International Journal on Science and Technology (IJSAT)

E-ISSN: 2229-7677 ● Website: www.ijsat.org ● Email: editor@ijsat.org

Hybrid Machine Learning Based Recommendation System for OTT Platform

Harsh Porwal[1], Mrs. Kiran Patel[2]

[1]Computer Engineering Department, Gujarat Technological University
Chandkheda, Ahmedabad, [2]Information Engineering Department, KITRC, Kalol
[1]porwalharsh007@gmail.com, [2]kiranmpatel178@gmail.com

Abstract:
This paper is dedicated to the creation of a recommendation system on Over-the-Top (OTT) platform based on the analysis of user behavior and multiple filtering strategies. The system incorporates collaborative filtering, content-based filtering and demographic methods to come up with personalized movie recommendations. The handling of missing values and the preprocessing of raw user data and the use of statistical techniques, which include the use of cosine similarity and TF-IDF vectorization, are used to identify the meaningful patterns in the viewing preferences. There are metadata attributes connected with the recommendations, including casting, directors, keywords, genres, etc. that enhance the recommendations and increase the prediction quality.
The study points to the fact that hybrid recommendation systems can address the weaknesses of classic ones, including sparsity, cold-start issues, and extreme computation fees. The outcomes of experimental evidence prove the hypothesis that metadata and behavioral analysis are better than individual methods of producing more relevant and varied recommendations. In addition to technical input, this publication focuses on the commercial worth of recommender systems and demonstrates the way in which they can help to boost customer satisfaction, to improve user engagement, and assist data-based marketing efforts in a highly competitive OTT landscape.

Keywords: Recommendation system, OTT platforms, collaborative filtering, content-based filtering, hybrid model, data mining.

I. INTRODUCTION

Recommendation systems have become a large component of the digital economy and are helping organizations to customize user experience and boost engagement. Based on user information like browsing history, rating and demographics, these systems can forecast and suggest content that is in line with personal preferences. They are important in various platforms such as the e-commerce and social media by guiding the users on large volumes of information.

Recommendation systems are specifically relevant in OTT platforms like Netflix, Amazon Prime, and Spotify because of the number of content available. They decrease the fatigue in decision-making by recommending relevant and new information, thus enhancing user satisfaction, boosting engagement, and leading to customer retention.

The common methods, such as the collaborative and content-based filtering, are not free of issues as the cold start problem, data sparsity, and scalability are significant problems with them. To overcome such problems, hybrid recommendation systems that unify various techniques and use both user behavior and metadata of content objects are being embraced.
Recommendation systems can also guide OTTs to improve their marketing strategies, purchase of content and revenues besides contributing to better user experience. Therefore, the creation of a

recommendation system of the next level is not only a technical need but also a strategic objective in the competitive industry of OTT.

A. Scope

The study is aimed at creating and pilot testing a recommendation system that is specific to OTT platforms. It makes use of user rating, review, and metadata datasets of movies. The system structures diversity of approaches; demographic filtering, content-based filtering, collaborative filtering and integrates them together. It only covers movies and other metadatas without expanding to other OTT content like live shows, sports or short-form videos.

The research focuses on technical application, such as the data pre-processing, similarity measurement, and algorithmic modeling. It is not about business-specific operational issues, including licensing or content acquisition. The results are, however, directly applicable to the improvement of the user experience and business strategy in OTT services.

B. Research Problem

Recommendation systems have a number of issues, particularly regarding OTTs, even though they are used extensively. Data sparsity is one of the primary problems as user ratings are scarce, and, therefore, it is hard to assume what people like and dislike. This is tightly connected with the cold-start issue, in which new users or objects do not have adequate data to be used in giving meaningful recommendations. Also, the conventional techniques such as collaborative and content based filtering are not scalable and computationally heavy with massive data sets resulting in real time delays in the recommendations. Even the evaluation metrics like MAE can not be comprehensive enough to measure performance in cases where performance patterns are not fixed.

The other weakness is that it is too specialised and systems are prone to offering users similar content many times and the user ought to have more variety and discovery. The low use of rich metadata diminishes further on the recommendation quality. These difficulties emphasize the necessity of hybrid methods which would incorporate several techniques to enhance the accuracy, efficiency and diversity.

C. Research Question

- What might preprocess and analysis of the user data contribute to the quality of recommendations to OTT platforms?
- Which are the relative advantages and disadvantages of content-based, collaborative, and hybrid filtering approaches to the OTT domain?
- What can metadata (cast, crew, genre) add to recommendation models to enable it to overcome the difficulties associated with sparsity and cold-start problems?
- How much does a hybrid recommendation system predict better and more satisfyingly than a single-method model?

D. Objectives

- **To perform data preprocessing and cleaning** on Netflix Movies and TV Shows datasets, including handling missing values, normalizing textual data, and transforming raw data into a structured format suitable for machine learning models.
- **To analyze user behavior and content metadata** in order to identify patterns in viewing preferences, using attributes such as genres, cast, directors, and descriptions through Exploratory Data Analysis (EDA).
- **To implement content-based filtering techniques** by applying methods such as TF-IDF vectorization to convert textual metadata into numerical representations, enabling similarity-based recommendations.

- **To apply collaborative filtering approaches** by simulating user-item interactions and utilizing matrix factorization techniques, specifically Truncated Singular Value Decomposition (SVD), to uncover latent relationships.
- **To develop a hybrid recommendation model** that combines content-based and collaborative filtering methods, addressing challenges such as cold-start problems, data sparsity, and lack of diversity in recommendations.
- **To evaluate the performance of the proposed system** using appropriate metrics such as Precision@10, ensuring the relevance and effectiveness of the recommendations.
- **To demonstrate the applicability of the proposed system** in OTT platforms by improving personalization, scalability, and diversity, thereby enhancing user engagement and supporting data-driven decision-making.

II. LITERATURE REVIEW

A. Problem Statement

OTTs present thousands of shows and movies, which makes it hard to find something to watch that suits the user. Even though recommendation systems are meant to address such a problem, most of them fail because of the lack of user information, incomplete metadata, and scalability, which in most cases leads to irrelevant or repetitive suggestions and poor engagement of users.

Conventional approaches such as collaborative and content-based filtering work well in a controlled setting, but are practical in textbooks, cold-start, and high computation issues in real-life settings. These restrictions do not allow OTT platforms to offer really personal experiences.

Hence, scalable hybrid recommendation system integrating various methods, using rich metadata and can successfully counter these problems to yield more precise, varied and user-friendly recommendations is needed.

