Journal of New Media

Tech Science Press

DOI: 10.32604/jnm.2023.037583
Article

Accurate Machine Learning Predictions of Sci-Fi Film Performance

Amjed Al Fahoum[1,*] and Tahani A. Ghobon[2]

[1]Biomedical Systems and Informatics Engineering Department, Hijjawi Faculty for engineering Technology, Yarmouk University, Irbid, 21163, Jordan
[2]School of Engineering Technology, Al Hussein Technical University, Amman, Jordan
*Corresponding Author: Amjed Al Fahoum. Email: afahoum@yu.edu.jo
Received: 09 November 2022; Accepted: 15 May 2023; Published: 16 June 2023

Abstract: A groundbreaking method is introduced to leverage machine learning algorithms to revolutionize the prediction of success rates for science fiction films. In the captivating world of the film industry, extensive research and accurate forecasting are vital to anticipating a movie's triumph prior to its debut. Our study aims to harness the power of available data to estimate a film's early success rate. With the vast resources offered by the internet, we can access a plethora of movie-related information, including actors, directors, critic reviews, user reviews, ratings, writers, budgets, genres, Facebook likes, YouTube views for movie trailers, and Twitter followers. The first few weeks of a film's release are crucial in determining its fate, and online reviews and film evaluations profoundly impact its opening-week earnings. Hence, our research employs advanced supervised machine learning techniques to predict a film's triumph. The Internet Movie Database (IMDb) is a comprehensive data repository for nearly all movies. A robust predictive classification approach is developed by employing various machine learning algorithms, such as fine, medium, coarse, cosine, cubic, and weighted KNN. To determine the best model, the performance of each feature was evaluated based on composite metrics. Moreover, the significant influences of social media platforms were recognized including Twitter, Instagram, and Facebook on shaping individuals' opinions. A hybrid success rating prediction model is obtained by integrating the proposed prediction models with sentiment analysis from available platforms. The findings of this study demonstrate that the chosen algorithms offer more precise estimations, faster execution times, and higher accuracy rates when compared to previous research. By integrating the features of existing prediction models and social media sentiment analysis models, our proposed approach provides a remarkably accurate prediction of a movie's success. This breakthrough can help movie producers and marketers anticipate a film's triumph before its release, allowing them to tailor their promotional activities accordingly. Furthermore, the adopted research lays the foundation for developing even more accurate prediction models, considering the ever-increasing significance of social media platforms in shaping individuals' opinions. In conclusion, this study showcases the immense potential of machine learning algorithms in predicting the success rate of science fiction films, opening new avenues for the film industry.

Keywords: Film success rate prediction; optimized feature selection; robust machine learning; nearest neighbors'; algorithms

1 Introduction

Films, online videos, and television are the most prevalent forms of entertainment worldwide [1]. The film industry is a significant contributor to international commerce and advertising, as well as one of the most critical entertainment industries on a global scale. Films are potent tools that can alter people's perspectives and behaviors for the better. The film industry requires substantial financial, time, and physical resources [2]. The film industry produces hundreds of films annually. Consequently, it is essential to anticipate a film's success in its early stages. A film's success or failure depends on several factors. Both positive and negative influences from movies on adolescent behavior were found in a study by the authors [3]. According to [4], moviegoers who already smoke are further persuaded to pick up the habit after viewing a film containing a smoking character. Locations are other features that can attract film's viewers, such places were found so attractive that tourists will keep coming to them long after the movie has been released, no matter the season or the climate. As a result, cinema may soon become a popular inspiration for vacations [5]. Popular and significant genres include science fiction. Many samples of significant technical and artistic advancements in science fiction films exist. Awards of note are being considered for films in the science fiction genre. Table 1 demonstrates that all of the genres mentioned above and others not included can be present in a science fiction film. There is a cultural and creative impact from watching science fiction films. Middle school pupils who watched science fiction films had more technological imagination and product improvement ideas [6]. Another study shows that people's conceptions of what is achievable with technology have been greatly influenced by science fiction. People's interest in robotics and artificial intelligence (AI) will also be piqued [7]. Several examples of science fiction films have changed people's minds, such as The Day After Tomorrow, which depicts the fast evolution of the World's environment into an ice age. The film left audiences feeling more responsible for protecting the planet and its climate [8]. Aside from the film's aesthetic and cultural value, the director should think about how much money it could make when it is released. The success of a film can be assessed by its box office performance or its reception by critics. Several aspects determine a film's box office performance, including the film's genre, writer(s), director(s), actor(s), running time, year of release, production firm, and producers. The film's success is tied to how well it is marketed and advertised [9]. There are far too many variables at work to reliably forecast a film's success. Nevertheless, a model that can predict a film's performance at the box office can be proposed by studying the successes and failures of similar films. The purpose of this research is to use the existing data to make accurate early predictions about how successful a film will be at the box office. The Internet provides access to various movie-related data, including IMDB ratings, movie titles, actors, critic reviews, user reviews, directors, social media ratings, production information, writers, available budget, genre, YouTube opinions and numbers for movie trailers, Facebook likes, and Twitter followers. The initial few weeks of a film's box office run are crucial to its success during this period [10]. Internet reviews and film evaluations have a significant impact on opening-week earnings. Because the first few weeks of a film's release are crucial to its success, the film's production team spends much time on publicity and developing public opinion. The Internet Movie Database (IMDb) contains information about practically all films. The prediction of the success of a movie can be achieved utilizing supervised machine learning techniques. The best model for the feature is selected by comparing the accuracy of its prediction in terms of root mean squared error (RMSE), maximum

