability of models across different linguistic, cultural, and contextual settings, which remains a major gap in current research. Moreover, adding user data such as reviews could enrich the datasets used in genre classification and thereby contribute to a robust outcome.

Additionally, the search for pertinent studies was executed on just one occasion, which could have led to the omission of relevant papers published after the specified period. Second, the exclusion of studies that were not fully accessible and not written in the English language could have affected the findings and generalizability of this research. Moreover, this systematic review only included papers sourced from a few selected databases. Therefore, studies in other databases were overlooked, although an attempt was made to overcome this limitation by using the Google Scholar search engine. Finally, the use of predefined keywords may have restricted the inclusion of other studies. Hence, the use of different keywords could expand the search results and enrich the study with a higher volume of publications.

REFERENCES

[1] Toomela, A. (2020). Psychology today: Still in denial, still outdated. Integrative Psychological and Behavioral Science, 54: 563-571. https://doi.org/10.1007/s12124-020-09534-3

[2] Telotte, J., Duchovnay, G. (2011). Science Fiction Film, Television, and Adaptation: Across the Screens (Vol. 34). Routledge, UK.

[3] Shafaei, M., Samghabadi, N.S., Kar, S., Solorio, T. (2019). Rating for parents: Predicting children suitability rating for movies based on language of the movies. ArXiv Preprint ArXiv: 1908.07819. https://doi.org/10.48550/arXiv.1908.07819

[4] Tickle, J.J., Beach, M.L., Dalton, M.A. (2009). Tobacco, alcohol, and other risk behaviors in film: How well do MPAA ratings distinguish content? Journal of Health Communication, 14(8): 756-767. https://doi.org/10.1080/10810730903295567

[5] Calvert, S.L., Wilson, B.J. (2009). The Handbook of Children, Media, and Development. John Wiley & Sons. https://doi.org/10.1002/9781444302752

[6] Kublenz-Gabriel, L. (2016). Rating the rating systems: A comparison of media rating systems worldwide. Master's thesis, Behavioural, Management and Social Sciences University of Twente.

[7] Lachowitzer, C. (2017). (Re) Defining movie ratings: Acceptability, access, and boundary maintenance. Master's thesis, Department of communication studies, Colorado State University.

[8] Grant, B.K. (2007). Film Genre: From Iconography to Ideology (Vol. 33). Wallflower Press. London, UK.

[9] Neale, S. (Ed.). (2019). Genre and Contemporary Hollywood. Bloomsbury Publishing. London, UK.

