ORIGINAL ARTICLE

Hybrid deep learning approach for multi-label classification problem: genre prediction

Fatıma Zehra Ünal · Mehmet Serdar Güzel · Metehan Ünal · Fatih Ekinci · Tunç Aşuroğlu · Koray Açıcı

Received: 11 July 2025 / Accepted: 12 January 2026
© The Author(s) 2026

Abstract
The primary aim is to develop a model that can achieve high accuracy in solving multi label classification problems by training, testing, and analyzing deep learning models that utilize both image and text data. In this paper, we propose a novel hybrid model that jointly considers textual and visual data for the task of movie genre prediction, which is a representative example of multi label classification problems. In the proposed model, textual features extracted from movie summaries are obtained using the DistilBERT model, while visual features derived from movie posters are extracted using the ConvNeXt deep learning model. These features from the two modalities are then combined using the XGBoost machine learning algorithm to perform genre prediction. This approach aims to achieve higher accuracy and better generalizability in movie genre classification by integrating information from different modalities through a late fusion method. The ConvNeXt architecture was adapted to the problem using transfer learning and fine-tuning techniques. To achieve the highest performance from the DistilBERT model, optimization was performed for the token length and threshold hyperparameters, and the model with a token length of 256 and a threshold of 0.5 was used in the hybrid model. Furthermore, to maximize the overall performance of the novel hybrid model, optimization was conducted using the Grid Search algorithm. All three proposed models were trained and tested on a dataset obtained from the IMDB website. The performances of the models were evaluated using hamming loss, precision and F1 score metrics. Experimental results revealed that, overall, the text-based model outperformed the image-based model, while the proposed hybrid model achieved higher performance than both individual models. It was demonstrated that textual and visual features complement each other and positively enhance the overall performance. This paper presents an original and effective study for multi label movie genre classification, combining the fields of computer vision, natural language processing, and machine learning.

Keywords Multi label classification · Genre prediction · Deep learning · Artificial intelligence



Published online: 06 March 2026

1 Introduction

Multilabel classification problems are important because many real-world datasets can simultaneously belong to multiple categories. For example, an article may also be grouped under the biology, computer science, medicine. A movie poster might represent drama and thriller genres at the same time. Similarly, a patient can have multiple diseases concurrently. Complex structures that cannot be accurately defined by a single label can be more precisely categorized through multilabel classification. Multilabel classification applications provide practical solutions for addressing the complexity of real-world problems.

The capacity of features obtained from different types of data to complement each other is a topic of interest for researchers. One application of multilabel classification is the automatic prediction of movie genres. In this study, the problem chosen is the automatic determination of a movie genre because a movie can have more than one genre at the same time.

Early studies on genre classification predominantly relied on single modality inputs, typically using textual data such as movie plots or synopses. With the rise of deep learning, Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have been widely applied to text-based genre classification, achieving moderate success. More recent approaches have explored multimodal learning, combining textual and visual cues such as posters or trailers to improve predictive performance. Notable examples include the works of [1, 2], and [4] which integrated CNN- and Transformer-based architectures through various fusion strategies. Although these studies achieved promising results, they generally relied on conventional architectures such as ResNet and BERT, with limited exploration of modern, optimized networks like ConvNeXt and DistilBERT. Furthermore, most fusion strategies in the literature are neural-based (early or late fusion), while tree-based or boosting-based fusion mechanisms remain largely unexplored. This limits both the interpretability and flexibility of multimodal integration.

In addition to architectural constraints, several methodological gaps remain. Many studies overlook data imbalance and multilabel stratification, using random splits that result in overrepresentation of certain genres. Evaluation metrics are also often limited to accuracy or F1 score, whereas Hamming loss, a metric tailored for multilabel classification, is rarely considered. Moreover, while multimodal approaches have demonstrated improved performance, few studies have analyzed the individual contribution of textual and visual modalities, leaving the question of their relative impact unresolved.

While significant progress has been made in the field of multimodal movie genre classification in recent years, the majority of existing work is still based on classical deep learning architectures (e.g., CNN and BERT). These models fail to sufficiently integrate the interaction between visual and textual modalities, limiting the interpretability of intermodal relationships. Furthermore, existing fusion methods are mostly neural network-based and exhibit limitations in generalization performance on imbalanced, multi-label datasets. These limitations formed the primary motivation for our study.

In this context, a novel deep hybrid model is proposed that combines a text-based classification model with an image-based classification model trained on the same dataset. The proposed hybrid model aims to examine the effect of complementarity of different types of data on the classification performance by utilizing both image and text features.

For the study, a text-based classification model and image-based classification model were initially created and evaluated separately. For the image-based classification model, state of the art pretrained ConvNeXt network was adapted to the genre prediction problem using transfer learning and fine-tuning techniques. The model's performance was evaluated and recorded based on metrics such as Hamming loss, F1 score and precision. For the text-based classification model, the DistilBERT network, commonly used in Natural Language Processing applications, was adapted for this specific problem. In the second stage of the study, novel deep hybrid model was proposed, combining features obtained from both image-based and text-based models. The goal was to investigate the potential positive or negative effects of this multimodal deep hybrid model, created with different types

of data, on classification performance. Each of the three proposed models was trained and tested on the dataset obtained from IMDB website. During the training and evaluation of the image-based model, the movie plots and synopsis corresponding to the posters in the dataset were obtained using a Python script. This process resulted in a multimodal dataset containing both text and images, which can be used for multi label classification problems.

The contributions and innovations of this article can be listed as follows:

- A novel multimodal hybrid deep network was developed by combining the ConvNeXt network, used in the image-based model, with the DistilBERT network (token length 256, threshold 0.5), used in the text-based model. In the studies conducted so far, these models have not been used together for the genre prediction problem. Additionally, DistilBERT was selected for the first time as the starting point for genre prediction models based on text-based movie summaries. In multimodal approaches that combine visual and textual modalities, classical CNN or RNN/LSTM-based architectures have generally been preferred. In our study, ConvNeXt's deep normalization and modern convolutional architecture and DistilBERT's lightweight Transformer architecture, which minimizes information loss, are integrated in a complementary manner. This allows for a more robust and balanced representation of both visual and textual features.
- In contrast to direct feature fusion approaches commonly used in the literature, our study uses the XGBoost algorithm to combine visual and textual representations at the decision level. This method reduces the risk of overfitting, increases interpretability, and enables heterogeneous processing of information obtained from different modalities, resulting in a more robust and generalizable hybrid model.
- In our study, we applied the stratified 5-fold cross validation and Multi-Label Stratified Shuffle Split (MLSS) method to reduce data imbalance, used binary cross-entropy loss, and systematically optimized hyperparameters using Grid Search. Furthermore, model performance was comprehensively evaluated using multidimensional metrics such as not only accuracy but also Hamming loss, precision, and F1 score. This approach provides a more balanced and reliable evaluation compared to previous multimodal studies.
- In our study, we compared not only the performance of the hybrid model but also the results of the text-based and image-based single models. This clearly demonstrates the relative contributions of the modalities: the text modality is more powerful than the image modality, but the combination of both modalities achieves the lowest Hamming loss (0.0694) and the highest precision (0.7319).
- An existing movie poster dataset has been extended by integrating corresponding movie plots and synopses, thereby transforming it into a comprehensive multimodal dataset suitable for both image—and text-based classification tasks. This enhanced dataset, which combines visual and textual modalities, is expected to serve as a valuable resource for future research in multimodal movie genre classification and related fields [38].
- In conclusion, this study proposes a high performance, interpretable, and methodologically novel hybrid model for multi-label movie genre classification by combining the ConvNeXt–DistilBERT–XGBoost architecture. In this respect, it advances existing multimodal approaches in the literature in both architectural and optimization aspects.

The remainder of the paper is structured as follows. The Literature Review section summarizes previous studies addressing the multi-label classification problem, discusses existing research gaps, and explains how the proposed model addresses them. Additionally, a comprehensive literature review table is provided to give readers a clearer overview of prior works. The Materials and Methods section introduce the proposed image-based, text-based, and hybrid models, and details the training and optimization procedures, evaluation metrics, and implementation settings. In the Results and Discussion section, the performance of the proposed models is evaluated according to the selected metrics, and the study most closely related to our work is discussed in detail. Finally, the Conclusion section summarizes the proposed model and presents the research limitations, future studies, and the managerial insights of our research.

2 Literature review

Automatic movie genre prediction, a multi-label learning problem, is one of the most intensively studied topics in both natural language processing and computer vision. The goal here is to classify a movie into multiple genres simultaneously by leveraging features obtained from different modalities. This section presents multimodal, text-based, and image-based approaches found in the literature for this purpose. We discuss existing gaps and how our proposed model addresses them.