B. Research Gaps

1. Low Efficacy of Univariate Methods: The current systems predominantly depend on either content-based or collaborative filtering which possesses limitations of over-specialization and reliance on large user data. Strong hybrid models that can be effective to combine both approaches are lacking.
2. Cold-Start Problems and Data Sparsity: The small amount of user interaction data and the presence of information to the new users/items decrease the accuracy of the recommendation. Existing solutions are either computationally intensive or do not use the available metadata exhaustively.
3. Poor Evaluation Measures: The conventional performance indicators such as MAE and RMSE are not indicative of the real user satisfaction. The partial application of measures like Precision, Recall and diversity causes partial evaluation of performance.
4. Inadequate use of Rich Metadata: Some of the metadata (genres, cast, directors, etc.) is not fully utilized to limit the capacity of the system to create contextual and relevant recommendations.
5. Scale and Computer Problems: Most of them are efficient with small data sets, but they do not scale effectively in larger OTT platforms because of the high level of computation complexity.
6. Inadequacy in Diversity and Novelty: The accuracy is emphasized in most systems, which leads to frequent repeats of recommendations and decreased interaction of the user in the long run.
7. Lack of Domain-Specific Optimization: The current solutions are generic and not specific to OTT platforms, which feature peculiarities such as multimedia content, viewing patterns, and user interaction patterns.

C. Research Paper

The paper (Nirmani et al., 2025) conducted a review of crowdsourced software engineering systems of task recommendations which are divided into content-based, collaborative, hybrid, and context-based

methods. The article emphasized the need to combine skills, reputation and task difficulty, with little empirical support in the field and deprivation of behavioral aspects. It found out that existing systems are simplistic and need more adaptive and scalable systems.

(Bamne & Agrawal, 2025) conducted a survey of product recommendation systems, with a focus on AI-based methods (including deep learning, NLP, reinforcement learning, etc.). Mixed methods contributed significantly to accuracy and personalization, on sites such as Amazon and Netflix. Nevertheless, such problems as cold-start, overfitting, and unethical concerns are significant problems.
A retrieval-based recommendation model suggested by (Nguyen et al., 2024) is used to recommend the candidate items and rank them in order of relevance and diversity. The model was also more precise, high recall, and quicker to compute than the traditional approaches. Yet, it is strongly based on quality of data and is not scalable and interpretable in real life.

Scholarly recommendation systems have been explored (Zhang et al., 2023), with the prevalence of content-based filtering and an increase in the use of hybrid and graph-based approaches. The study also established gaps in the recommendations of datasets and funding opportunities. It has some limitations: small datasets, the absence of benchmarks, poor attention to ethics and personalization.
The article (Ajmal et al., 2023) conducted a review of data mining-based recommender systems on the basis of social network data, with a preference towards hybrid methods of collaborative and content-based filtering. The most common technique was identified to be collaborative filtering. But some of the weaknesses are absence of standardized evaluation procedures and reporting of experimental limitations is inadequate.

(Valencia-Arias et al., 2024) investigated the use of AI in e-commerce recommender systems, which revealed such techniques as CNNs, sentiment analysis, and optimization algorithms. Accuracy and scalability were reported to have been improved in the study. Nonetheless, such problems as transparency, bias, and the lack of diversity in the datasets are still there.
There is a review of recommender systems in OTT platforms (Vicente and Burnay, 2024), which also emphasizes the usefulness of hybrid and artificial intelligence-based methods in enhancing engagement and personification. Other problems that were noted by the study include filter bubbles, privacy concerns, and transparency. Majority of the studies were not practically validated and ethically conducted.

The paper (Shinde et al., 2024) examined recommender systems in the context of big data and IoT, focusing on machine learning, deep learning, and edge computing to provide real-time recommendations. The researchers found enhanced accuracy and low latency in the study. However, there are the high cost of computing, privacy concerns and uninterpretability.
A detailed literature review of the recommender systems was presented by (Raza et al., 2024), which indicated that deep learning is more accurate and hybrid models are more diverse and scalable. These challenges included bias and lack of transparency as well as high computational requirements, which were pointed out in the study. Real world application is not much.

The overcoming of cold-start, sparsity, and scalability problems were the focus of analysis in (Trabelsi et al., 2021) of hybrid recommendation systems. Other hybridization methods such as weighting and switching were brought up. Although performance has improved, research biases, overfitting, and redundancy are still an issue.

The set of concepts of the AI-based recommender systems reviewed in (Masciari et al., 2024) is centered around such ethical issues as bias, privacy, and transparency. Although AI enhances scalability and

accuracy, it also creates ethical dangers. The paper has highlighted the importance of explainable and fairness-conscious systems.

The article (Necula and Pavaloaia, 2023) studied an example of AI-based recommendation systems in e-commerce and reported the enhancement of both personalization and customer engagement rates with the help of machine learning and graph-based models. Nonetheless, there are problems of transparency, bias, and generalizability that restrict practice.

Arban et al. (2023) created an OTT movie recommendation model based on the cosine similarity and various combinations of features. The system scored highly on similarity and enhanced user contentment. Nevertheless, it did not have standard evaluation metrics, scalability and comparison with high-level models.
The video recommender systems were reviewed (Lubos et al., 2023), with the focus on the implementation of hybrid and deep learning methods to increase the accuracy and variety. Inadequate explainability, problems with fairness, and bias of data were the major challenges mentioned in the study. Ethical issues are not well investigated.

(Engstrom et al., 2024) examined the users behavior of adopting recommender system, the analysis revealed that trust and perceived usefulness influences adoption, particularly in the entertainment system. Nonetheless, this study was based on self-reported information and had no technical assessment. Algorithms should be combined with behavioral understanding.

The reviewed hybrid recommendation systems showed increased accuracy, scaling, and diversity (Chaudhari et al., 2024). The paper has focused on the use of a combination of techniques with the aim of overcoming historical constraints. Nevertheless, there is a cost of high computation and uninterpretability.

Social relationships and similarity scores were used to propose a friendship-based recommendation model, as (Khalique et al., 2022) suggested. The system augmented accuracy of recommendation through user interaction and trust factor. Practical applicability is however limited by scalability concerns and size of dataset.

Reviewing deep learning-based recommender systems, (Li et al., 2023) emphasized the fact that these systems are able to capture the complex patterns and increase the accuracy. Deep learning models that were hybrids performed better. There are, however, issues of high computational complexity and inexplicability.

The evolution of the Amazon recommender system was also discussed in (Smith & Linden, 2017), with the focus on item-based collaborative filtering, which is scalable and provides real-time recommendations. The system enhanced user interaction and sale substantially. Nevertheless, such problems as cold-start and over-personalization still exist.
In recommender systems, (Kumara & Jadona, 2023) reviewed the machine learning methods and suggested the hybrid and context-aware methods as the best techniques to improve performance. Some of the challenges that were found in the study include bias, scalability, and non-existent evaluation assessment frameworks. The issue of ethics is not well investigated.
The systematic review of recommender systems in various fields presented in (Roy and Dutta, 2022) revealed the prevalence of collaborative and content-based filtering. There was better performance with hybrid and deep learning models. Nevertheless, there are issues of non-personalization, scalability, and standard benchmarks.