[10] Langford, B. (2019). Film Genre: Hollywood and

Beyond. Edinburgh University Press. Edinburgh, UK. https://doi.org/10.1515/9781474470131
[11] Elsaesser, T., Hagener, M. (2015). Film Theory: An Introduction Through the Senses. Routledge, London, UK. https://doi.org/10.4324/9781315740768
[12] Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews, BMJ, 372: n71. https://doi.org/10.1136/bmj.n71
[13] Yu, F., Liu, C., Sharmin, S. (2022). Performance, usability, and user experience of Rayyan for systematic reviews. Proceedings of the Association for Information Science and Technology, 59(1): 843-844. https://doi.org/10.1002/pra2.745
[14] Lee, J. (2019). Movie genre classification based on plot description. Doctoral dissertation. College of Natural Sciences.
[15] Ghawi, R., Pfeffer, J. (2019). Movie genres classification using collaborative filtering. In ACM International Conference Proceeding Series, pp. 35-44. https://doi.org/10.1145/3366030.3366034
[16] Van Der Lee, C. (2017). Text-based video genre classification using multiple feature categories and categorization methods. Master Thesis, Department of Communication and Information Sciences, Tilburg University, Tilburg, Netherlands.
[17] Fei, N., Zhang, Y. (2019). Movie genre classification using TF-IDF and SVM. In ACM International Conference Proceeding Series, pp. 131-136. https://doi.org/10.1145/3377170.3377234
[18] Akbar, J., Utami, E., Yaqin, A. (2022). Multi-label classification of film genres based on synopsis using support vector machine, logistic regression and naïve Bayes algorithms. In 2022 6th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE), Yogyakarta, Indonesia, pp. 250-255. https://doi.org/10.1109/ICITISEE57756.2022.10057828
[19] Doshi, P., Zadrozny, W. (2018). Movie Genre Detection Using Topological Data Analysis. In Statistical Language and Speech Processing. SLSP 2018. Lecture Notes in Computer Science, pp. 117-128. https://doi.org/10.1007/978-3-030-00810-9_11
[20] Fourati, M., Jedidi, A., Gargouri, F. (2017). Generic descriptions for movie document: An experimental study. In Proceedings of IEEE/ACS International Conference on Computer Systems and Applications, AICCSA, Hammamet, Tunisia, pp. 766-773. https://doi.org/10.1109/AICCSA.2017.164
[21] Van der Meer, S.T. (2019). Multi-label classification of movie genres using text-based features and WordNet hypernyms. Radboud University Nijmegen. https://theses.ubn.ru.nl/handle/123456789/12560.
[22] Guehria, S., Belleili, H., Azizi, N., Belhaouari, S.B. (2020). 'One vs All' classifier analysis for multi-label movie genre classification using document embedding. In International Conference on Intelligent Systems Design and Applications, pp. 478-487. https://doi.org/10.1007/978-3-030-71187-0_44
[23] Mehedi Hasan, M., Tamim Dip, S., Rahman, T., Sonia Akter, M. (2021). Multilabel movie genre classification from movie subtitle: Parameter optimized hybrid classifier. In 2021 4th International Symposium on Advanced Electrical and Communication Technologies (ISAECT), Alkhobar, Saudi Arabia, pp. 1-6. https://doi.org/10.1109/ISAECT53699.2021.9668427
[24] Wehrmann, J., Lopes, M.A., Barros, R.C. (2018). Self-attention for synopsis-based multi-label movie genre classification. In The 31th International FLAIRS Conference, 2018, Brasil, pp. 236-241.
[25] Nakano, Y., Ohshima, H., Yamamoto, Y. (2019). Film genre prediction based on film content and screenplay structure. In Proceedings of the 21st International Conference on Information Integration and Web-Based Applications & Services, pp. 151-155. https://doi.org/10.1145/3366030.3366100
[26] Nyberg, A. (2018). Classifying movie genres by analyzing text reviews. ArXiv Preprint ArXiv: 1802.05322. https://doi.org/10.48550/arXiv.1802.05322
[27] Agarwal, A., Das, R.R., Das, A. (2021). Machine learning techniques for automated movie genre classification tool. In 2021 4th International Conference on Recent Developments in Control, Automation and Power Engineering, RDCAPE, Noida, India, pp. 189-194. https://doi.org/10.1109/RDCAPE52977.2021.9633422
[28] Wang, J. (2020). Using machine learning to identify movie genres through online movie synopses. In Proceedings - 2020 2nd International Conference on Information Technology and Computer Application, ITCA 2020, Guangzhou, China, pp. 691-697. https://doi.org/10.1109/ITCA52113.2020.00008
[29] Hoang, Q. (2018). Predicting movie genres based on plot summaries. ArXiv Preprint ArXiv: 1801.04813. https://doi.org/10.48550/arXiv.1801.04813
[30] Noersasongko, E., Salvana Ervan, D., Santoso, H.A., Supriyato, C., Al Zami, F., Soeleman, M.A. (2019). The use of particle swarm optimization to obtain n-gram optimum value for movie genre classification. Journal of Theoretical and Applied Information Technology, 97(20): 2441-2451.
[31] Ertugrul, A.M., Karagoz, P. (2018). Movie genre classification from plot summaries using bidirectional LSTM. In Proceedings - 12th IEEE International Conference on Semantic Computing, ICSC 2018, Laguna Hills, CA, USA, pp. 248-251. https://doi.org/10.1109/ICSC.2018.00043
[32] Portolese, G., Feltrim, V.D. (2018). On the use of synopsis-based features for film genre classification. Anais Do XV Encontro Nacional de Inteligncia Artificial e Computacional, 892-902. https://doi.org/10.5753/eniac.2018.4476
[33] Kumar Reddy, G., Shriram, R. (2021). Genre classification of Telugu and English movie based on the hierarchical attention neural network. International Journal of Intelligent Engineering and Systems, 14(1): 54-62. https://doi.org/10.22266/IJIES2021.0228.06
[34] Battu, V., Batchu, V., Reddy, R.R., Reddy, M.K., Mamiji, R. (2018). Predicting the genre and rating of a movie based on its synopsis. In Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation, pp. 52-62.
[35] Nareti, U.K., Adak, C., Chattopadhyay, S. (2023). Demystifying visual features of movie posters for multi-label genre identification. ArXiv Preprint ArXiv: 2309.12022. https://doi.org/10.1109/TCSS.2024.3481157
[36] Barney, G., Kaya, K. (2019). Predicting Genre from Movie Posters. Stanford CS 229: Machine Learning.