The trend that has become leading in the literature in recent years is multimodal approaches that combine textual, visual and auditory features. Different multimodal strategies were evaluated by the study [1] to classify movie genres using poster, summary, subtitle and the highest result was obtained in the late fusion method with an F1 score of 0. 649.It has been observed that complementarity between modalities improves performance. The study of [2] obtained a value of 0.732 according to the AU(PRC) metric with the intermediate fusion method called LAVO-S in their four-mode system including poster, synopsis, audio and video data. In another multimodal study [3], in addition to text and visuals, metadata features including actor, director and author information were included and an F1 score of 0.65 was achieved with the weighted average ensemble model. In the study conducted by [4], a knowledge graph-supported IDKG model was developed, a self-supervised attention mechanism was designed that models director, actor and genre relationships, and a value of 0.849 was obtained in the Micro-F1 score. In another multimodal study [5] performed genre prediction using image, audio, and text modalities on movie trailers. Transformer-based fusion was applied by combining pre-trained CLIP, AudioTag, and DistilBERT multimodal networks. The hybrid model evaluated on the MovieNet dataset achieved a mAP value of 0.66. This study [6] developed an InceptionResNetV2+GRU based model for genre classification using video and audio data together and obtained a value of 0.79 according to the F1 Score metric with the Maximum Gap Thresholding (MGT) algorithm, which is a multi-label decision making method and provides dynamic thresholding. A multimodal method for automatic genre prediction using movie posters and summaries was proposed [7] Visual features were obtained with the DenseNet-19 model with Class Activation Mapping (CAM) module, and textual features were obtained with the BERT model. The features obtained from each model were sent as input to the multimodal fusion module for genre prediction. The fusion model was a simple Feed Forward Neural Network consisting of 3 linear layers and a final classification layer. MM-IMDB dataset was used for model evaluation. As a result of the evaluations, it was seen that the model was successful according to F1-Weight and F1-Macro metrics. The study of [8] methods were proposed for the classification of multimodal data and multimodal fusion methods were investigated and these methods were compared in terms of accuracy and calculation speed. Text and image data were classified separately, and multimodal classification was performed by using them together. FastText and ResNet architectures were used to obtain textual and visual features, respectively. Multimodal datasets named Food101, MM-IMDB and FlickrTag were used to test the models. As a result of the experiments, it was seen that including the image instead of just the text increased the performance in the classification task. Multi-label classification of movie genres was addressed in a multimodal manner [9]. A dataset consisting of movie titles in the TMDb movie database; trailers, video clips, subtitles, summaries, and movie posters was created. Various feature extractors were calculated for different data types and classifiers were trained. Classifier and feature combinations were examined using the late fusion strategy. As a result of the experiments, it was seen that combining an LSTM trained on movie summaries and another LSTM trained on movie subtitles achieved the best results in F-Score and AUC-PR metrics. In the multimodal study [10], the evaluation of the usefulness of visual, textual and metadata-based information to predict the genre of the movie and analyze its visual features using movie trailers was presented. Moviescope dataset was used for the experiments. Linear Regression, K Nearest Neighbors, Decision Tree, Random Forest and Artificial Neural Networks are just some of the classification algorithms used and compared in this research study. A large-scale study comparing the effectiveness of visual, audio, text and metadata-based features for predicting high-level information about movies, such as genre or budget estimate, was presented [11]. A comprehensive study on feature extraction methods for representing video and

text was presented and also the extent to which different modalities complement each other was shown. A new large-scale dataset, Moviescope, consisting of 5000 movies with associated movie trailers, posters, plots and metadata was also presented. In another multimodal classification study, a new movie genre classification method was proposed using a combination of problem transformation techniques, text vectorizers and machine learning classifier models [12]. In study of multi-label movie genre model, Support Vector Machine was combined with Decision Tree which optimizes the value of parameters and Parameter Optimized Hybrid Classifier was added [13]. A Multi-Modal approach leveraging Shot information was proposed to effectively classify video genres [14]. MMShot leverages both auditory and verbal information in addition to visual modality to further improve the performance. In their study [15] examined whether the combination of a movie's poster and summary could provide more accurate results for genre prediction than using either the poster or summary alone. GloVe was used for extracting textual features, while the ResNet50 model was utilized for extracting visual features.

When the multimodal studies reviewed so far are examined, it is seen that combining features extracted from different modalities provides a systematic improvement in genre classification.

Text-based approaches rely on plot synopses or subtitles as the main source of information. SAS-MC model was suggested by the study of [16] for multi-label movie genre classification based on movie summary. SAS-MC used the self-attention mechanism designed to automatically learn the importance of features at each time step by analyzing different views. The study of [17] suggested a new multi-label model that performs genre classification using classical machine learning methods using subtitles of movies. However, this classical machine learning approach could not adequately represent the contextual meaning. Later, in the study by [18], an explainable Transformer ensemble model was proposed by combining BERT, RoBERTa and DistilBERT language models for movie genre prediction with movie summaries. The model achieved a value of 0.75 according to the F1 Score metric and made the decision processes of the model interpretable with the LIME method.

These studies show that language-based models are powerful in contextual meaning inference, with Transformer architectures in particular becoming decisive in genre prediction.

In visual-based studies, researchers used visual or auditory cues obtained from movie posters and trailers in genre classification. A multimodal deep recurrent model was proposed for movie genre classification [19] A new Gated Recurrent Unit (GRU) based structure was designed to derive the spatial temporal visual features of movie frames. Then, by combining video features with audio features processed with 1D-CNN, a prediction model with increased accuracy, particularly for small sample sizes, was proposed, which was relatively low in computation cost. Multi-label trailer database (LMTD) was used as the dataset. This study emphasizes that genre emerges not at the frame level but within the context of the entire trailer, and advocates for explicit modeling of the temporal dimension. In this study [20] extracted twelve semantic design features, such as color, composition, emotion, and lighting, from movie posters and used a framework called SEMPD to predict movie genres. They performed a multi-label classification task using Naive Bayes and RakEL machine learning methods.

One of the first studies [21] to predict genres from trailer audio components, converted audio components into spectrograms using the Librosa library and analyzed them with a 1D CNN model. This method achieved 79% accuracy.

These studies show that visual and auditory attributes can also play an important role in genre prediction, and emotional tone and compositional information are particularly decisive.

Outside the multimodal field, some recent studies have addressed the issues of intermodal knowledge transfer and generative modeling. Multimodal book genre classification study using book cover images and book titles was proposed [22] ResNeXt-101 network with added Gram and SE layers was used to predict genres from cover images and EXplicit InterActive Network (EXAN) network from book titles. A multimodal fusion architecture was designed for multimodal classification, utilizing an attention mechanism across modalities. This study [23] proposed a cross-domain learning model based on RoBERTa and DistilBERT by matching book descriptions with movie summaries based on genre similarities. The model generated movie recommendations from book data and produced values of 0.908 and 0.96 according to the LRAP and Precision metrics. In the study by [24] developed a generative system called Spectrogrand that generates visual and auditory scenes from text prompts,

demonstrating that multimodal representations can be used not only in classification but also in generative applications. In another text-based study [25], Transformer-based language models such as BERT, RoBERTa, and DistilBERT were adapted to multi-label text classification and achieved a significant increase in F1 metric performance by proposing learnable threshold layers instead of fixed thresholds.

One of the multi-label multi-modal studies is music genre prediction [26]. Music genre prediction was made with deep learning techniques using audio, text and image data. MuMu dataset was proposed for this study. Additionally, an approach for multi-label genre classification based on the combination of feature embeddings learned with the latest deep learning methodologies was proposed. Experiments showed that the multi-modal approach outperformed single-modality approaches and the combination of three modalities gave the best result. In a similar study [27] compared BiLSTM, BERT, and DistilBERT models when classifying musical genres from song lyrics, demonstrating that Transformer-based models outperform RNNs in both speed and accuracy. This study [28] used DistilBERT to predict song genre, popularity, and release year, demonstrating that Transformer models can capture genre-distinguishing semantic patterns even in artistic texts.

Table 1 presents a summary literature review of multimodal studies on multi-label movie genre classification. This table provides a summary of the main multimodal studies in the field of multi-label movie genre classification, as described in the literature review. Each study is compared in terms of the modalities, core architectures, fusion types, datasets, best performance metric results, and unique contributions. The ConvNeXt–DistilBERT–XGBoost-based hybrid model proposed in this study is also included. This table aims to highlight the innovative aspects of the proposed model by highlighting trends and existing research gaps in the literature. This review demonstrates that existing research is generally limited to classical CNN and BERT-based models, while fusion strategies are mostly neural network-based and have limited handling of data imbalance and multi-label structures. Therefore, the proposed ConvNeXt–DistilBERT–XGBoost-based hybrid model addresses these shortcomings and brings innovation to the literature at the architectural, fusion, and optimization levels.

The literature review reveals that despite significant advances in multi-label movie genre classification, several research gaps remain. The majority of existing studies [2, 4, 9] are limited to classical CNN (e.g., ResNet, VGG) and Transformer (e.g., BERT)—based architectures. The lack of systematic evaluation of modern and optimized networks such as ConvNeXt, Swin Transformer, and DistilBERT for this problem constitutes a significant architectural shortcoming. Furthermore, the fusion strategies used in multimodal learning have not been analyzed in depth. Early or late fusion methods are generally preferred in the literature, while tree-based or boosting-based fusion mechanisms are almost never used. This limits the ability to interpretably model intermodal relationships. Another limitation, data imbalance and multi-label data structures have not been adequately addressed in most studies; training and test sets are randomly separated, and layered balancing techniques such as Multi Label Stratified Shuffle Split (MLSS) are not utilized. This has led to the overrepresentation of some genres and a decrease in the generalizability of the model. Moreover, performance evaluations have generally used only accuracy or F1 metrics, ignoring metrics specific to multi-label classification, such as the Hamming loss. Additionally, many multimodal studies in the literature discuss the performance contribution of the combined use of textual and visual modalities only in general terms, without systematically analyzing the impact of each modality individually. Therefore, the question of which modality contributes more to classification accuracy remains an open research problem. In this context, our study addresses these research gaps and makes an original contribution to the literature by first combining modern network architectures such as ConvNeXt and DistilBERT, proposing an XGBoost-based late fusion strategy, applying robust optimization approaches such as MLSS and Grid Search for imbalanced datasets, and empirically analyzing modality contribution.

3 Material and method

In this section, the architectures of the proposed models, the training and optimization methods, the evaluation metrics and implementation details are described in detail.