III. METHODOLOGY

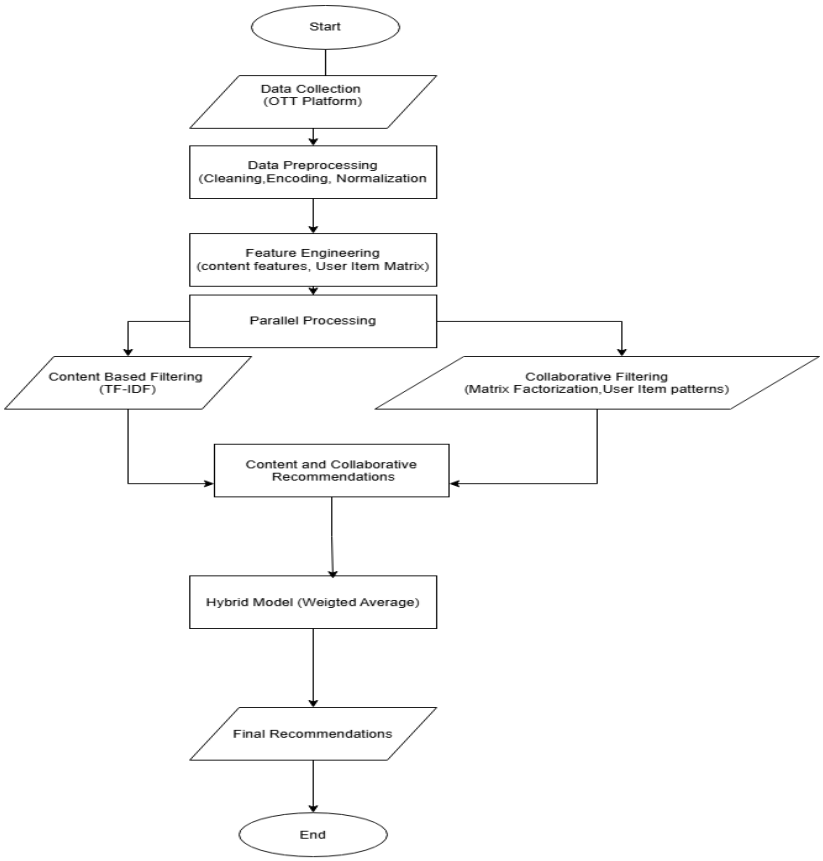

Fig.1 Proposed Flow of the Research

The proposed research will start by gathering the data on the user interaction, rating, reviews, and movie metadata on the OTT platforms or open sources, then clean and preprocess the data to guarantee the accuracy. The metadata is converted into machine-readable formats in order to solve such problems as cold-start or sparsity of data. The collaboration filtering and other methods like the clustering and matrix factorization are used to capture user behavior and increase its scalability. To increase the accuracy and diversity, a hybrid recommendation model that is built on collaborative, content-based, and demographic is then constructed. Lastly, the system is tested based on such metrics as accuracy, recall, and F1-score, which proves that the system is effective to enhance personalization and user engagement on OTT platforms.

A. Parameter Details

The Kaggle Netflix Movies and TV Shows dataset can be useful in creating a recommendation system because of the extensive metadata, such as ShowID and Type to differentiate content and Title, Director, and Cast to recommend something personal to the user. Other attributes like Country, Release Year and Date Added can be used to make region and time-based recommendations and Rating and Duration can be used to make preferences based on age and length. Recommendations are additionally refined by the Genre and Description fields by using content classification and text-based similarity algorithms such as TF-IDF to increase the overall accuracy and personalization.

IV. IMPLEMENTATION

A. Data Loading

The data employed in this project is the Netflix Movies and TV Shows data that has a total of 8,807 records. This metadata will contain all information on the content on the site including movies and television shows. The entries are considered to be unique titles and have many attributes including title, cast, director, genre, release year and a concise description on the contents.

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13	90 min	Documentaries	As her father nears the end of his life, filmm...
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries	After crossing paths at a party, a Cape Town t...
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...	To protect his family from a powerful drug lor...
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	September 24, 2021	2021	TV-MA	1 Season	Docuseries, Reality TV	Feuds, flirtations and toilet talk go down amo...
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, Romantic TV Shows, TV ...	In a city of coaching centers known to train I...

Fig 2: Sample entries from the Netflix Movies and TV Shows dataset

The data set of this project contains 8,807 records and 12 attributes, and each record is a title assigned a ShowID which is a unique identifier and contains both a categorical and a textual data. Characteristics of the content that are captured by key elements include type, title, director, cast, country and release year whereas features like rating and duration assist in the appropriateness of the content to the audience and the length of content. The description and genre provide the opportunity to perform the analysis at the advanced level with the help of the NLP techniques, so the dataset can be used in the hybrid recommendation systems. But the missing values in such fields as director, cast, and country point to the fact that it is necessary to preprocess data properly to guarantee quality and model behavior.

B. Data Cleaning and Preprocessing

Preprocessing and cleaning of data are important processes of ready the data to be used in machine learning operations. These procedures guarantee that the information is accurate, comprehensive, and error free, which could produce adverse consequences on the model operation. Missing or incoherent values are the norm in real-world databases, and unless addressed appropriately, may be the cause of biased outcomes, wasteful feature extraction, or even exception-based errors when running the model.

	title	release_year	genres_clean
0	Dick Johnson Is Dead	2020	documentaries
1	Blood & Water	2021	international tv shows, tv dramas, tv mysteries
2	Ganglands	2021	crime tv shows, international tv shows, tv act...
3	Jailbirds New Orleans	2021	docuseries, reality tv
4	Kota Factory	2021	international tv shows, romantic tv shows, tv ...
...	...	...	...
8802	Zodiac	2007	cult movies, dramas, thrillers
8803	Zombie Dumb	2018	kids' tv, korean tv shows, tv comedies
8804	Zombieland	2009	comedies, horror movies
8805	Zoom	2006	children family movies, comedies
8806	Zubaan	2015	dramas, international movies, music musicals

8807 rows × 3 columns

Fig 3: Data Cleaning and Preprocessing

Preprocessing of Netflix dataset was done to clean and prepare the data to be used in feature engineering by addressing the missing values in such fields as title, description, genre, director, and cast with blank values. Regular expression was used to normalize texts in order to make them lowercase, remove unnecessary spaces and unwanted characters to make them similar. To improve the recommendations, genre data was cleaned to be better tokenized and a new feature called credits which included both the

director and cast was created. The year of release was standardized, and the ultimate dataset of 8,807 records was optimized in such methods as TF-IDF to improve the quality of the data on the whole and the readiness of the model.

C. Exploratory Data Analysis (EDA)

Before recommendation techniques are applied, Exploratory Data Analysis (EDA) is very important in determining how data is structured, the patterns, and possible biases. This paper examined the Netflix data and identified the essential trends and direct feature engineering to a hybrid system. It is analyzed that the movies are the most dominant of the dataset, with only approximately 30.4% of the content being TV shows.

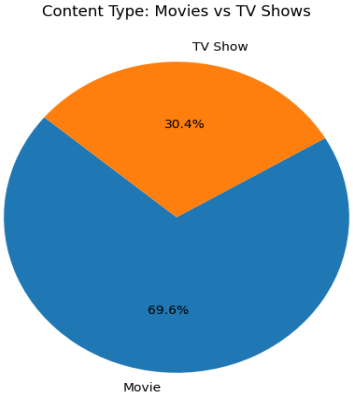

Fig 4 : Netflix Movies vs TV Shows Distribution.

This asymmetry has significant consequences in terms of recommendation systems since the models can be biased towards movies. Thus, it needs some extra plans that will guarantee that TV Shows are fairly represented in recommendations. The temporal analysis of content in terms of release year shows that there was a low content production until the 1980s, which then slowly accelerated to the 1990s and early 2000s, and then grew very fast after 2010, reaching its peak in 2017 to 2019.