[37] Unal, F.Z., Guzel, M.S., Bostanci, E., Acici, K., Asuroglu, T. (2023). Multilabel genre prediction using deep-learning frameworks. Applied Sciences (Switzerland), 13(15): 8665. https://doi.org/10.3390/app13158665

[38] Wi, J.A., Jang, S., Kim, Y. (2020). Poster-based multiple movie genre classification using inter-channel features. IEEE Access, 8: 66615-66624. https://doi.org/10.1109/ACCESS.2020.2986055

[39] Álvarez, F., Sánchez, F., Hernández-Peñaloza, G., Jiménez, D., Menéndez, J.M., Cisneros, G. (2019). On the influence of low-level visual features in film classification. PLoS ONE, 14(2): e0211406. https://doi.org/10.1371/journal.pone.0211406

[40] Popat, A., Gupta, L., Meedinti, G.N., Perumal, B. (2023). Movie poster classification using federated learning. Procedia Computer Science, 218: 2007-2017. https://doi.org/10.1016/j.procs.2023.01.177

[41] Hossain, N., Ahamad, M.M., Aktar, S., Moni, M.A. (2021). Movie genre classification with deep neural network using poster images. In 2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD), Dhaka, Bangladesh, pp. 195-199. https://doi.org/10.1109/ICICT4SD50815.2021.9396778

[42] Dewidar, M. (2019). Inferring movie genres from their poster. Learning, 1.

[43] Sirattanajakarin, S., Thusaranon, P. (2019). Movie genre in multi-label classification using semantic extraction from only movie poster. In Proceedings of the 7th International Conference on Computer and Communications Management, pp. 23-27. https://doi.org/10.1145/3348445.3348475

[44] Ivasic-Kos, M., Pobar, M. (2017). Multi-label Classification of Movie Posters into Genres with Rakel Ensemble Method. In Artificial Intelligence XXXIV. SGAI 2017. Lecture Notes in Computer Science, pp. 370-383. https://doi.org/10.1007/978-3-319-71078-5_31

[45] Immanuel, J., Isa, S.M. (2021). Multi-label classification of genre film based on poster with the convolutional neural networks (CNN) method. ICIC Express Letters, 15(2): 109-115. https://doi.org/10.24507/icicel.15.02.109

[46] Simões, G.S., Wehrmann, J., Barros, R.C., Ruiz, D.D. (2016). Movie genre classification with convolutional neural networks. In 2016 International Joint Conference on Neural Networks (IJCNN), Vancouver, BC, Canada, pp. 259-266. https://doi.org/10.1109/IJCNN.2016.7727207

[47] Marcellus, M., Herwindiati, D.E., Hendryli, J. (2021). Movie poster genre classification with convolutional neural network. In Proceedings - 2021 IEEE 7th International Conference on Multimedia Big Data, BigMM 2021, pp. 74-77. https://doi.org/10.1109/BigMM52142.2021.00020

[48] Ivasic-Kos, M., Pobar, M., Mikec, L. (2014). Movie posters classification into genres based on low-level features. In 2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, pp. 1198-1203. https://doi.org/10.1109/MIPRO.2014.6859750

[49] Kundalia, K., Patel, Y., Shah, M. (2020). Multi-label movie genre detection from a movie poster using knowledge transfer learning. Augmented Human Research, 5(1): 1-9. https://doi.org/10.1007/s41133-019-0029-y

[50] Pobar, M., Ivasic-Kos, M. (2017). Multi-label poster classification into genres using different problem transformation methods. In Computer Analysis of Images and Patterns: 17th International Conference, CAIP 2017, Ystad, Sweden, August 22-24, 2017, Proceedings, Part II 17, pp. 367-378.