Table 1 A comprehensive literature review table

Authors Year	Modalities	Core models	Fusion strategy	Dataset	Best results	Key contribution
[26] 2017	Audio + Text + Image	CNN (Audio) + TF − IDF (Text) + ResNet-101 (Image) + Cosine	Late fusion	MuMu	AUC = 0.931	Music genre classification One of the pioneering examples of multimodal deep learning in the field of music information retrieval (MIR)
[8] 2018	Image + Text	ResNet (Image), FastText (Text)	Additive / Gated Fusion / Bilinear Fusion (best)	Food-101, MM-IMDb, FlickrTag	F1 = 0.62 (MM-IMDb)	Multimodal classification Scalable fusion strategies
Authors Year	Modalities	Core models	Fusion strategy	Dataset	Best results	Key Contribution
[16] 2018	Synopsis	Convolutions through time (CTT) TF-IDF + SVM CNN + LSTM	Label correlation layer	LMTD	Hamming loss = 0.077	Movie genre classification SAS-MC = 1D time-based convolution approach
[20] 2019	Poster	SEMPD + LP + MNB	–	MovieLens100K	Jaccard = 0.4178	Movie genre classification Semantic extraction from poster design
[11] 2019	Video + Audio + Text + Poster + Metadata	fastText (GloVe) (Text model) fastVideo (Video model) VGG16 (Image model) XGBoost & RandomForest (Metadata model)	Attention based fusion	Moviescope	mAP = 75.3	Movie genre and budget estimation Large scale multimodal film analysis
[15] 2020	Poster + Synopsis	VGG16 + Word2Vec ResNet50 + Word2Vec VGG16 + GloVe ResNet50 + GloVe (best)	Late fusion (concatanetion)	TMDb - 42k	F1 = 0.674	Movie genre classification Early multimodal example
[27] 2021	Lyrics	BiLSTM, BERT, DistilBERT	–	Kaggle: Detkov, MetroLyrics	Acc = 0.7129	Music genre classification Transformer vs RNN comparison
[7] 2021	Poster + Synopsis	DenseNet − 169 + CAM (Class Activating Module) (Image model) BERT (Text model)	MMF- Multi module fusion (concatanetion)	MM-IMDb	F1 - Micro= 0.675	Movie genre classification Visual–text integrated network
[13] 2021	Subtitle	TF − IDF + SVM + DT	–	YIFY Subtitle	F1 = 0.96	Movie genre classification A hybrid system (POHC) based on parameter optimization of SVM and Decision Tree
[1] 2022	Poster + Synopsis + Subtitle + Trailer	fastText (Skip-gram) + LSTM (Text model) ResNet-152 (Visual model)	IMF, DMF, Late fusion (Sum/Product/ Max)	TMDb	F1= 0.649	Movie genre classification Comparison of multimodal fusion strategies
[17] 2022	Subtitle	SelectKBest+ chi square + KNN, NN, LR, SVM, NB	–	yifysubtitles. com	Precision= 0.78	Movie genre classification Subtitle-based innovation

Table 1 (continued)

Authors Year	Modalities	Core models	Fusion strategy	Dataset	Best results	Key contribution
[25] 2022	Text	BERT, DistilBERT, RoBERTa , DeBERTa for English Camem-BERT, FlauBERT for French	Threshold and custom layers	for English; AAPD, Reuters-21578 for French; MFHAD (new)	F1 = 0.91 (Reuters) 93.94 (AAPD)	Multi-label text classification Transformer adaptation, Threshold optimization strategy
[14] 2022	Visual + Audio + Language	MM-Shot (Multi-Modality approach leveraging Shot)	Early / Intermediate (Best) / Late fusion	MovieNet , condensed movies	mAP= 61.57	Movie genre classification MMShot - Learning visual, audio and linguistic features together
[9] 2022	Video + Audio + Poster + Subtitle + Synopsis	InceptionV3 (Visual Model) LSTM (Text Model)	Late Fusion	TMDb 10K	F1 = 0.674	Movie genre classification Multimodal late fusion classification
Authors Year	Modalities	Core models	Fusion strategy	Dataset	Best results	Key contribution
[10] 2022	Video + Text	VGG-16 (Visual model) + TF − IDF (Text model) + NB, KNN, DT, RF, ANN (best)	Late Fusion	MovieScope	F1 = 0.67	Trailer genre classification Simplified multi-modal approach
[4] 2023	Poster + Plot + Metadata (actors, directors)	IDKG (Incorporating Domain Knowledge Graph) = Knowledge Graph (KG) + Attention Teacher (AT) + Genre − Centroid Anchored Contrastive Learning (G- CAL)	Knowledge graph- based multimodal fusion	MM- IMDb 2.0	F1 (Micro) = 0.85	Movie genre classification Knowledge graph and contrastive learning
[3] 2023	Poster + Synopsis + Metadata (actor, director, writer)	CNN / VGG-16 (Image Model) LSTM (Text Model)	Weighted average ensemble	IMDB + Kaggle movies.csv	F1 = 0.65	Movie genre classification Actor-director-writer information was included in the classification process for the first time.
[19] 2023	Visual + Audio (Trailer)	ResNet-152 + GRU + SVM	Early fusion	LMTD	Hamming loss = 0.13	Movie genre classification The first GRU-SVM based model to propose a visual + audio integrated structure
[12] 2023	Metin	LP + TF-IDF + SVM	–	IMDb	Acc = 0.95	Movie genre classification Binary relevance and label powerset comparison
[22] 2023	Book Cover + Metadata (title, author, publisher)	EXAN (Explicit interActive Network) (Text Model) GR-ResNeXt-101 (Image Model)	Attention- based fusion	BookCover-28-ext, Arabic BookCover-28	Acc = 0.6909	Book genre classification Multimodal network with book covers
[2] 2024	Poster + Synopsis + Audio + Video	GloVe + Bi-LSTM (Text Model) ResNET-50 (Visual Model) VGGish + Bi-LSTM (Audio Model)	Decision level fusion Intermediate fusion (best)	LMTD-9	AU(PRC) = 0.732	Movie genre classification 4 modality integration + different fusion comparison

Table 1 (continued)

Authors Year	Modalities	Core models	Fusion strategy	Dataset	Best results	Key contribution
[21] 2024	Trailer (Audio- Spektrogram)	1D CNN	–	IMDb (2020-2022)	Acc = 0.79	Trailer genre classification Audio-based film genre classification
[5] 2024	Trailer (Video + Audio)	CLIP (Video Model) + AudioTag (Audio Model)	Multi-transformer fusion	MovieNet	mAP = 66	Trailer genre classification Pretrained multi-modal fusion, long sequence processing
Authors Year	Modalities	Core models	Fusion strategy	Dataset	Best results	Key contribution
[24] 2024	Text	AudioLDM2 (Text to Music) Stable DiffusionXL (Text to Audio) Spectrogrand=Diffusion + GANAI	–	HouseX, BOTOR-NOT, SDXL	F1= 0.90	Audiovisual generation Spectrogrand-a spectrogram-based Generative AI (GenAI) system
[28] 2024	Lyrics	DistilBERT	–	Genius	Genre Acc = 0.65, Popularity Acc=0.79	Song Lyric genre, popularity and release year estimation Evaluation of Distil-BERT in the lyrics domain
[6] 2025	Video + Audio	InceptionResNetV2 + GRU (Visual Model) + Spectral Centroid (SC) (Audio Model) + MGT (Maximum Gap Thresholding)	Early intermediate fusion	MovieLens 20M -> Refined - MTSD (R- MTSD)	F1 = 0.79	Movie scene classification Scene-level multi-modal analysis, MGT decision making method
[18] 2025	Plot	BERT + RoBERTa + DistilBERT = Transformer Ensemble Model	Soft voting fusion	Trailers12K, LMTD9	F1 = 0.75	Movie genre classification The first interpretable transformer ensemble
[23] 2025	Text + Image	RoBERTa + DistilBERT	Cross-domain fusion	Goodreads + IMDb (Kaggle)	LRAP= 0.908	Book- movie recommendation Cross-domain knowledge transfer (Book-Movie)
Our Model 2025	Poster + Synopsis	ConvNeXt + Distil-BERT + XGBoost	Late fusion	IMDb	Hamming loss =0.0694, Precision = 0.7319, F1 = 0.62	Movie genre classification A hybrid model where modern architectures (ConvNeXt and DistilBERT) are used together for the first time, XGBoost is applied with tree-based late fusion, and Grid Search and MLSS are optimized for imbalanced multi-label data.

3.1 Image-based model: ConvNeXt

One of the innovative aspects of this study is the image-based genre prediction model developed using the ConvNeXt architecture. To the best of our knowledge, this model has not previously been used within a hybrid framework for movie genre prediction, and it is applied to the genre prediction problem for the first time in this study.

With the rising popularity of Vision Transformers (ViTs), the ConvNeXt architecture was developed in 2022 to enhance the performance of CNNs. ConvNeXt preserves the design simplicity of traditional convolutional networks and represents a progressively modernized version of the ResNet architecture. Key modernization steps include the use of 7×7 kernels, which enable the network to learn a broader context; replacing Batch Normalization with Layer Normalization; utilizing Depthwise Convolution for computational efficiency; and adopting the GELU activation function instead of ReLU [29].