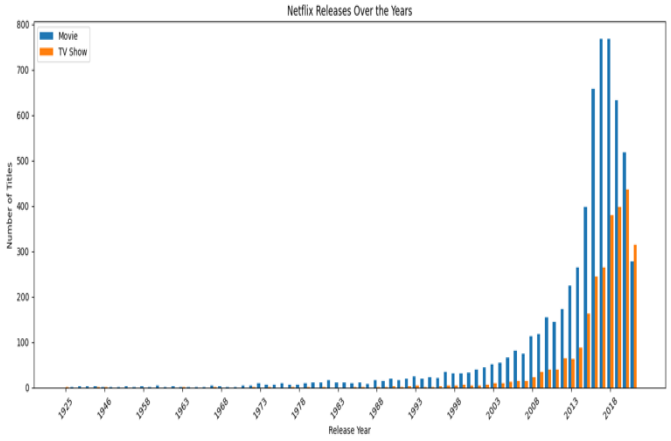

Fig 5: Netflix releases by the type of content over the years.

Movies are still dominant but over the last few years television programs are increasingly becoming popular, and there is a move towards series, a form of content that brings in recency bias that a model would have to deal with. Also, the content rating analysis demonstrates that the most widespread rating is TV-MA, then there is TV-14 and TV-PG, which means that the emphasis is made on adult viewers.

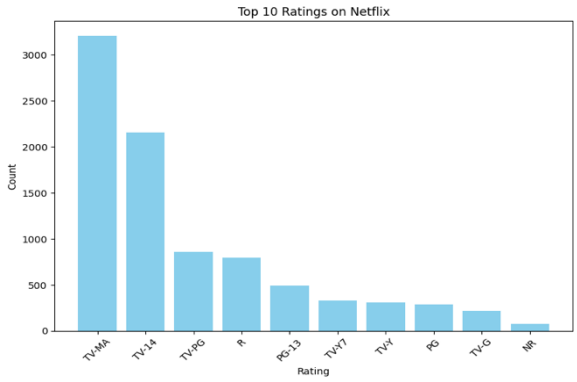

Fig 6: Top 10 content ratings on Netflix.

However, children-oriented ratings like TV-Y and TV-G are lower, indicating that rating features could be used to make sure that the right kind of ratings are applied.

The geographic distribution analysis demonstrates that the United States is the country to provide the most titles, with such nations as India, the United Kingdom, and Canada coming next.

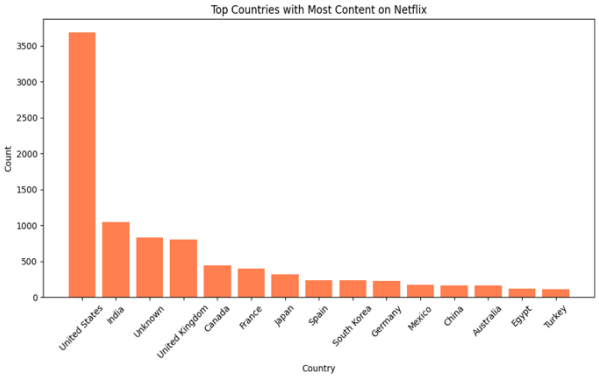

Fig 7: Top countries with most content on Netflix.

The existence of an “Unknown” category indicates incomplete metadata and the prevailing influence of U.S. content can easily cause regional bias unless some normalization methods are used.

Genre distribution analysis shows that the most common ones are International Movies and Dramas, then Comedies and International TV Shows are next.

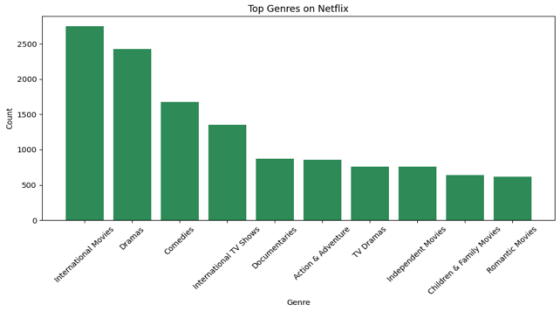

Fig 8: Top genres on Netflix

Although this is indicative of Netflix investing in a wide range of content and focusing on content from all over the world, the prevalence of specific genres could lead to biased recommendations, so genre-based normalization is critical to diversity.

The trends in temporal genres reveal strong variations in content strategy of Netflix with time. In the beginning, in the early 2000s, Comedies and Dramas were in the limelight, International Movies started

to gain significance around 2010, and by 2021, the International TV Shows, TV Dramas, and Docuseries gained relevance.

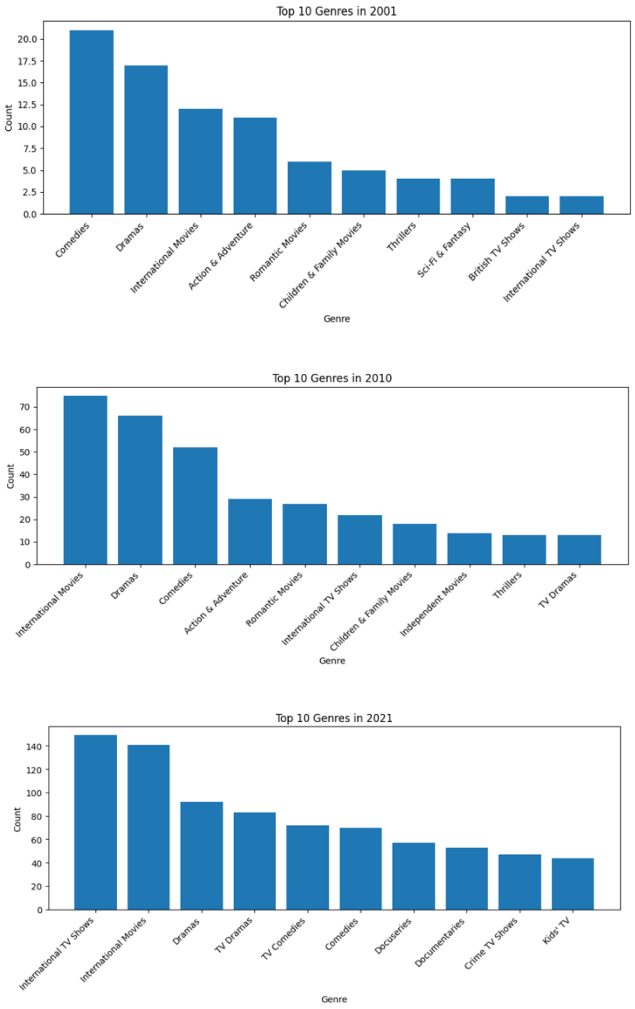

Fig 9: Temporal patterns in the most popular genres in various years (2001, 2010, and 2021).

Such changes underscore the necessity of using temporal dynamics when creating recommendation models so as to adapt to the variability of user preference.

Content analysis of textual descriptions shows that there is repetition of words like family, life, love, world and friend, which means that the emphasis is made on relationships and emotional narration.