absolute error (MAE), minimum variance, and R2 score. Then the selected feature set is used via various machine learning techniques to achieve a robust prediction. Furthermore, social media sites like Facebook, Instagram, and Twitter have grown to be very influential in shaping people's beliefs. These sources, crucial for acquiring movie evaluations, generate large data. Reviews and remarks are the most common ways individuals express their opinions regarding a film. Therefore, it is possible to become inundated by this quantity of data. Therefore, it is natural to question whether it is possible for computers to deduce emotions from current data. Combining machine learning and natural language processing (NLP), a model is developed to analyze the language of online content such as messages and reviews. Positive reviews and remarks are separated from negative reviews and comments. Several machine learning algorithms were applied to the problem, and the results were compared to determine the most precise model for predicting the commercial success of a film. Combining the prediction model, as mentioned earlier, and the emotion analysis model yields a prediction of the hybrid success rating. In this paper, among other machine learning algorithms a special purpose one is developed and applied to predict the success of a movie considering many factors. Different types of machine learning algorithms are applied and all results are then discussed and compared with each other. By integrating the features of existing prediction models and social media emotion analysis models, the proposed work provides a more precise prediction of the movie's success than previously published work. The remainder of the paper is organized as follows: Section 2 surveys the relevant literature, and Section 3 describes the structure and components of the newly devised algorithm. The results will be presented in Section 4, along with a discussion of the lessons learned during the implementation process and their relationship to recently published works. The fifth section concludes the study and outlines directions for future research in this field.

Table 1: Science fiction movies genres

Movie	Genres
About time	Fantasy, comedy, romance, Sci-Fi, drama
Alien	Sci-Fi, horror
The martian	Sci-Fi, drama, adventure
Boss level	Thriller, Sci-Fi, mystery, action
Inception	Adventure, thriller, action, Sci-Fi
Okja	Action, Sci-Fi, adventure, drama
A clockwork orange	Crime, Sci-Fi, drama
The day	Horror, action, Sci-Fi, drama, thriller
Coherence	Thriller, Sci-Fi, drama, mystery, horror,
Mac and me	Fantasy, adventure, Sci-Fi, family
Only	Romance, drama, Sci-Fi
Iron sky	Action, Sci-Fi, adventure, comedy
Dune	Sci-Fi, action, adventure,
Megaforce	Sci-Fi, action
Maggie	Horror, drama, Sci-Fi