- The system was developed using Python, along with the TensorFlow and Keras frameworks.
- The data used for the image-based model were obtained from the IMDB dataset [30]. The genres selected for classification are Action, Adventure, Animation, Biography, Comedy, Crime, Drama, Documentary, Fantasy, Family, History, Horror, Music, Mystery, Romance, Science Fiction, Short, Sport, Thriller, and War. Due to the relatively low quality of movie posters produced prior to 1970, this year was set as the threshold, and the dataset comprises films released between 1970 and 2018. The dataset exhibits a significant class imbalance across genres; for example, the Drama category contains 14,585 films, whereas the Short Film category includes only 551. This imbalance makes the problem even more challenging.
- Due to the nature of the problem, the dataset has a multi label structure. For example, while a film may belong to the adventure genre, it can also be categorized as action at the same time. Each film in the dataset has at least one and at most three genres. Furthermore, the number of films per genre is not balanced. For instance, while the drama genre contains the highest number of films, genres such as war, short film, and sport are represented by fewer films. This imbalance is one of the challenging aspects of the study.
- In order to preserve the genre balance within the dataset and prevent data loss, the dataset was split into training and test sets in an 80%-20% ratio using the iterative stratified 5-fold cross validation technique [31] %20 of the training set was further separated as a validation set. Subsequently, the genre values were converted into numerical vectors that can be interpreted by the model using the multi hot encoding method.
- The ConvNeXt model, obtained using the Keras library, is loaded into the system along with its weights. The model takes 224×224 sized RGB images as input. To benefit from the feature extraction capability of the pretrained ConvNeXt model, the layers before the classification layer are frozen, and the final layer is replaced with a sequence of fully connected layers consisting of 1024, 512, 256, 128, and 20 units, respectively, tailored to the requirements of our problem. All added fully connected layers use the ReLU activation function. The output layer utilizes the Sigmoid activation function, which is suitable for multi-label classification tasks. Additionally, batch normalization layers are inserted between the fully connected layers to prevent overfitting. The training diagram of the image-based model is presented in Fig. 1.
- As illustrated in the figure, the stratified 5-fold cross-validation method was applied to the dataset. The entire dataset was split into five folds of equal size, and each fold was trained for 10 epochs.
- The newly added fully connected layers (FCL) are trained using the four folds designated as the training set. After 10 epochs, the model is evaluated on the validation set based on accuracy and loss metrics, and the results are recorded. The model trained for the selected fold is saved as "Model v1." This process is repeated separately for each of the five folds. In each case, the results obtained according to the evaluation metrics are saved as "Model v2," … "Model v5." Among these versions, the model with the highest performance is considered the "Final Model."
- In the selected final model, the previously frozen layers are unfrozen and fine-tuned. The model is retrained using a very low learning rate, such as 0.001. The purpose of this step is to adapt the pretrained features to the genre classification problem. At this stage, both the pretrained layers of the ConvNeXt network and the

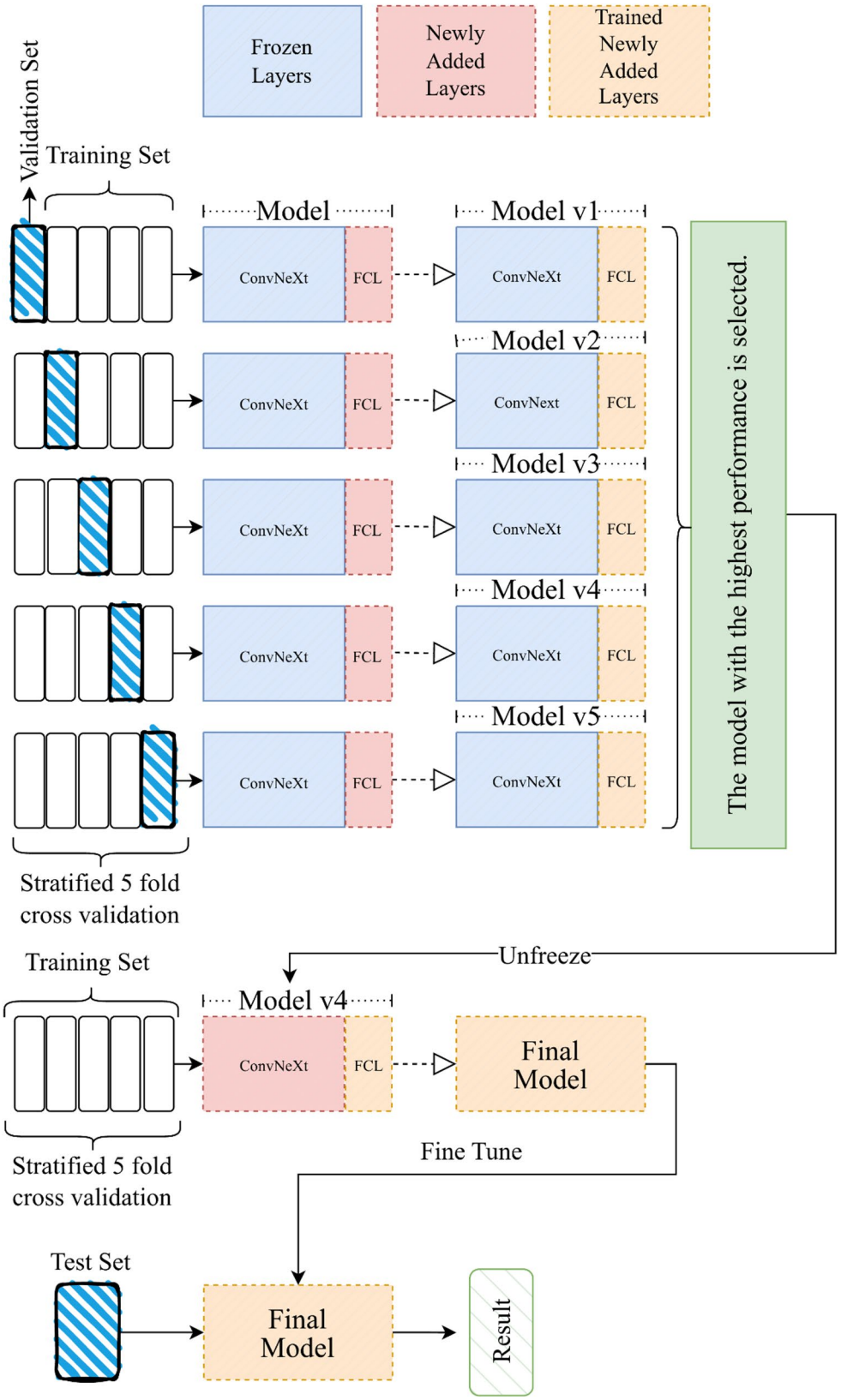

Fig. 1 The training diagram of the image-based model

newly added layers are trained. In this way, the final model is better adapted to the specific requirements of the problem.

- Model hyperparameters were: Binary Cross Entropy loss, Adam optimizer (learning rate 0.001), batch size 32, and 10 epochs.
- The final model obtained through fine-tuning was evaluated on the test dataset using Hamming loss, Precision and F1 score metrics and the results were recorded.

3.2 Text-based model: distilbert

Proposed by the Google team in 2017, the Transformer architecture has become the foundation for pretrained language models in natural language processing and has led to major advancements in the field. Based on the self-attention and multi head mechanisms, this model processes data in parallel rather than sequentially, unlike traditional models. This property makes the model faster, more flexible, and more effective [32].

BERT (Bidirectional Encoder Representations from Transformers) is a pretrained deep learning model that has pioneered many innovations in natural language processing applications. Unlike previous language models, the Transformer based BERT model learns word representations bidirectionally, considering both left and right contexts, rather than only in a single direction [33].

DistilBERT is a machine learning model proposed by the Hugging Face team as a smaller, faster, cheaper, and lighter version of BERT for use in natural language processing applications. It is a lightweight and efficient language model that can be fine-tuned for other tasks and requires less computational resources. DistilBERT is trained using the principle of knowledge distillation, which involves training a smaller "student" model under the supervision of a larger "teacher" model. In this case, BERT serves as the teacher and DistilBERT as the student. DistilBERT learns by distilling knowledge from the outputs and predictions of the BERT model [34].

The Transformer based DistilBERT was chosen as the starting point for the text-based movie genre classification model.

- The dataset for the image-based genre prediction model was expanded by adding movie plots and summaries, making it suitable for text-based classification as well. Plots and summaries were collected using web scraping with a Python script based on movie Ids [38]. The plot and summary were then combined into a single column in the dataset. Both the plot and synopsis texts were collected and merged because some of the movies contained only the plot while others included only the synopsis. This approach ensured the inclusion of the maximum number of films with sufficiently long summaries. Subsequently, the summaries were cleaned, and those containing fewer than ten words were removed from the dataset. As a result, the final dataset consisted of 29,286 movies.
- The dataset was split into training 80% and test 20% sets, with 20% of the training set further used for validation. The splitting was performed using the Multi Label Stratified Shuffle Split (MLSS) method [35]. This approach ensures balanced label distributions across the training, validation, and test sets in multi-label classification problems. MLSS prevents rare genres from being excluded from any subset and guarantees that each genre is adequately represented in all sets.
- After the data preprocessing stage, the entire dataset was tokenized using the " DistilBertTokenizer ". This tokenizer converts raw text into numerical IDs that the DistilBERT model can process by splitting the text into words, sub words, or characters. After tokenization, the token length for each film was set to 128.
- The DistilBERT model for multi label classification was defined using " DistilBertForSequenceClassification " from the Hugging Face " transformers " library.
- After adding the model, a "CustomDataSet" class was created to load our own dataset for training and testing. The model was trained using the " Trainer" API with the following hyperparameters: token length 128, 5 epochs, batch size 8, and threshold 0.3.