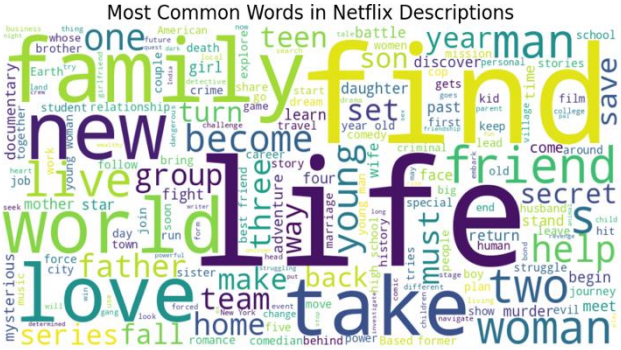

Fig 10: The most frequent words in Netflix.

The use of NLP techniques such as TF-IDF to filter content based on the use of NLP is justified by the presence of other similar words in the themes of the storytelling such as team, young, father, secret, and adventure.
Director analysis indicates that creators such as Rajiv Chilaka, Raul Campos and Jan Suter have made the largest contribution to the content, as well as internationally known directors like Steven Spielberg and Martin Scorsese.

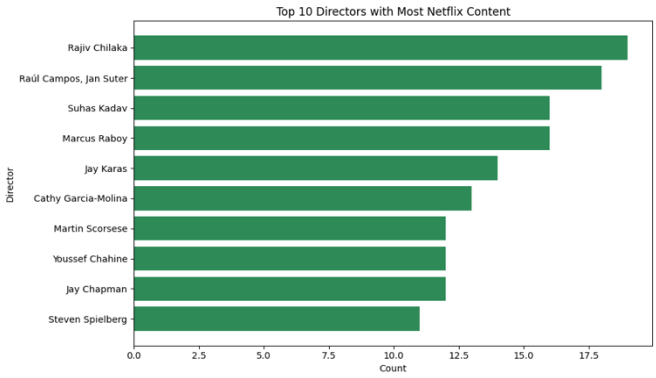

Fig 11: Ten most content directors on Netflix.

This means that there is a combination of international and regional creators and the inclusion of director metadata may increase personalization in recommendation.
On the whole, the EDA demonstrates that the dataset is abundant in the information, but it also has imbalances in the content type, region, genre, and ratings. The points above demonstrate the necessity of a hybrid recommendation strategy that integrates collaborative filtering recommendations with content-based strategies. With the help of metadata, which includes genres, descriptions, ratings, and directors, the system will reduce bias, enhance personalization, and have high quality and a variety of recommendations.

D. Feature Extraction and Engineering

Feature extraction and engineering is an important part in converting raw data into meaningful numerical form that can be used in machine learning models especially in systems of recommendations that depend on content-based and collaborative filtering models. In the given work, textual metadata was transformed into numerical values with the help of TF-IDF, giving words importance depending on their frequency and peculiarity between documents.

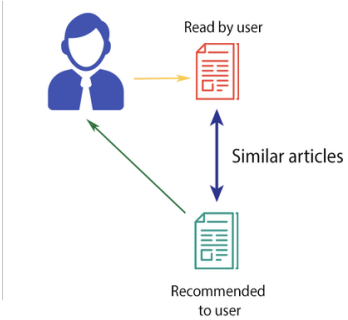

Fig 12: Content based Architecture.

TF-IDF allows focusing on the meaningful words and downplaying the common words, which makes it a computationally efficient and interpretable method of semantic information representation.
In order to obtain several content dimensions, different TF-IDF vectorizers were used on several key metadata fields including title, description, genres, and credits (as a single vector and combines director

and cast, including the two).

```
TITLE_FEAT  ← 1500
DESC_FEAT   ← 4000
GENRE_FEAT  ← 800
CREDIT_FEAT ← 800

tf_title  ← TfidfVectorizer(max_features=TITLE_FEAT, ngram_range=(1,2), stop_words='english')
tf_desc   ← TfidfVectorizer(max_features=DESC_FEAT,  ngram_range=(1,2), stop_words='english')
tf_genre  ← TfidfVectorizer(max_features=GENRE_FEAT, token_pattern=genre_token_pattern)
tf_credit ← TfidfVectorizer(max_features=CREDIT_FEAT, token_pattern=credit_token_pattern)

X_title  ← tf_title.fit_transform(df['title_clean'])
X_desc   ← tf_desc.fit_transform(df['desc_clean'])
X_genre  ← tf_genre.fit_transform(df['genres_clean'])
X_credit ← tf_credit.fit_transform(df['credits_clean'])

w_title, w_desc, w_genre, w_credit ← chosen weights (e.g., 0.3, 0.5, 0.15, 0.05)
X_content ← hstack([X_title * w_title, X_desc * w_desc, X_genre * w_genre, X_credit * w_credit])
X_content ← L2_normalize_rows(X_content)
```

Fig 13: Pseudo Code

Each field had feature limits established to provide a tradeoff between richness and efficiency, with those dimensions representing descriptions being larger than those representing structured fields such as genres and credits.

Once the features were vectorized, the weighted stacking of all the feature matrices was performed according to the importance of the feature and the L2 normalization was performed in order to assure constant scaling of the vectors and so that similarity could be effectively computed. The outcome of this process is a high-dimensional feature matrix of each title and a way of overcoming cold-start issues by using metadata.

Real user interaction data was not available, so there were synthetic interactions that were created to make realistic user interaction to use collaborative filtering.

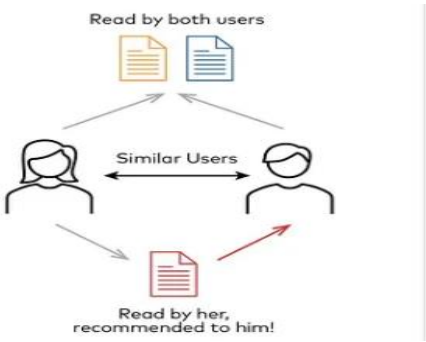

Fig 14: Architecture of Collaboration.

To bring more real-world consumption rates into the simulation, the biases included genre popularity and content recency.

The process of interaction generation entailed, probability scoring items depending on the frequency of the genre and the year of release and then sampling out interactions of synthetic users in a specific range.

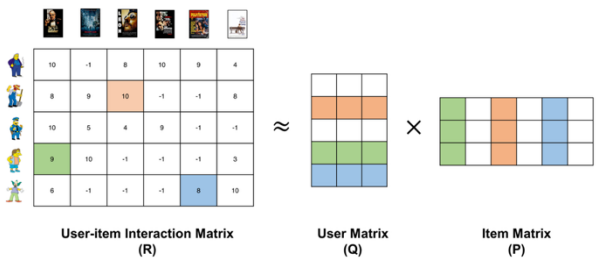

Fig 15: User Item Interaction Architecture.