2 Literature Background

There are a number of academic articles that use machine learning algorithms to create estimates about a film's success, popularity, and box office take. Numerous studies on film rating prediction and review emotion analysis have been undertaken. Machine learning techniques were used to classify the obtained features whereas natural language processing is used in the majority of sentiment analysis. The following are research illustrations that attempt to predict whether movie reviews will be positive or negative. Using unsupervised learning, Turney [11] classified movies by their average semantic orientation. Support vector machine (SVM) classifiers performed the best in comparing machine learning algorithms used to analyze movie evaluations conducted by Pang et al. [12]. Mishne et al. [13] analyzed bloggers' opinions to forecast box office receipts. Machine learning was utilized in conjunction with a straightforward method incorporating the tally of reviewers' positive and negative word counts [14]. A new analysis attempts to predict the box office success of films [15]. It has been argued [16,17] that the selection of model variants and features influenced by the employer substantially affects the efficacy of the text classification algorithm that serves as the baseline for machine learning. In [18], it was suggested that employing sentiment analysis could improve collaborative filtering. In order to predict emotions, we combined generative and discriminative models [19]. Using a clustering method, they presented an enhanced sentiment analysis of online movie evaluations [20]. In [21,22], A combined CNN and a long short-term memory (LSTM) technique has been proposed as an alternative to straightway methods for sentiment analysis. The 'Senti ALSTM' model for sentiment analysis was presented in [23], and it is an attention-based model for the LSTM algorithm. The Bollywood film prediction model is presented in [24]. A lexicon and neural networks for sentiment analysis were proposed [25]. In [26], they proposed the random forest (RF) technique model as a highly accurate predictive model to evaluate the box office performance of a film. While in [27] they developed a method for rating prediction using decision trees [26]. LSTM-based sentiment analysis has been proposed for film evaluations [27]. In [28], a news recommendation system was presented using a machine learning (ML) classifier based on a multinomial Nave Bayes (MNB) classifier. The methodology proposed by [29] for predicting the box office performance of a film uses a combination of social network analysis, text mining, and machine learning. Using their model, numerous categories of attributes can be extracted. They examined the box office from the audience's perspectives, the film's release, and the film itself. Both IMDB and Box Office Mojo provide information for their investigation. In [30], they analyzed and associated the performance and significance of three models that use supervised learning methodologies to predict future revenues. Wikipedia, IMDB, and Rotten Tomatoes all contribute to their respective databases. In [31], they proposed methods for forecasting the success of movies by comparing six machine learning models using the same method as [32]: categorizing films based on their IMDB rating as horrible, poor, average, or outstanding. Researchers in [32] presented machine learning techniques employing intrinsic features, correlation coefficients, and other machine learning tools to forecast movie popularity categories, whereas [33] proposed a data mining strategy to evaluate and forecast movie ratings. By means of neural networks and regression data mining techniques, researchers in [34] developed a method for estimating profit and forecasting box office income using neural networks. Initially, Opus Data is used, then, information from IMDb, IDMB, Metacritic, and Mojo 2016a was added to the initial set. The authors of [35] construct a mathematical model to predict the popularity of films in both the Hollywood and Bollywood markets using the KNN technique. The data came from various online sources, including IMDB and social media platforms. In [36], the authors anticipated a film's revenues to assist with prior budgetary considerations. The group proposes a method that utilizes analysis of social networks and mining of texts to extract information autonomously about a film's cast, narrative, and date of release from

various data sites. Their findings demonstrated that the system was preferable to other methods. They utilized a feature selection technique that significantly improved the forecast's accuracy. They intend to develop a practical decision-making tool by analyzing the industry's dominant factors. They developed a mathematical model [37] that accurately predicted the box office performance of forthcoming films. The movie star, budget, producer, location, narrative writer, cast, trailer, director, release date, concurrently released films, music, announcement location, and target viewers can all impact a film's success. They constructed a model by analyzing the correlations between individual attributes. The prediction was made after giving each factor due consideration and designating it a weight. They also demonstrated how the method could select a group of profit-maximizing actors, demonstrating its prescriptive potential. Researchers in [38] presented a method for predicting a film's box office performance before its premiere, thereby eradicating the need to rely on the opinions of critics and others. Using the IMDb Movie Dataset, this paper explains how to estimate an IMDb rating. Multiple methodologies were utilized in their study, but the RF classifier developed the highest accuracy. Numerous factors significantly influenced an IMDb rating, including people who voted for the film, reviewers who rated it, the number of likes on social media received, the film's running time, and its total box office earnings. Typically, the finest films in a particular genre are dramas and biopics. Using regression techniques, the authors of [39] created a model that predicts the box office performance of future films based on a number of factors. In [40], algorithms were devised to predict a film's box office earnings before its release. The financial success or failure of the film was determined using heuristics to set a revenue threshold. They were eliminated since YouTube comments on trailers and previews are informative when evaluating a film. Keywords were extracted from user reviews using NLP, and the reviews' affective content was examined to determine if they were good or bad. An RF algorithm was trained using data mined from IMDb to forecast the triumph of a movie. Additionally, the Naive Bayers model was implemented to train the model to estimate the rating of a movie by analyzing YouTube user ratings. The models' accuracy was measured by how well they performed on real-world datasets. At this point, two things are certain: (1) a film's success can be forecast using internet data or characteristics, and (2) a new film's rating cannot be predicted using observations on its trailers and teasers gathered from YouTube. The models' overall efficacy is on par with that of other examples in the literature. With a 70% overall accuracy, they urged that their proposed model can be utilized as a preceding evaluation instrument to predict the movie performance, which would be good for both the film industry and moviegoers. The many machine learning models used to forecast a movie's success are compared and contrasted in [41]. They examined the predictor models' efficacy and statistical significance to pick the best one. Some of the aspects that contribute to a film's box office performance are also discussed. Multiple models were examined, including a neural network, a time series model, a regression model, and a machine learning model. The accuracy of the neural network was around 86%. The testing also includes a review of the 2020 movie release schedule. The authors of [42] took into account genre, release date, actors, and box office gross while deciding the fate of a film's cast. Numerous factors can affect the success or failure of a film's release, and business proprietors in the film industry must be aware of these risks. In light of this, they presented an ensemble learning approach to evaluating such comprehension, whereby predictions from earlier guided attribute computations can be used to enhance the success/failure accuracy of later evaluations. The results are compared and contrasted in the literature using a number of different approaches. Several machine learning algorithms, including SVM, KNN, Naive Bayes, Boosting Ensemble Method, Stacking Ensemble Technique, Voting Ensemble Technique, and MLP Neural Network, have been applied to this dataset to predict a film's box office success. Using many algorithms and trends, the authors of [42] could foresee the upshot of a film and demonstrate that their proposed approach is superior to the present standard of practice. Their research showed that