- To optimize performance, training was repeated with thresholds of 0.3, 0.35, 0.4, 0.45, and 0.5 for both 128 and 256 token lengths.
- The sigmoid activation function was used in the output layer.
- Model performance was evaluated based on Hamming loss, Precision and F1 score metrics and the results were recorded.

3.3 Hybrid model: feature fusion with XGBoost

The image model constructed with the ConvNeXt network was combined with the text model created using DistilBERT under the optimal conditions determined by hyperparameter optimization (token length of 256 and threshold value of 0.5) via the XGBoost machine learning algorithm. The integration of ConvNeXt and DistilBERT networks for movie genre classification is one of the innovative aspects of this study. To date, no previous research has combined these two models for this problem using the XGBoost algorithm. The presence of features from two different modalities allows the models to compensate for each other in cases where either visual or textual information may be insufficient. In this way, more accurate and reliable classification can be achieved. The schematic of the proposed multimodal hybrid model is presented in the Fig. 2.

XGBoost is a popular and powerful machine learning algorithm based on decision trees. Essentially, it is an optimized and extended version of the Gradient Boosting algorithm. In Gradient Boosting, the goal is to add a new model at each step that reduces the errors of the previous model [36].

Grid Search is an optimization method used to find the set of hyperparameters under which the model produces the best results. The Grid Search method systematically evaluates all possible combinations within user defined hyperparameter ranges, assesses the model's performance for each combination, and selects the parameter set that yields the highest performance [37].

- The combined data obtained from the image and text-based models was split into training 80% and test 20% sets using the iterative stratification method, which ensures a balanced division of the dataset.
- The most important features were extracted from the poster data using the ConvNeXt model, and from the summary data using the DistilBERT model.
- The visual features are represented as a 20-dimensional probability distribution vector, and the textual features are also expressed as a 20-dimensional probability distribution vector.

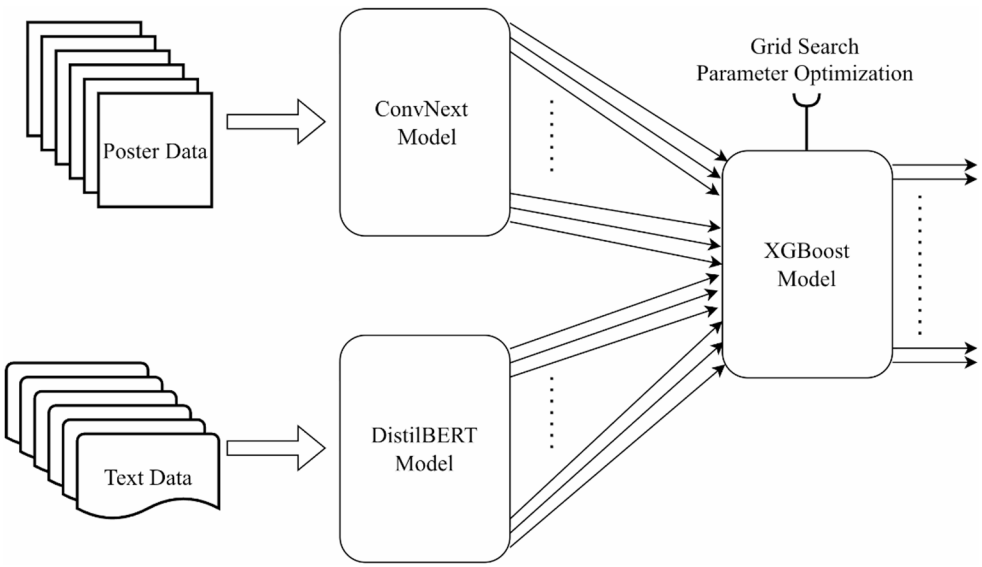

Fig. 2 Multimodal hybrid model

- These two vectors are concatenated, resulting in a 40-dimensional probability distribution vector, which along with the 20-dimensional output vector, the training set, and the XGBoost model parameters is provided to the Grid Search algorithm.
- The input parameters for the Grid Search algorithm, which was used for hyperparameter optimization of the multimodal hybrid model, were defined as follows:

  o Tree depth (estimator__max_depth): [3, 5, 7]
  o Learning rate (estimator__learning_rate): [0.01, 0.1, 0.2]
  o Number of trees (estimator__n_estimators): [50, 100, 200]

- Hamming loss was selected as the main metric for the Grid Search algorithm to optimize. Additionally, evaluations were performed using precision, recall, and F1 score metrics.
- Grid Search generates all possible combinations of the specified hyperparameter values.
- For each combination, the XGBoost model is trained, and its performance is evaluated.
- The highest performance according to the Hamming loss metric was achieved with a tree depth of 3, a learning rate of 0.1, and 50 trees.
- Grid Search returns the XGBoost model trained with these optimal hyperparameters.
- The final model is tested, and results are obtained based on Hamming loss, Precision and F1 score metrics.

4 Result and discussion

The performances of the three proposed models for multi label genre classification are presented in Table 1, 2 according to the hamming loss, precision, and F1 score metrics.

Hamming loss measures the average error rate of a model by calculating the percentage of incorrectly predicted labels across all samples in the dataset, the lower the value, the better the model's performance. According to the results, it can be observed that the lowest hamming loss value was achieved by the hybrid model with 0.0694. This indicates that the hybrid model, which utilizes both visual and textual features, has a lower error rate compared to models based solely on a single data type such as image or text. The text-based model achieved a lower hamming loss with 0.0753 than the image-based model, suggesting that textual information may be more decisive than visual data for film genre prediction. The image-based model, with a hamming loss of 0.0971, had the highest error rate, supporting the notion that visual data alone is not sufficient and that approaches utilizing text or multimodal data yield better results. Overall, all three models achieved over 90% accuracy, demonstrating that the genre prediction results of these models are acceptable and successful according to hamming loss metric.

Precision measures the proportion of true positives that are actually correct. A high precision value indicates that the model produces a low rate of false positives. According to the precision metric, the text-based model outperforms the image-based model, while the hybrid model surpasses both. The hybrid model achieves the highest precision rate, indicating that 73.19% of its predicted genres are truly correct.

The text-based model ranks second, with a precision value of 0.6491, and performs significantly better than the model based solely on visual data. The image-based model, with a precision value of 0.3310, has a rather low precision, demonstrating a high tendency to produce false positives meaning that most of its predicted genres

Table 2 Comparison of model performance based on metrics

Model	Hamming loss	Precision	F1 Score
Image-based model	0.0971	0.3310	0.4028
Text-based model	0.0753	0.6491	0.6213
Hybrid model	0.0694	0.7319	0.6200

are incorrect. Although this rate may be acceptable for incorrectly predicted genres in a film genre classification application, it is clear that the model needs improvement in terms of this metric.

In conclusion, the fact that the hybrid model has the highest precision shows that combining textual and visual information significantly increases prediction accuracy.

The F1 score is the harmonic mean of precision and recall, and it is especially used to measure overall model performance in cases of class imbalance. A high F1 score indicates that the model makes accurate predictions while also minimizing misclassifications. When comparing the models based on the F1 score, the highest value was obtained by the text-based model with the value of 0.6213. The F1 score of the hybrid model with the value of 0.6200 is very close to that of the text-based model, and both models perform significantly better than the image-based model with the value of 0.4028. This finding demonstrates that textual information contributes more to genre prediction than visual information, and that combining textual and visual features in the hybrid model has a positive impact on overall performance. The low F1 score of the image-based model indicates that both its precision and recall are relatively weak. In conclusion, the F1 scores of both the text-based and hybrid models are close to each other, indicating that textual information is particularly important for multi label film genre classification, although the hybrid approach also provides competitive performance.

To maximize the performance of the text-based genre classification model, various optimization procedures were performed on certain model parameters. There are two key parameters that can be adjusted before training and that affect performance: token lengths (L, length) and the threshold value (T, threshold). In this study, token lengths of 128 and 256, and threshold values of 0.30, 0.35, 0.40, 0.45, and 0.50 were selected. The results obtained from the metrics based on these parameters are presented in the graphs below. Figures 3, 4 and 5.

Examining the Hamming loss metric graph reveals that all values are below 0.085. It is clearly observed that as the threshold value increases, the Hamming loss decreases. Additionally, when the token length is set to 256, the Hamming loss is lower across all threshold values compared to when the token length is 128. The lowest Hamming loss value of 0.0753 was obtained when the token length was set to 256 and the threshold was 0.50. A loss value of 0.0753 indicates that the model achieved an accuracy rate of 92.47%, which is a very good result for genre prediction. As shown in the figure, the hyperparameter optimization experiment improved the Hamming loss by approximately 0.01 between the best and worst cases.