This created a sparse user item interaction data that is very similar to real world systems.
Truncated SVD was used to reduce the dimensionality of the interaction matrix retaining significant relationships among the items in order to extract meaningful latent patterns.

```
n_items ← number of rows in df
N_USERS ← chosen number of synthetic users (e.g., 2000)

genre_popularity ← COUNTS of genres across df
item_genre_score ← for each item: sum of its genre popularities + 1
year_score ← (df.release_year - min_year) + 1
raw_score ← item_genre_score * year_score
probabilities ← raw_score / sum(raw_score)

rows = []
FOR user_id in 0 .. N_USERS-1:
    k ← random int between min_watch and max_watch (e.g., 8..35)
    sampled_items ← sample k unique items using probabilities (without replacement)
    FOR item in sampled_items:
        rows.append((user_id, item, 1.0))

interactions_sim ← DataFrame(rows, columns=['userId','item_idx','watch'])

user_item ← COO_to_CSR(matrix with shape (N_USERS, n_items) using interactions_sim)
item_user ← transpose(user_item)

k_factors ← chosen latent dimension (e.g., 64)
svd_model ← TruncatedSVD(n_components=k_factors).fit(item_user)
item_factors ← svd_model.transform(item_user)
item_factors_norm ← L2_normalize_rows(item_factors)
```

Fig 16: Truncated SVD Simulated Collaborative Filtering.

The model, with 64 latent dimensions, had a tradeoff between computational efficiency and representational power, allowing the similarity computation to be done using the vector operations.
It produced in the simulation around 41540 interactions among 2000 users and the simulation gave us sparse matrices that represent real-life recommendation situations.

```
Simulated interactions rows: (41540, 3)
user_item shape: (2000, 8807) item_user shape: (8807, 2000)
```

Fig 17: Statistics Simulated User-Item Interaction Matrix.

These matrices certify the existence of sparsity and yet have adequate data to extract patterns that would be meaningful.
Altogether, the presented feature engineering scheme is effective in combining the representation of content by the use of TF-IDF with simulated collaborative filtering and dimensionality reduction with the help of SVD, thus creating a potent hybrid scheme that enhances scalability, captures semantic and behavioral patterns, and gives a solid background to accurate, diverse, and effective recommendation systems.

E. Hybrid Recommendation Framework

A hybrid recommendation system is designed to combine the strengths of content-based filtering (CBF) and collaborative filtering (CF) to overcome their individual limitations and provide more accurate, diverse, and scalable recommendations. Content-based methods utilize item metadata such as title, description, genres, and cast to compute similarity, making them effective in cold-start scenarios, while collaborative filtering captures user behavior patterns and identifies hidden relationships between items, though it struggles with sparse interaction data.

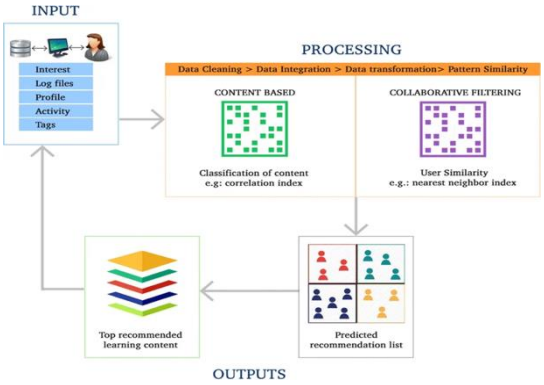

Fig 18: Hybrid Architecture

By integrating both approaches, the hybrid model ensures robust and balanced recommendations even when one component is weak.
The system first computes content-based similarity by generating TF-IDF embeddings for selected seed items and calculating their average representation, which is then normalized for cosine similarity comparison.

```
FUNCTION cf_scores_svd(seed_indices):
    user_cf_vec ← mean rows item_factors_norm[seed_indices]
    cf_scores ← item_factors_norm dot user_cf_vec
    RETURN flatten(cf_scores)

FUNCTION hybrid_scores(seed_indices, alpha, cf_method='svd'):
    content_scores ← content_similarity_for_seeds(seed_indices)
    IF cf_method == 'svd':
        cf_scores ← cf_scores_svd(seed_indices)
    ELSE:
        cf_scores ← item-item cooccurrence similarity (optional)
    stacked ← stack_columns(content_scores, cf_scores)
    scaled ← MinMaxScaler.fit_transform(stacked)
    hybrid ← alpha * scaled[:,cf_column] + (1-alpha) * scaled[:,content_column]
    RETURN hybrid

FUNCTION find_title_index(title_query):
    TRY exact match (case-insensitive) on df['title_clean']
    ELSE fuzzy match (difflib or rapidfuzz) with threshold
    IF no match: raise error / return None
    RETURN matched_index

FUNCTION recommend_by_title(title_query, topK, alpha, cf_method='svd'):
    idx ← find_title_index(title_query)
    hybrid ← hybrid_scores([idx], alpha, cf_method)
    hybrid[idx] ← -inf
    top_indices ← topK indices with largest hybrid scores
    RETURN df rows for top_indices with scores and metadata

FUNCTION recommend_for_user(liked_titles_list, topK, alpha, cf_method='svd'):
    liked_indices ← [find_title_index(t) for t in liked_titles_list]
    hybrid ← hybrid_scores(liked_indices, alpha, cf_method)
    set hybrid[liked_indices] ← -inf
    top_indices ← topK indices with largest hybrid scores
    RETURN df rows for top_indices with scores and metadata
```

Fig 19: Recommendation Generation using SVD-Based CF and Content Similarity

Similarity scores are computed using the dot product between the seed vector and all item vectors, ensuring recommendations can be generated even without user interaction data.
The suggested system employs a hybrid scoring system, which integrates the content based and collaborative filtering with the aid of a weighted parameter (a), which provides the flexibility of semantic similarity vs user behavior. It has a title-matching option based on precise and fuzzy means to make it more user-friendly, as well as can show one or more recommendations. Altogether, the given model is more accurate, diverse, and customized and solves cold-start problems, which makes it a scalable and feasible solution to OTT platforms.

F. Evaluation and Results

Precision at 10 was used to evaluate the hybrid recommendation system, as it is a metric of the relevance of the top 10 recommendations, and it can help to adjust the weighting parameter (a). The leave-one-out strategy was used, in which a single interaction per user was to be tested and a hit was implemented in case the interaction was observed in top results. Because the actual data on interaction were not available, synthetic user interactions were created with the help of weighted sampling, whereas TF-IDF and Truncated SVD were implemented in order to derive content-based and collaborative characteristics, which guarantee credible performance analysis.

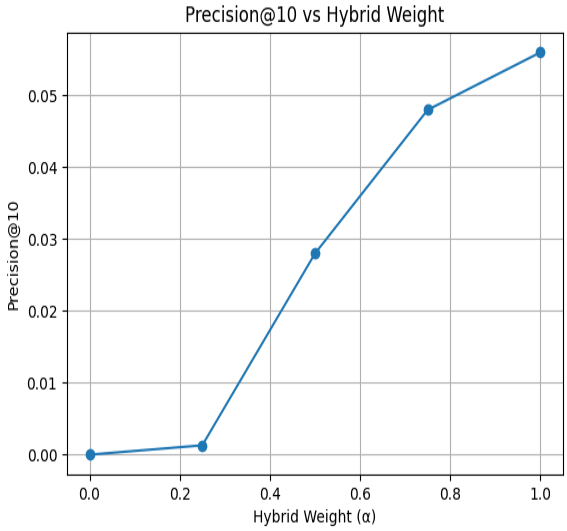

Fig 20: Analysis of Performance of Hybrid Model by Precision@10.