[51] Ivasic-Kos, M., Pobar, M., Ipsic, I. (2014). Automatic movie posters classification into genres. ICT Innovations 2014: World of Data, pp. 319-328. https://doi.org/10.1007/978-3-319-64698-5_31

[52] 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. https://doi.org/10.1145/1873951.1874068

[53] Hamed, A.A.M., Li, R., Zhang, X., Xu, C. (2013). Video genre classification using weighted kernel logistic regression. Advances in Multimedia, 2013(1): 653687. https://doi.org/10.1155/2013/653687

[54] 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. In ACM UMAP 2019 - Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization, pp. 303-308. https://doi.org/10.1145/3320435.3320470

[55] Deldjoo, Y., Elahi, M., Quadrana, M., Cremonesi, P. (2018). Using visual features based on MPEG-7 and deep learning for movie recommendation. International Journal of Multimedia Information Retrieval, 7(4): 207-219. https://doi.org/10.1007/s13735-018-0155-1

[56] Jiang, D. (2022). A hybrid PlacesNet-LSTM model for movie trailer genre classification. Journal of Theoretical and Applied Information Technology, 100(14): 5306-5316.

[57] Shambharkar, P.G., Thakur, P., Imadoddin, S., Chauhan, S., Doja, M.N. (2020). Genre classification of movie trailers using 3D convolutional neural networks. In 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, pp. 850-858. https://doi.org/10.1109/ICICCS48265.2020.9121148

[58] Yadav, A., Vishwakarma, D.K. (2020). A unified framework of deep networks for genre classification using movie trailer. Applied Soft Computing Journal, 96: 106624. https://doi.org/10.1016/j.asoc.2020.106624

[59] You, J., Liu, G., Perkis, A. (2010). A semantic framework for video genre classification and event analysis. Signal Processing: Image Communication, 25(4): 287-302. https://doi.org/10.1016/j.image.2010.02.001

[60] Kumar Vishwakarma, D., Jindal, M., Mittal, A., Sharma, A. (2021). Multilevel profiling of situation and dialogue-based deep networks for movie genre classification using movie trailers. arXiv: 2109.06488. https://doi.org/10.48550/arXiv.2109.06488

[61] Yu, Y., Lu, Z., Li, Y., Liu, D. (2021). ASTS: Attention based spatio-temporal sequential framework for movie trailer genre classification. Multimedia Tools and Applications, 80: 9749-9764. https://doi.org/10.1007/s11042-020-10125-y

[62] Kumar, V., Tripathi, V., Pant, B. (2019). Content based movie scene retrieval using spatio-temporal features. International Journal of Engineering and Advanced Technology, 9(2): 1492-1496. https://doi.org/10.35940/ijeat.B3495.129219

[63] Bi, T., Jarnikov, D., Lukkien, J. (2020). Video representation fusion network for multi-label movie genre classification. In Proceedings - International Conference on Pattern Recognition, Milan, Italy, pp. 9386-9391. https://doi.org/10.1109/ICPR48806.2021.9412480

[64] Martins, G.B., Almeida, J., Papa, J.P. (2015). Supervised video genre classification using optimum-path forest. Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications: 20th Iberoamerican Congress, CIARP 2015, Montevideo, Uruguay, November 9-12, 2015, Proceedings 20, pp. 735-742. https://doi.org/10.1007/978-3-319-25751-8_88

[65] Varghese, J., Ramachandran Nair, K.N. (2019). A novel video genre classification algorithm by keyframe relevance. In Information and Communication Technology for Intelligent Systems. Smart Innovation, Systems and Technologies, pp. 685-696. https://doi.org/10.1007/978-981-13-1742-2_68

[66] Luna, B. (2013). Automatic movie genre classification based on musical descriptors. http://hdl.handle.net/10230/22902.