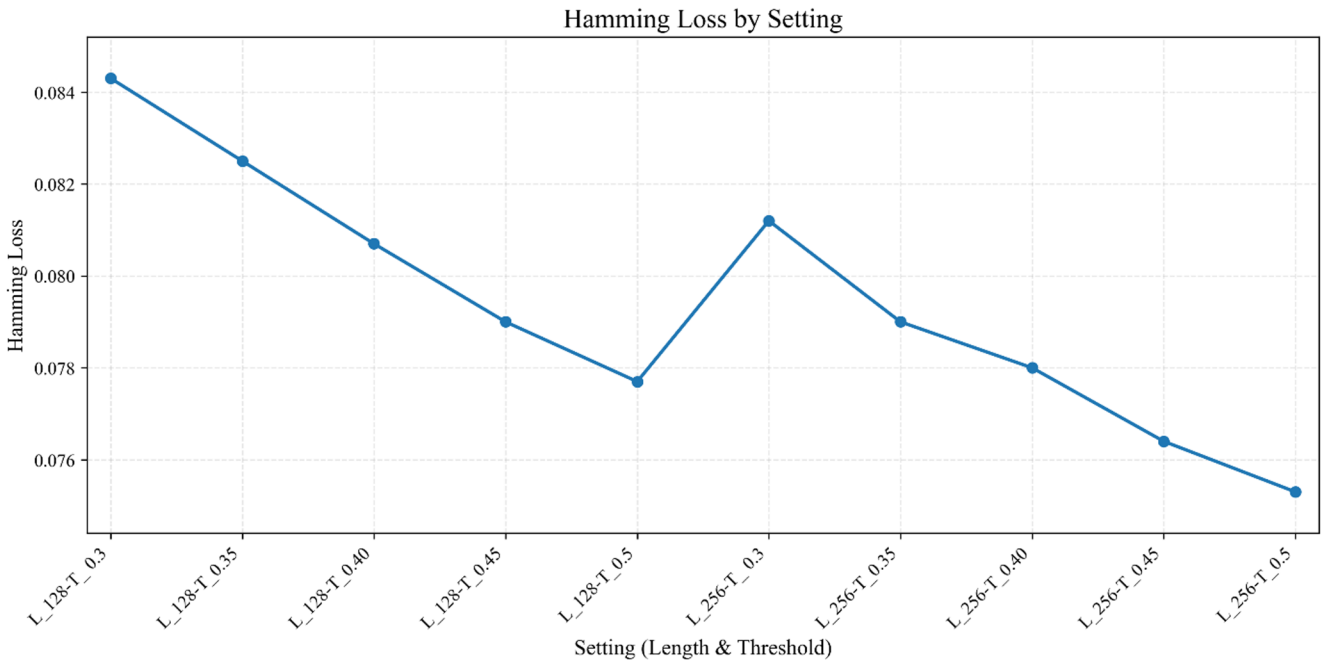

Fig. 3 Optimized model hamming loss graph

 https://doi.org/10.1007/s00521-026-11852-3

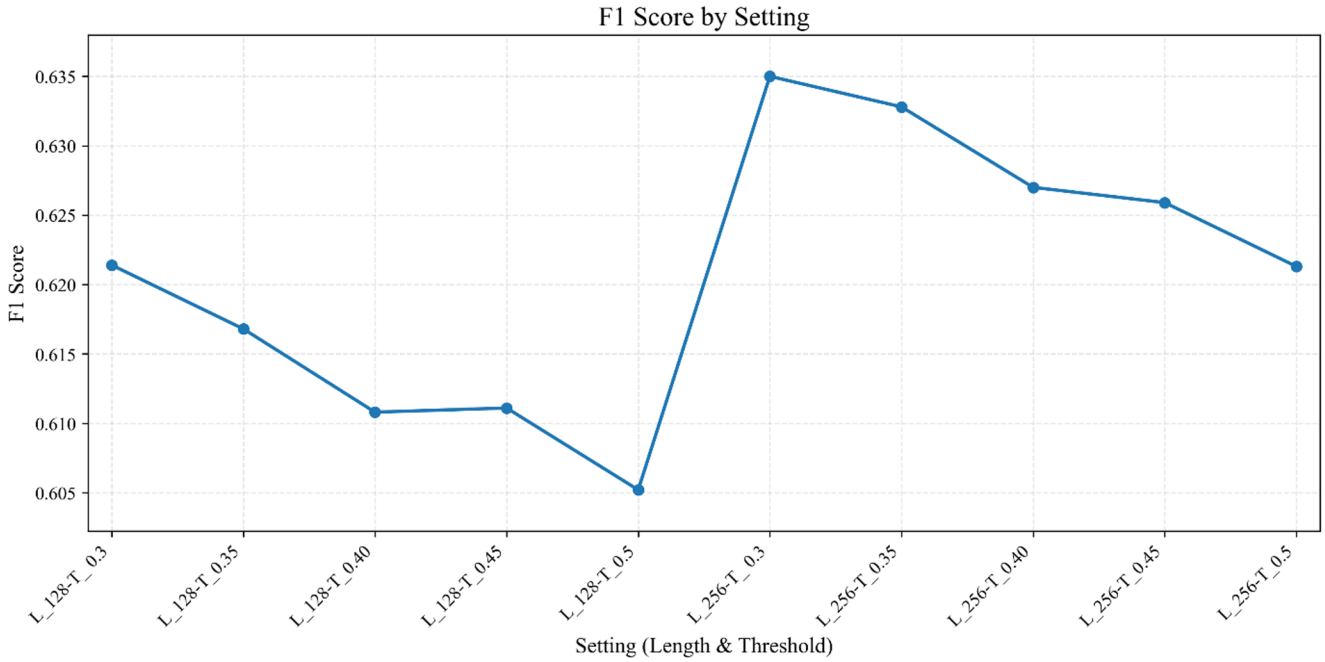

Fig. 4 Optimized model F1 score graph

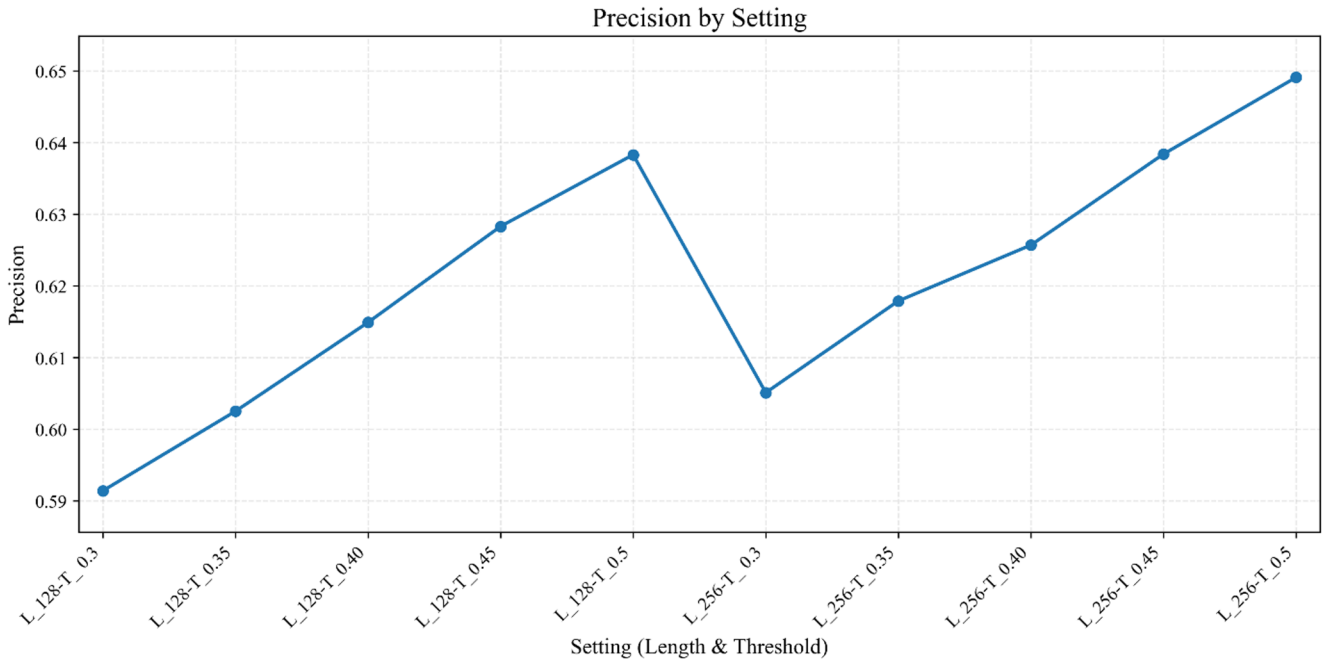

Fig. 5 Optimized model precision graph

When examining the results obtained for the F1 score metric, it is generally observed that increasing the threshold value leads to a decrease in F1 score. However, increasing the token length from 128 to 256 results in a higher F1 score. The highest score was achieved with a token length of 256 and a threshold value of 0.30. Overall, while there are no substantial differences, selecting a token length of 256 yields approximately a 1% increase in F1 score.

According to the precision metric, the highest performance was achieved with a token length of 256 and a threshold value of 0.5, reaching 0.64. The lowest performance was observed when the token length was 128 and the threshold was 0.3. The graph clearly shows that increasing the threshold value improves model performance when the token length is fixed. Additionally, increasing the token length from 128 to 256 also leads to higher model performance, as reflected in the results.

Based on all these evaluations, it was observed that the highest performance according to the Hamming loss metric, which is considered one of the most important evaluation criteria in multi label classification problems was achieved under the conditions of a token length of 256 and a threshold value of 0.5 (L_256_T_0.5).

This study proposed a hybrid multimodal model combining ConvNeXt and DistilBERT for multi-label movie genre classification. The main findings are discussed below.

When the existing studies are evaluated, a review of the existing literature indicates that the multimodal approach proposed by [2] can be considered the closest to our proposed approach. They also proposed a hybrid model combining visual and textual modalities for multi-label movie genre classification. In their study, they extracted visual features from movie posters using a ResNet50 network and textual features from movie summaries using a BERT-based model. They combined these two features using a multi-layer fusion method to perform genre prediction. Similarly, both studies used IMDB-based datasets, targeted multi-label genre classification, and evaluated performance using the F1 metric. In these respects [2], is highly similar to this research in terms of its methodological framework, data types, and problem definition.