gradient boosting obtained the highest success rate (84.129%). The study's abstract suggests that machine learning algorithms could be useful for predicting the box office performance of a film by considering all relevant criteria. Therefore, there is an opportunity to advance previous work by creating an algorithm that improves the efficacy of tried and true methods. The objective is to propose an algorithm based on machine learning that outperforms existing methods and accurately predicts the probability of success given the available data (including prior knowledge, user behavior, and the algorithm's track record). The prescriptive value of our method will be demonstrated by recommending the agents with the highest potential revenue. This study discusses the potential for the use of efficient forecasting tools and adaptive data analytics to shape the future of the science fiction film business.

3 Methodology

The proposed model in this study will follow the steps below to improve the prediction of the movie's success rate.

- Data gathering and cleansing
- Selection of features and data modeling
- Application of several Machine Learning algorithms
- Compare the outcomes of various algorithms

3.1 Data Gathering and Cleansing

All of the data utilized in this paper was culled from IMDB. Information such as cast members, directors, authors, narrative synopses, producers, movie titles, genres, box office earnings, trailers, and more can be found on IMDB. This online database provides information about movies, television shows, home videos, video games, and streaming material. Movies on IMDB can be rated anywhere from 1 to 10. After extracting the data, we removed duplicates, junk, and movies that did not fit the sci-fi genre, TV shows, short films (less than 1 h), or films with less than 1000 ratings.

3.2 Selection of Features and Data Modeling

The results of the ratings have been divided into four categories for analysis and classification, as shown in Table 2. The method described in [43] is replicated here. Analyzing the data and investigating the connections between the various elements that influence the film's financial success is crucial. Important details can be gleaned from the subsequent two figures. We constructed several narratives from the data we gathered on science fiction films and then used that information to inform our genre analysis. The feature selection process is based on the performance of each feature which was evaluated based on composite metrics such as root mean squared error (RMSE), maximum absolute error (MAE), minimum variance, and R2 score. The critical feature histogram is depicted in Fig. 1. In Fig. 2, the IMDB score and highlights are shown.

Table 2: Types of ratings

Scores	Classification
$\leq$3.4	Flop
3.5–5.9	Below average
6.0–7.5	Average
$<$7.5	Hit