[67] Sageder, G., Zaharieva, M., Breiteneder, C. (2016). Group feature selection for audio-based video genre classification. In MultiMedia Modeling. MMM 2016. https://doi.org/10.1007/978-3-319-27671-7_3

[68] Austin, A., Moore, E., Gupta, U., Chordia, P. (2010). Characterization of movie genre based on music score. In 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, Dallas, TX, USA, pp. 421-424. https://doi.org/10.1109/ICASSP.2010.5495763

[69] Ionescu, B.E., Rasche, C., Vertan, C., Lambert, P., Seyerlehner, K. (2012). Video genre categorization and representation using audio-visual information. Journal of Electronic Imaging, 21(2): 1. https://doi.org/10.1117/1.jei.21.2.023017

[70] Shambharkar, P.G., Doja, M.N. (2020). Movie trailer classification using deer hunting optimization based deep convolutional neural network in video sequences. Multimedia Tools and Applications, 79(29-30): 21197-21222. https://doi.org/10.1007/s11042-020-08922-6

[71] Jiang, Y. (2020). Video game genre classification based on deep learning.

[72] Behrouzi, T., Toosi, R., Akhaee, M.A. (2023). Multimodal movie genre classification using recurrent neural network. Multimedia Tools and Applications, 82(4): 5763-5784. https://doi.org/10.1007/s11042-022-13418-6

[73] Cascante-Bonilla, P., Sitaraman, K., Luo, M., Ordonez, V. (2019). Moviescope: Large-scale analysis of movies using multiple modalities. arXiv: 1908.03180. https://doi.org/10.48550/arXiv.1908.03180

[74] Cai, Z., Ding, H., Wu, X., Xu, M., Cui, X. (2023). Hierarchical transformer for multi-label trailer genre classification. In ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, Rhodes Island, Greece, pp. 1-5. https://doi.org/10.1109/ICASSP49357.2023.10095502

[75] Montalvo-Lezama, R., Montalvo-Lezama, B., Fuentes-Pineda, G. (2022). Improving transfer learning with a dual image and video transformer for multi-label movie trailer genre classification. arXiv: 2210.07983. https://doi.org/10.48550/arXiv.2210.07983

[76] Li, J., Qi, G., Zhang, C., Chen, Y., Tan, Y., Xia, C., Tian, Y. (2023). Incorporating domain knowledge graph into multimodal movie genre classification with self-supervised attention and contrastive learning. In MM 2023 - Proceedings of the 31st ACM International Conference on Multimedia, pp. 3337-3345. https://doi.org/10.1145/3581783.3612085

[77] Ionescu, B.E., Seyerlehner, K., Mironică, I., Vertan, C., Lambert, P. (2014). An audio-visual approach to web video categorization. Multimedia Tools and Applications, 70(2): 1007-1032. https://doi.org/10.1007/s11042-012-1097-x

[78] Huang, YF., Wang, SH. (2012). Movie Genre Classification Using SVM with Audio and Video Features. In Active Media Technology. AMT 2012. Lecture Notes in Computer Science, pp. 1-10. https://doi.org/10.1007/978-3-642-35236-2_1

[79] Giannakopoulos, T., Makris, A., Kosmopoulos, D., Perantonis, S., Theodoridis, S. (2010). Audio-visual fusion for detecting violent scenes in videos. In Artificial Intelligence: Theories, Models and Applications. SETN 2010. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-12842-4_13

[80] Rodríguez-Bribiesca, I., Pastor López-Monroy, A., Montes-Y-Gómez, M. (2021). Multimodal weighted fusion of transformers for movie genre classification. In Proceedings of the Third Workshop on Multimodal Artificial Intelligence, pp. 1-5. https://doi.org/10.18653/v1/2021.maiworkshop-1.1

[81] Fu, Z., Li, B., Li, J., Wei, S. (2015). Fast film genres classification combining poster and synopsis. In Intelligence Science and Big Data Engineering. Image and Video Data Engineering. IScIDE 2015. https://doi.org/10.1007/978-3-319-23989-7_8

[82] Zhang, Z., Gu, Y., Plummer, B.A., Miao, X., Liu, J., Wang, H. (2022). Movie genre classification by language augmentation and shot sampling. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 7275-7285.

[83] Türköz, I. (2022). Multi-label multi-modal classification of movie scenes. Ph.D. dissertation, The Graduate School of Engineering and Science, Bilkent University.