However, our study differs from [2] in several key respects. First, our model uses ConvNeXt and DistilBERT architectures together; these two architectures have not been evaluated together in the context of movie genre classification before. ResNet50 and BERT, used in [2], are relatively older architectures; in our study, ConvNeXt utilizes deeper, normalized convolutional structures, achieving high representational power in visual feature extraction. On the textual side, the DistilBERT model, despite being lighter than BERT, minimizes information loss, speeds up training and produces more efficient results. Furthermore, while [2] used a neural network-based fusion approach, our study combined visual and textual features with the XGBoost algorithm. This allows for learning correlations between features in the fusion layer using a tree-based structure, improving interpretability and generalization. In addition, in our study, the combined application of optimization and evaluation approaches such as Grid Search, Multi Label Stratified Shuffle Split (MLSS) and Hamming Loss ensured the stability of the model in a balanced data distribution and allowed systematic improvement of the hyperparameters.

In conclusion [2], study is the closest to our study in terms of problem definition; however, our study provides a more powerful hybrid model in terms of performance and efficiency, advancing existing multimodal approaches in the literature, thanks to the integration of more modern network architectures, tree-based fusion and advanced optimization techniques.

Studies found in the literature review [1, 7, 9, 11, 15] appear to be similar to our work. However, these studies have significant limitations. On the visual modality side, modern CNN architectures such as ConvNeXt were not used, and on the text side, older RNN/LSTM approaches were preferred over Transformer-based contextual models such as DistilBERT. Our study achieves higher accuracy with fewer modalities by using more modern architectures and a learnable late combination (XGBoost). Additionally, optimizing a metric specific to multi-label classification, such as Hamming Loss, during the training process takes our work beyond similar studies in the literature.

Based on these comparisons, we analyzed the internal behavior and limitations of our model in more detail. The results obtained in our study indicate that the text-based model (DistilBERT) achieves higher accuracy in genre classification compared to the visual-based model (ConvNeXt). This finding can be explained by the fact that movie summaries contain explicit semantic cues that directly reflect genres. Movie posters, on the other hand, are less specific in terms of genre information because they often contain indirect visual elements such as artistic style, color palette, and cast composition. But the lower Hamming Loss achieved by the hybrid model suggests that visual features complement textual information. However, the model's relatively high reliance on textual data may reduce the system's robustness in situations where textual content is incomplete, inaccurate, or

noisy. Therefore, future work plans to apply pre-training strategies to posters to strengthen genre representations and evaluate the performance of the hybrid model under conditions where textual data is partially masked or corrupted. It is anticipated that such work will enhance the generalization ability and robustness of multimodal genre classification systems.

The movie summaries and posters used in this study are based on a specific dataset compiled from the IMDb database. Therefore, the high performance achieved by the model may be specific to this data distribution and may not yield similar results across different data sources or cultural contexts. Because the design style, color preferences, and language use of movie posters can vary across cultures, it is acknowledged that the model's generalization ability may be limited. Therefore, future work plans to evaluate the model's performance on broader and culturally diverse movie datasets and apply transfer learning-based domain adaptation approaches. Thus, the proposed hybrid model is expected to demonstrate higher generalization success across different movie genres and languages.

Additionally, the effect of the threshold value on model performance was analyzed. The effect of the threshold value used to convert the model's output probabilities to binary labels on multi-label classification performance was analyzed. Experiments in the 0.30–0.50 range showed similar trends across metrics. Hamming Loss values below 0.085 across the entire range indicate that the model has a low error rate and is stable against the decision threshold.

As the threshold value increases, precision increases, while a limited decrease in the F1 score is observed. This is due to the model making more conservative predictions at higher thresholds. When all results are evaluated, the best performance was achieved under the condition of $L=256$ and $T=0.50$, and this structure was selected as the final model for the study.

The hyperparameter ranges used in this study were determined based on the ranges reported in previous multi-label classification studies and preliminary experimental observations. Experiments with wider or different ranges resulted in limited performance gains, while training time was significantly extended. Therefore, the Grid Search process was limited to maintain a balance between computational efficiency and generalization. In future work, the model optimization process will be improved by using Bayesian Optimization that can scan larger hyperparameter spaces.

Combining the ConvNeXt and DistilBERT models with XGBoost was chosen to increase genre classification accuracy while preserving the individual representations of each modality. This approach specifically enabled the complementary nature of textual and visual representations. However, because the XGBoost model treats features as independent inputs, interactions between the two modalities are likely to be learned to a limited extent.

The results show that the hybrid model reduces the Hamming Loss value from 0.0971 to 0.0694 by combining visual and textual signals, demonstrating that the contribution of visual information is limited but significant. However, to increase the interpretability of the model, future studies plan to implement attention-based early or deep fusion approaches. This will enable learning of interactions between modals and strengthen the explainability of predictions.

In general, these findings display that the proposed hybrid model outperforms previous multimodal approaches and provides a solid foundation for more robust genre classification systems in future research.

5 Conclusion

In this study, a novel hybrid model is proposed for multi-label movie genre classification, which combines ConvNeXt and DistilBERT networks two architectures that have not previously been used together in this context thus leveraging the complementary strengths of both textual and visual features. Additionally, two separate models based solely on visual and textual features were also proposed and evaluated for the extraction of these respective features.

For image-based classification, a pre-trained ConvNeXt model was selected to classify movie posters into 20 predefined genres. The varying number of movies per genre and the fact that a single movie can belong to multiple genres are among the challenges of this study. To mitigate the effect of the imbalanced dataset, binary cross-entropy loss was employed. During training, stratified 5-fold cross-validation was applied to identify the best hyperparameters. To enable genre prediction with ConvNeXt, the final layers were removed and replaced with fully connected layers consisting of 1024, 512, 256, 128, and 20 units.

For text-based classification, the Transformer-based DistilBERT network which enabled major advancements in natural language processing was adapted for the task of genre prediction using movie summaries. The dataset used in the image-based model was expanded by adding movie summaries, and this enriched dataset was also utilized for training and testing the text-based model. The Multi Label Stratified Shuffle Split (MLSS) method was employed to ensure a balanced and adequate distribution of each genre across the training, validation, and test sets. During model training, token lengths of 128 and 256 and threshold values of 0.3, 0.35, 0.4, 0.45, and 0.5 were assigned separately as hyperparameters. Based on all evaluations, the highest performance was achieved when the token length was set to 256 and the threshold to 0.5. Therefore, in the hybrid model, these values were chosen for extracting textual features.

In the final stage of the study, the visual and textual features extracted from the image-based and text-based models were combined and provided to the XGBoost algorithm for genre prediction. The outputs obtained from both models were converted to binary values (0 or 1) for each genre by applying a threshold. The dataset was split into training and test sets using the iterative stratification method. Grid Search was employed for hyperparameter optimization. After running Grid Search, the optimal hyperparameters were found to be a tree depth of 3, a learning rate of 0.1, and 50 trees. Hamming loss was chosen as the primary metric for Grid Search to optimize.

Hamming loss, F1 score, and precision metrics which are suitable for multi-label classification tasks were evaluated for all three models.

When evaluating the Hamming loss metric, it was observed that the hybrid model achieved the lowest error rate with the value of 0.0694. The text-based model ranked second with a Hamming loss of 0.0753, while the image-based model had the highest error rate with a Hamming loss of 0.0971. These results demonstrate that combining textual and visual features improves multi-label film genre classification performance, and that textual features alone are more effective than visual features.

Considering the precision metric, the hybrid model achieved the highest precision with a value of 0.7319. The text-based model ranked second with a precision of 0.6491, while the image-based model had the lowest precision at 0.3310. These findings indicate that the combined use of textual and visual features significantly increases precision in film genre prediction, and that the text-based model performs much better than the image-based model.

In terms of the F1 score, the text-based model demonstrated the highest performance with a value of 0.6213. The F1 score of the hybrid model ,0.6200, was very close to and competitive with the text-based model. In contrast, the image-based model showed significantly lower performance with an F1 score of 0.4028 compared to the other two models. These results indicate that textual information is more effective for film genre prediction, and that combining it with visual features can also yield high performance.

The aim of this study was to demonstrate whether combining data from different modalities has a positive impact on performance. The experimental results showed that, overall, the text-based model outperformed the image-based model, while the hybrid model achieved even higher performance than both. Thus, the objective of the study was successfully achieved.

The proposed model also has several practical and managerial benefits. Specifically, in the film production and distribution industry, the proposed system can reduce content management costs by accelerating automatic genre tagging processes. Furthermore, in recommendation systems developed for digital broadcasting platforms, more accurate classification of movie genres can improve user experience and recommendation accuracy. Moreover, it has the potential to increase data consistency and system efficiency by reducing the reliance on human labor in updating large scale movie databases.

Although the hybrid model based on ConvNeXt and DistilBERT proposed in our study provides successful results in multi-label movie genre classification, it has some limitations. First, the model was tested on an IMDB-based dataset containing only visual (movie poster) and textual (movie synopsis) modalities. Therefore, future studies require validation of the model with additional modalities, such as trailers, audio, or subtitles from different sources. Furthermore, the class imbalances and multi-label nature of the dataset used led to some performance fluctuations. The model's complexity resulted in high computational costs, particularly when training the ConvNeXt and DistilBERT components together. However, these limitations provide a basis for future research to improve the model with larger datasets, additional modalities, and lower-cost optimization techniques.

The findings of our study demonstrate that multimodal hybrid approaches offer significant contributions to multi-label movie genre classification. However, future work remains to address current limitations and further improve performance. Future work plans to extend the proposed hybrid model to include different modalities (e.g., trailers, audio, or subtitles). Furthermore, testing the model not only on the IMDB dataset but also on larger and more diverse datasets will contribute to assessing its generalization capacity. In the fusion phase, examining transformer-based fusion approaches or attention mechanisms, in addition to XGBoost, are among the potential future research directions. It is also planned to implement dynamic (label-wise adaptive) thresholding approaches for different genre labels, enabling the model to perform more accurately, especially in unbalanced label distributions. Moreover, integrating the model into real-time recommendation systems and evaluating its genre prediction performance with user interaction also holds significant potential for future work.