The hybrid model was tested in relation to the varying values of the weighting parameter (a), it can be seen that pure content-based filtering (a = 0) did not work well with almost zero Precision@10 and the more it added weight to the collaborative filtering, the better the performance was. Pure collaborative filtering (a = 1) resulted in the best result (Precision@10 = 0.0560), which means that interaction data are more predictive. Despite the fact that hybrid models were better than content-based ones, collaborative filtering was prevalent, with content features primarily serving diversity and cold-start management.

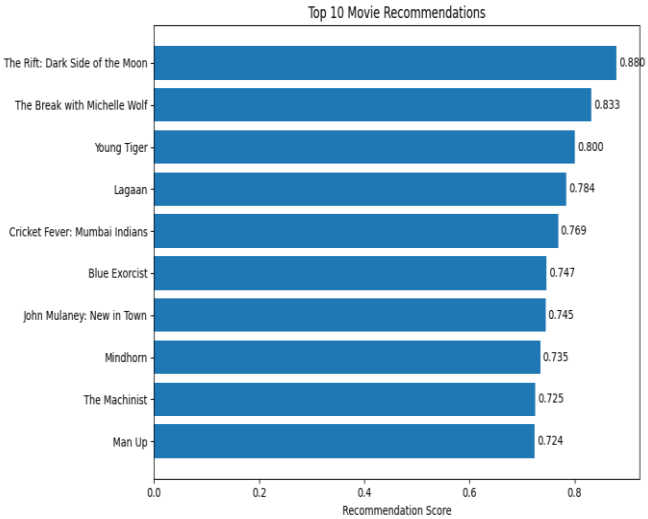

Fig 21: The 10 best recommended movies based on score.

In a qualitative assessment conducted with the help of the movie Sankofa, it was revealed that the recommendations were of various genres: drama, comedy, thriller, sports, and anime; some of them were

thematically similar, which confirms the usefulness of collaborative filtering. In general, the findings reveal that although collaborative filtering is more effective even using simulated data, the hybrid model is better at increasing diversity and moderate relevance and content discovery.

Table: Result Summary Table

Exper.	Method Used	α (Weight)	Precision@10	Key Observation
1	Content-Based Filtering	0.0	~0.0000	Very poor performance due to weak textual similarity
2	Hybrid Model	0.50	0.0280	Moderate improvement using combined features
3	Hybrid Model	0.75	0.0480	Better performance with higher collaborative influence
4	Collaborative Filtering	1.0	0.0560	Best performance using user interaction patterns
5	Overall Hybrid System	Variable	Up to 0.0560	Balances accuracy and diversity effectively

V. CONCLUSION

The present research was able to build a hybrid recommendation system in OTT sites by integrating content-based and collaborative filtering to address some of the major issues in these systems, including cold-start, data sparsity, and diversity. The model showed that collaborative filtering is more accurate, whereas the hybrid method is more robust and diverse in recommendations with the use of the Netflix dataset and simulated interactions between the user and the system. The effectiveness of the system was confirmed by experimental results on Precision at 10 closely followed by information obtained by the exploratory data analysis which informed the design of the model. Generally, the suggested structure presents a flexible and realistic approach to personalization and user interaction, and in the future, it is possible to consider real user data, sophisticated algorithms, and explainable AI to achieve a better performance and transparency.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to all individuals and organizations who contributed to the successful completion of this research work on the hybrid recommendation system for OTT platforms.

We extend our heartfelt appreciation to our mentors and academic guides for their continuous support, valuable insights, and constructive suggestions throughout the research process. Their guidance played a vital role in shaping the direction and quality of this work.

We also acknowledge the contributors of open-source libraries and tools such as **TensorFlow, Keras, NumPy, Pandas, and Scikit-learn**, which significantly facilitated the implementation of machine learning models and data processing tasks in this study.
Special thanks to the **Kaggle platform** for providing the Netflix Movies and TV Shows dataset, which served as the foundation for experimentation, analysis, and model development in this research.
We are equally grateful to our institution for providing the necessary infrastructure and resources required to carry out this work successfully.
Finally, we would like to thank our peers, friends, and family members for their constant encouragement, motivation, and support throughout the completion of this research paper.

REFERENCES:

[1] S. Nirmani, M. Shahin, H. Khalajzadeh and X. Liu, "A systematic literature review on task recommendation systems for crowdsourced software engineering," Information and Software Technology, vol. 184, pp. 1-27, 2025.
[2] P. Bamne and N. Agrawal, "A Survey of Product Recommendation System for Online Platforms," International Journal of Scientific Research & Engineering Trends, vol. 11, no. 3, pp. 1-8, 2025.
[3] D.-N. Nguyen, V.-H. Nguyen, T. Trinh, T. Ho and H.-S. Le, "A personalized product recommendation model in e-commerce based on retrieval strategy," Journal of Open Innovation: Technology, Market, and Complexity, vol. 10, pp. 1-14, 2024.
[4] Z. Zhang, B. G. Patra, A. Yaseen, J. Zhu and R. Sabharwal, "Scholarly recommendation systems: a literature survey," Knowledge and Information Systems, vol. 65, pp. 4433-4479, 2023.
[5] S. Ajmal, M. Awais, K. S. Khurshid, M. Shoaib and A. Abdelrahman, "Data mining-based recommendation system using social networks - an analytical study," PeerJ Computer Science, pp. 1-39, 2023.
[6] A. Valencia-Arias, H. Uribe-Bedoya, J. D. Gonz´alez-Ruiz and G. S. Santos, "Artificial intelligence and recommender systems in e-commerce. Trends and research agenda," Intelligent Systems with Applications, vol. 24, pp. 1-15, 2024.
[7] P. N. Vicente and C. D. Burnay, "Recommender Systems and Over-the-Top Services: A Systematic Review Study (2010–2022)," Journalism and Media, pp. 1259-1278, 2024.
[8] A. V. Shinde, A. V. Shinde and D. D. Patil, "A COMPREHENSIVE SURVEY ON RECOMMENDER SYSTEMS TECHNIQUES AND CHALLENGES IN BIG DATA ANALYTICS WITH IOT APPLICATIONS," RGSA – Revista de Gestão Social e Ambiental, vol. 10, pp. 1-30, 2024.
[9] S. Raza, M. Rahman, S. Kamawal, A. Toroghi, A. Raval and F. Navah, "A COMPREHENSIVE REVIEW OF RECOMMENDER SYSTEMS: TRANSITIONING FROM THEORY TO PRACTICE," arXiv, 2024.
[10] F. Z. Trabelsi, A. Khtira and B. E. Asri, "Hybrid Recommendation Systems: A State of Art," 2021.
[11] E. MASCIARI, A. UMAIR and M. H. ULLAH, "A Systematic Literature Review on AI-Based Recommendation Systems and Their Ethical Considerations," IEEE, pp. 23-42, 2024.
[12] S.-C. Necula and V.-D. Pavaloaia, "AI-Driven Recommendations: A Systematic Review of the State of the Art in E-Commerce," Applied Science, vol. 13, pp. 1-22, 2023.
[13] N. J. P. Arban, P. C. B. Arce, R. K. R. Bernabe, J. W. C. Kim and K. Y. C. Solomon, "An Exploratory Study of OTT Platform Movie Recommendation using Cosine Similarity," Research Congress, pp. 1-7, 2023.
[14] S. Lubos, A. Felfernig and M. Tautschnig, "An overview of video recommender systems: state-of-the-art and research issues," Frontiers in big data, pp. 1-22, 2023.