[84] Nambiar, G., Roy, P., Singh, D. (2020). Multi modal genre classification of movies. In 2020 IEEE International Conference for Innovation in Technology (INOCON), Bangluru, India, pp. 1-6. https://doi.org/10.1109/INOCON50539.2020.9298385

[85] Khant, P., Tidke, B. (2023). Multimodal approach to recommend movie genres based on multi datasets. Indian Journal of Science and Technology, 16(30): 2304-2310. https://doi.org/10.17485/IJST/v16i30.1238

[86] Wehrmann, J., Barros, R.C. (2017). Movie genre classification: A multi-label approach based on convolutions through time. Applied Soft Computing Journal, 61: 973-982. https://doi.org/10.1016/j.asoc.2017.08.029

[87] Muneesawang, P., Guan, L., Amin, T. (2010). A new learning algorithm for the fusion of adaptive audio-visual features for the retrieval and classification of movie clips. Journal of Signal Processing Systems, 59(2): 177-188.

https://doi.org/10.1007/s11265-008-0290-7
[88] Mangolin, R.B., Pereira, R.M., Britto, A.S., Silla, C.N., Feltrim, V.D., Bertolini, D., Costa, Y.M.G. (2020). A multimodal approach for multi-label movie genre classification. Multimedia Tools and Applications, 81: 19071-19096. https://doi.org/10.1007/s11042-020-10086-2
[89] Lebaron, D. (2021). Multi-label movie genre classification using multiple modalities. https://api.semanticscholar.org/CorpusID:246482479.
[90] Rouvier, M., Oger, S., Linarès, G., Matrouf, D., Merialdo, B., Li, Y. (2015). Audio-based video genre identification. IEEE Transactions on Audio, Speech and Language Processing, 23(6): 1031-1041. https://doi.org/10.1109/TASLP.2014.2387411
[91] Shafaei, M., Smailis, C., Kakadiaris, I.A., Solorio, T. (2021). A case study of deep learning based multi-modal methods for predicting the age-suitability rating of movie trailers. ArXiv Preprint ArXiv: 2101.11704. https://doi.org/10.48550/arXiv.2101.11704
[92] LeCun, Y., Bengio, Y., Hinton, G. (2015). Deep learning. Nature, 521(7553): 436-444. https://doi.org/10.1038/nature14539
[93] Giannakopoulos, T., Pikrakis, A., Theodoridis, S. (2010). A multimodal approach to violence detection in video sharing sites. In 2010 20th International Conference on Pattern Recognition, Istanbul, Turkey, pp. 3244-3247. https://doi.org/10.1109/ICPR.2010.793
[94] Patil, R., Boit, S., Gudivada, V., Nandigam, J. (2023). A survey of text representation and embedding techniques in NLP. IEEE Access, 11: 36120-36146. https://doi.org/10.1109/ACCESS.2023.3266377
[95] Devlin, J., Chang, M. W., Lee, K., Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, Minnesota, pp. 4171-4186. https://doi.org/10.18653/v1/N19-1423
[96] Chandana, R.K., Ramachandra, A.C. (2022). Real time object detection system with YOLO and CNN models: A review. arXiv Prepr. arXiv2208, 773. https://arxiv.org/pdf/2208.00773
[97] Shafaei, M. (2021). Detecting objectionable content in online media. Ph.D. dissertation. Department of Computer Science, University of Houston, Texas m USA.
[98] Sokolova, M., Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing and Management, 45(4): 427-437. https://doi.org/10.1016/j.ipm.2009.03.002
[99] Chicco, D., Jurman, G. (2023). The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. BioData Mining, 16(1): 1-23. https://doi.org/10.1186/S13040-023-00322-4
[100] Letcher, S., Isla, M.Z. (2018). Comparing sets of patterns with the Jaccard index. Australasian Journal of Information Systems, 22. https://doi.org/10.3127/AJIS.V22I0.1538
[101] Khan, S.A., Ali Rana, Z. (2019). Evaluating performance of software defect prediction models using Area under Precision-Recall Curve (AUC-PR). In 2019 2nd International Conference on Advancements in Computational Sciences, ICACS 2019, Lahore, Pakistan, pp. 1-6. https://doi.org/10.23919/ICACS.2019.8689135

APPENDIX

Evaluation metrics

This appendix provides the formal definitions of the evaluation metrics commonly applied in multi-label movie genre classification. The accuracy metric is computed according to Eq. (1):

$$\text{Accuracy}= \frac{\text{Number of correct predictions}}{\text{Total Number of predictions}} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

True positives (*TP*): Cases where the model correctly predicts the positive class.