Acknowledgements No acknowledgments are applicable for this study.

Author contributions Fatıma Zehra Ünal, Mehmet Serdar Güzel; Methodology: Fatıma Zehra Ünal, Mehmet Serdar Güzel, Fatih Ekinci, Koray Açıcı; Formal analysis and investigation: Fatıma Zehra Ünal, Metehan Ünal, Tunç Aşuroğlu; Writing—original draft preparation: Fatıma Zehra Ünal, Fatih Ekinci, Koray Açıcı; Writing—review and editing: Fatıma Zehra Ünal, Fatih Ekinci, Koray Açıcı, Tunç Aşuroğlu; Supervision: Mehmet Serdar Güzel, Fatih Ekinci, Koray Açıcı, Tunç Aşuroğlu

Funding Open access funding provided by the Scientific and Technological Research Council of Türkiye (TÜBİTAK). This research received no external funding.

Data availability The datasets analysed during the current study are available in the Kaggle repository, https://www.kaggle.com/datasets/neha1703/movie-genre-from-its-poster.

Declarations

Competing interests The authors state no Conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Paulino MAD, Costa YMG, Feltrim VD (2022) Evaluating multimodal strategies for multi-label movie genre classification. In: Proceedings of the 29th International Conference on Systems, Signals and Image Processing (IWSSIP 2022), pp. 1–6. IEEE. https://doi.org/10.1109/IWSSIP55020.2022.9854451
2. Cai Z, Ding H, Wu J, Xi Y, Wu X, Cui X (2024) Multi-label movie genre classification based on multimodal fusion. Multimed Tools Appl 83(19):36823–36840. https://doi.org/10.1007/s11042-023-16121-2

3. Khant P, Tidke B (2023) Multimodal approach to recommend movie genres based on multi datasets. Indian J Sci Technol 16(30):2304–2310. https://doi.org/10.17485/IJST/v16i30.1238
4. 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: Proceedings of the 31st ACM International Conference on Multimedia (MM'23), pp. 2814–2823. ACM. https://doi.org/10.1145/3581783.3612085
5. Sulun S, Viana P, Davies MEP (2024) Movie trailer genre classification using multimodal pretrained features. Expert Syst Appl 258:125209. https://doi.org/10.1016/j.eswa.2024.125209
6. Soykök IT, Güvenir HA (2025) Multi-label multi-modal classification of movie scenes. Knowl Based Syst 318:113459. https://doi.org/10.1016/j.knosys.2025.113459
7. Braz Leodécio et al (2021) Image-text integration using a multimodal fusion network module for movie genre classification. In: 11th International Conference of Pattern Recognition Systems (ICPRS 2021). Vol. IET, 2021
8. KIELA, Douwe et al (2018) Efficient large-scale multi-modal classification. In: Proceedings of the AAAI conference on artificial intelligence
9. Rafael MANGOLIN (2022) A multimodal approach for multi-label movie genre classification. Multimedia Tools Appl 81(14):19071–19096
10. Jha A, Kishore et al (2022) Multimodal and Multilabel Genre Classification of Movie Trailers. In: International Conference on Sustainable Computing and Data Communication Systems (ICSCDS). IEEE, 2022
11. Cascante-Bonilla P et al Moviescope: Large-scale analysis of movies using multiple modalities. *arXiv preprint arXiv*:(1908). *03180* (2019)
12. Kumar S et al (2023) Movie genre classification using binary relevance, label powerset, and machine learning classifiers. Multimedia Tools Appl 82(1):945–968
13. Hasan M, Mehedi et al (2021) Multilabel Movie Genre Classification from Movie Subtitle: Parameter Optimized Hybrid Classifier. In: 2021 4th International Symposium on Advanced Electrical and Communication Technologies (ISAECT). IEEE
14. Zhang Z et al (2022) Effectively leveraging Multi-modal Features for Movie Genre Classification. *arXiv preprint arXiv:2203.13281*
15. Nambiar G, Roy P, Singh D (2020) Multi Modal Genre Classification of Movies. In: 2020 IEEE International Conference for Innovation in Technology (INOCON). IEEE
16. Wehrmann J, Cerri R, Barros RC (2018) Movie genre classification: a multi-label approach based on convolutions through time. Neurocomputing 322:123–132. https://doi.org/10.1016/j.neucom.2018.09.042
17. Rajput NK, Grover BA (2022) A multi-label movie genre classification scheme based on the movie's subtitles. Multimedia Tools Appl 81(29):32469–32490. https://doi.org/10.1007/s11042-022-12961-6
18. Shaukat F, Ejaz N, Ashraf Z, Alnfiai MM, Alotaibi NN, Alnefaie SMM (2025) An interpretable multi-transformer ensemble for text-based movie genre classification. PeerJ Comput Sci 11:e2945. https://doi.org/10.7717/peerj-cs.2945
19. Behrouzi T, Toosi R, Mohammad Ali Akhaee (2023) Multimodal movie genre classification using recurrent neural network. Multimed Tools Appl 82(4):5763–5784
20. Sirirattanajakarin S, Thusaranon P (2019) Movie genre in multi-label classification using semantic extraction from only movie poster. In: Proceedings of the International Conference on Computer and Communication Management (ICCCM 2019), pp. 1–6. ACM. https://doi.org/10.1145/3348445.3348475
21. Visutsak P, Pensiri F, Netisopakul P (2024) Genre classification of movie trailers using spectrogram analysis and machine learning. In: Proceedings of the IEEE Black Sea Conference on Communications and Networking (BlackSeaCom 2024), pp. 1–6. IEEE. https://doi.org/10.1109/BLACKSEACOM61746.2024.10646256
22. RASHEED, Assad et al (2023) Cover-based multiple book genre recognition using an improved multimodal network. Int J Doc Anal Recognit (IJDAR) 26(1):65–88
23. Singhal S, Singh V (2025) Cross-domain learning framework for book—movie recommendation with RoBERTa and distilbert in action. Expert Syst (Preprint). https://doi.org/10.22541/au.174351635.54917844
24. Jaisankar V, Jayagopi DB (2024) Spectrogrand: Computational creativity driven audiovisuals' generation from text prompts. In: Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP 2024), pp. 1–9. ACM. https://doi.org/10.1145/3702250.3702280
25. Fallah H, Bellot P, Bruno E, Murisasco E (2022) Adapting transformers for multi-label text classification. In: Proceedings of CIRCLE 2022 – Joint Conference of the Information Retrieval Communities in Europe, pp. 1–12. HAL. https://hal.science/hal-03727927v1
26. Oramas S et al (2017) Multi-label music genre classification from audio, text, and images using deep features. *arXiv preprint arXiv:1707.04916*
27. Akalp H, Çiğdem EF, Yılmaz Ş, Bölücü N, Can B (2021) Language representation models for music genre classification using lyrics. In: Proceedings of the International Symposium on Electrical, Electronics and Information Engineering (ISEEIE 2021). ACM. https://doi.org/10.1145/3459104.3459171
28. Pizarro S, Zimmermann M, Offermann MS, Reither F (2024) Exploring genre and success classification through song lyrics using DistilBERT: A fun NLP venture. *arXiv preprint*, arXiv:2407.21068. https://arxiv.org/abs/2407.21068

29. Liu Z, Mao H, Wu C-Y, Feichtenhofer C, Darrell T, ve Xie S (2022) A convnet for the 2020s, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 11976–11986
30. Dataset (2024) Home page, https://www.kaggle.com/datasets/neha1703/movie-genre-from-its-poster, last accessed: 2024/01/01.
31. Sechidis K, Tsoumakas G, ve Vlahavas I (2011) On the stratification of multi-label data, Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2011, Athens, Greece, September 5–9, 2011, Proceedings, Part III 22, Springer, 145–158
32. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, pp 6000–6010
33. Devlin J, Chang M-W, Lee K, ve 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, volume 1 (long and short papers), 4171–4186
34. Sanh V (2019) DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108
35. Branescu I, Grigorescu O, ve Dascalu M (2024) Automated mapping of common vulnerabilities and exposures to mitre att&ck tactics. Information 15(4):214
36. Chen T, ve Guestrin C (2016) Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 785–794
37. Hastie T, Tibshirani R, ve Friedman J (2009) The elements of statistical learning. In: Citeseer
38. IMDb-Movie-Summaries-Dataset (2025) Home page, https://github.com/metehanunal/IMDb-Movie-Summaries-Dataset/, last accessed: 2025/11/07.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Fatıma Zehra Ünal[1] · Mehmet Serdar Güzel[2,3] · Metehan Ünal[2,4] · Fatih Ekinci[3] · Tunç Aşuroğlu[5,6] · Koray Açıcı[3,7]

✉ Koray Açıcı
kacici@ankara.edu.tr

[1] Optoelectronics and Electro-Optics Technologies Pr., Akyurt Vocational School, Ankara University, Ankara, Turkey

[2] Department of Computer Engineering, Ankara University, Ankara, Turkey

[3] Artificial Intelligence Institute, Ankara University, Ankara, Turkey

[4] Department of Computer Science and Mathematics, Keele University, Staffordshire, UK

[5] Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland

[6] VTT Technical Research Centre of Finland, Tampere, Finland

[7] Department of Artificial Intelligence and Data Engineering, Ankara University, Ankara, Turkey