[15] E. Engstr¨om, I. Vartanova, J. V. Johansson, M. Persson and P. Strimling, "Comparing and modeling the use of online recommender systems," Computers in Human Behavior Reports, vol. 15, pp. 1-16, 2024.
[16] A. CHAUDHARI, A. A. H. SEDDIG, A. SARLAN and R. RAUT, "A Hybrid Recommendation System: A Review," IEEEAccess, pp. 7-27, 2024.
[17] A. Khalique, M. K. I. Rahmani, M. Saquib, I. Hussain, A. W. Muzaffar and M. A. Ahad, "A Deterministic Model for Determining Degree of Friendship Based on Mutual Likings and Recommendations on OTT Platforms," Computational Intelligence and Neuroscience-Hindawi, pp. 1-11, 2022.
[18] C. LI, I. ISHAK, H. IBRAHIM, M. ZOLKEPLI and F. SIDI, "Deep Learning-Based Recommendation System: Systematic Review and Classification," IEEEAccess, pp. 90-136, 2023.
[19] B. Smith and G. Linden, "Two Decades of Recommender Systems at Amazon.com," The Test of Time, pp. 1-7, 2017.
[20] S. Kumara and P. Jadona, "Machine Learning Techniques and Advancement in Recommendation System," Sharda University, Greater Noida, 2023.
[21] D. Roy and M. Dutta, "A systematic review and research perspective on recommender systems," Journal of Big Data, pp. 1-36, 2022.
[22] P. Ong, S.-Y. Chiu, I.-L. Tsai, Y.-C. Kuan, Y.-J. Wang and Y.-K. Chuang, "Nondestructive egg freshness assessment using hyperspectral imaging and deep learning with distance correlation wavelength selection," Current Research in Food Science, vol. 11, pp. 1-9, 2025.
[23] P. P. Patil and V. N.Patil, "NIR Spectroscopy for Rapid Freshness Assessment and Quality Classification of Chicken Eggs," Jordan Journal of Agricultural Sciences, vol. 21, pp. 1-16, 2025.
[24] O. I. Olufemi, O. Ayeni and O. E. Olagoke-Komolafe, "INNOVATIVE MODELS FOR MANAGING SALMONELLA AND OTHER MICROBIAL RISKS IN EGG PROCESSING AND STORAGE IN THE USA," International Research Journal of Modernization in Engineering Technology and Science, vol. 7, no. 1, pp. 1-11, 2025.
[25] Z. Gao, J. Zheng and G. Xu, "Research Progress and Technological Application Prospects of Comprehensive Evaluation Methods for Egg Freshness," MDPI, vol. 14, pp. 1-19, 2025.
[26] I. C. Crivei, R. A. Marc, R. N. Rațu and A. N. Postolache, "Advanced risk and hazard analysis in the egg sorting-packing units industry from supplier selection to delivery in chain stores under global food safety initiative integrated food safety programs," Heliyon, pp. 1-31, 2025.
[27] O. Olagunju, M. Stump and Y. Li, "Machine learning enabled nondestructive quality analysis of animal protein based foods: a comprehensive review," Agricultural Products Processing and Storage, vol. 1, no. 7, pp. 1-28, 2025.
[28] E. Sheidaee and P. Bazyar, "Enhancing the destructive egg quality assessment using the machine vision and feature extraction technique," Research in Agricultural Engineering,, vol. 71, no. 2, pp. 95-105, 2025.
[29] M. Denli, E. Yavuzer, H. Tangüler and M. Köse, "Determination of Quality Changes of Hard-Boiled Chicken Eggs Due to Slow and Fast Cooling by Electronic Nose and Machine Learning," Turkish Journal of Agriculture - Food Science and Technology, vol. 13, no. 4, pp. 934-940, 2025.
[30] E. M. Atwa, S. Xu, A. K. Rashwan, A. M. Abdelshafy and G. ElMasry, "Advances in Emerging Non-Destructive Technologies for Detecting Raw Egg Freshness: A comprehensive Review," MDPI, pp. 1-27, 2024.

[31] E. Sheidaee and P. Bazyar, "Enhancing the destructive egg quality assessment using the machine vision and feature extraction technique," Research in Agricultural Engineering, vol. 71, pp. 1-10, 2024.
[32] Z. Wu, H. Zhang and C. Fang, "Research on machine vision online monitoring system for egg production and quality in cage environment," Poultry Science, pp. 1-18, 2025.
[33] R. G. H. Nivasini, A. J. J. Priscilla, M. Logeswari and N. Damini, "Thermal Imaging for Egg Freshness Classification using Convolutional Neural Networks with SVM-Based Accuracy Prediction," International Journal of Multidisciplinary Research Transactions, pp. 1-12, 2024.
[34] M. W. Ahmed, A. Khaliduzzaman, J. L. Emmert and M. Kamruzzaman, "An overview of recent advancements in hyperspectral imaging in the egg and hatchery industry," Computers and Electronics in Agriculture, vol. 230, pp. 1-16, 2024.
[35] T. O. S. Akowuah, E. Teye, J. Hagan and K. Nyandey, "Rapid and Nondestructive Determination of Egg Freshness Category and Marked Date of Lay using Spectral Fingerprint," Journal of Spectroscopy, pp. 1-11, 2023.
[36] M. Ahmed, S. J. Hossainy, A. Khaliduzzaman, J. L. Emmert and M. Kamruzzaman, "Non-destructive optical sensing technologies for advancing the egg industry toward Industry 4.0: A review," COMPREH ENSIVE REVIEW- WILEY, pp. 4378-4404, 2023.
[37] V. M. Nakaguchi and T. Ahamed, "Fast and Non-Destructive Quail Egg Freshness Assessment Using a Thermal Camera and Deep Learning-Based Air Cell Detection Algorithms for the Revalidation of the Expiration Date of Eggs," Sensors, pp. 1-22, 2022.
[38] Ms.PritiP.Patila and Dr.V.N.Patil, "Freshness Detection and Classification of Chicken Eggs-A Review," NeuroQuantology|, vol. 20, no. 11, pp. 8907-8914, 2022.
[39] M. S. Yogeswari, J. Selamat, N. N. Jambari, A. Khatib, M. H. M. Amin and S. Murugesu, "Metabolomics for quality assessment of poultry meat and eggs," Food Quality and Safety-Oxford, vol. 8, pp. 1-14, 2024.
[40] X. Han, Y.-H. Liu, X. Zhang and Z. Zhang, "Study on egg sorting model based on visible-near infrared spectroscopy," Systems Science & Control Engineering, vol. 10, pp. 733-741, 2022.