True negatives (*TN*): Cases where the model correctly predicts the negative class.

False positives (*FP*): Cases where the model incorrectly predicts the positive class.

False negatives (*FN*): Cases where the model incorrectly predicts the negative class.

The precision, recall, F1-Score, Specificity is computed according to Eqs. (2)-(5).

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

$$\text{F1-Score} = 2* \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

$$\text{Specificity} = \frac{TN}{TN+FP} \quad (5)$$

Another metric was used in some studies such as Jaccard similarity, it evaluates the similarity between the set of predicted positive instances and the set of actual positive instances as Eqs. (6)-(7).

$$\text{Jaccard Index}= \frac{|A \cap B|}{|A \cup B|} \quad (6)$$

$$\text{Jaccard Index}= \frac{TP}{TP+FP+FN} \quad (7)$$

where, $|A \cap B|$ is the number of elements in the intersection of sets A and B, and $|A \cup B|$ is the number of elements in the union of sets A and B.

Matthews Correlation Coefficient (MCC) [99] is generally regarded as a balanced measure that can even be used in an imbalanced class distribution, as in Eq. (8).

$$\text{MCC} = \frac{\text{TP} * \text{TN} - \text{FP} * \text{FN}}{\sqrt{(\text{TP} + \text{FN})*(\text{TN} + \text{FP})*(\text{TP} + \text{FP}) *(\text{TN} + \text{FN})}} \quad (8)$$

The mAP, uAP, sAP is obtained from Eqs. (9)-(11).

$$\text{Macro Averaged Precision (mAP)}= \frac{1}{C} \sum_{j=1}^{c} Precision\ j \quad (9)$$

where, c refers to the number of total genres and j represents individual movie genres.

$$\text{Micro Averaged Precision(uAP)}=\frac{\sum_{j=1}^{c} TP_j}{\sum_{j=1}^{c} (TP_j + FP_j)} \quad (10)$$

$$\text{sAP}= \frac{1}{N} \sum_{i=1}^{N} Precision_j \quad (11)$$

where, N is the total number of samples and $Precision_j$ is the Precision for the *i-th* sample.

The Exact Match (EM) metric is used to evaluate the performance of multi-label classification models and natural language processing (NLP) tasks such as answering questions and classifying text. This metric measures the percentage of instances where the predicted labels exactly match the true labels , as in Eq. (12):

$$\text{EM}= \frac{1}{N} \sum_{i=1}^{N} I(y_i = \hat{y_i})^2 \quad (12)$$

where, N is the total number of instances, $\hat{y_i}$ is the predicted label set for instance I, and y_i is the true label set for instance i. Here, I is the indicator function that returns '1' if $y_i = \hat{y_i}$ and otherwise, '0'.

The Correct Detection (CD) metric measures the performance of classification models by evaluating the proportion of correctly detected positive instances among all actual positive instances. This metric is closely related to Recall but is specifically focused on the correct identification of positive cases as define in Eq. (13).

$$\text{CD} = \frac{TP}{TP+FN} \quad (13)$$

Hamming Loss assesses the accuracy of a model by identifying each sample according to all relevant labels, as shown in Eq. (14):

$$\text{Hamming Loss} = \frac{Number\ of\ incorrectly\ predicted\ labels}{Total\ number\ \text{of}\ labels} \quad (14)$$

Mean Absolute Error (MAE) [54] metric measures the average absolute difference between actual and predicted values, as determined by Eq. (15):

$$\text{MAE}= \frac{1}{N} \sum_{j=1}^{N} |y_i - \hat{y_i}| \quad (15)$$

Moreover, Root Mean Squared Error (RMSE) [54] takes the square root of the Mean Squared Error (MSE) as in Eq. (16), which measures the average squared difference between actual and predicted values, as demonstrated in Eq. (17):

$$\text{RMSE}= \sqrt{MSE} \quad (16)$$

$$\text{MSE}= \frac{1}{N} \sum_{j=1}^{N} (y_i - \hat{y_i})^2 \quad (17)$$

since N = the total number of instances, $\hat{y_i}$ is the predicted label set for instance i, and y_i is the true label set for instance i.