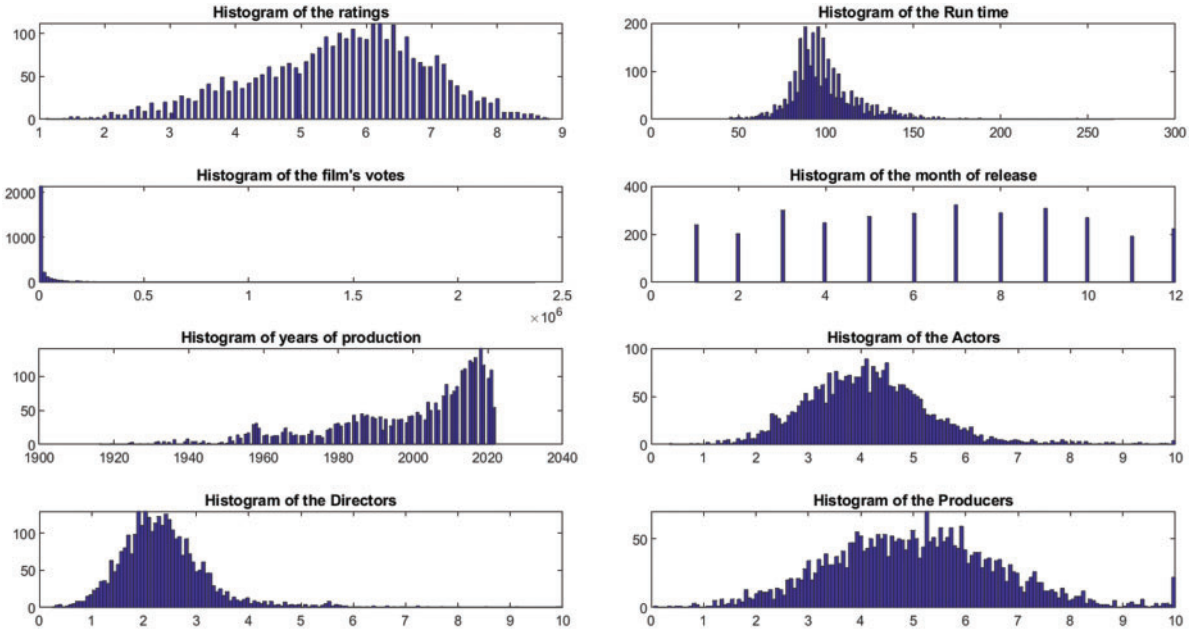

Figure 1: The histogram of the main features

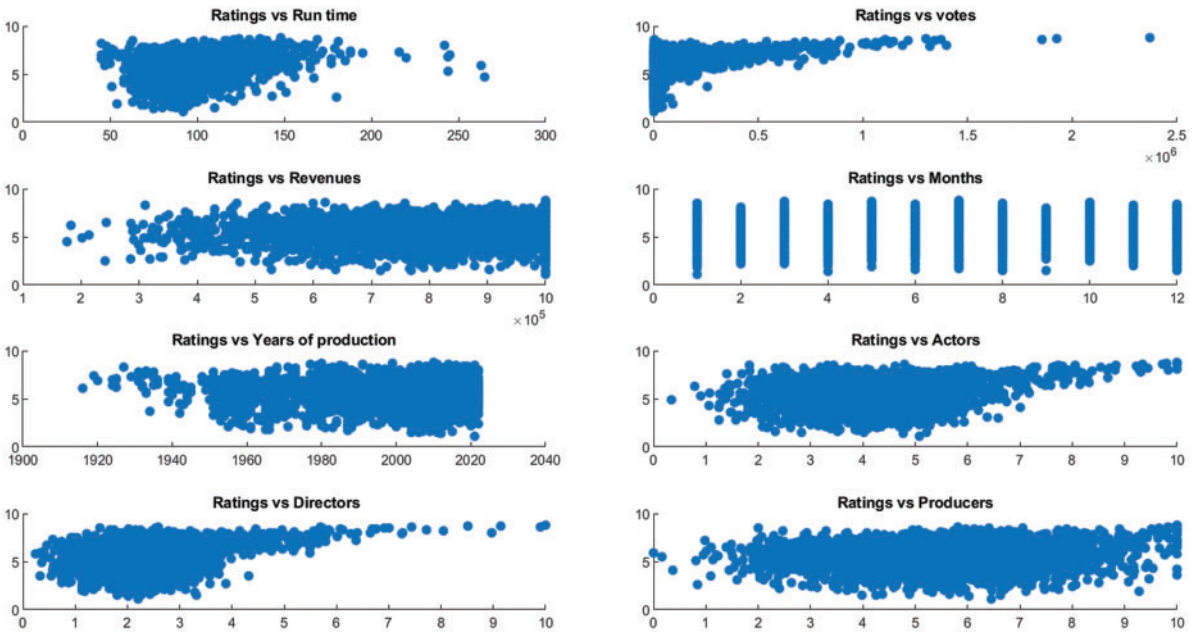

Figure 2: The distribution of the IMDB Rating and a number of modeling features

Table 3 displays the top 10 science fiction films in terms of worldwide box office receipts.

Table 3: The top 10 science fiction films in terms of worldwide box office earnings

Movie	Box office
Avatar	2,847,246,203 $
Star wars: the force awakence	2,068,223,624 $
Jurassic world	1,670,516,444 $
Star wars: the last jedi	1,332,539,889 $
Jurassic world: fallen kingdom	1,308,467,944 $
Transformers: dark of the moon	1,123,794,079 $
Transformers: age of extinction	1,104,054,072 $
Star wars: the rise of skywalker	1,074,144,248 $
Rogue one: a star wars story	1,056,057,273 $
Jurrasic park	1,033,928,303 $

3.3 Machine Learning Algorithms

An abundance of documents, articles, images, files, scientific data, and other forms of data have resulted from the widespread adoption of IT across many industries. One needs a strategy for obtaining meaningful insights from enormous datasets in order to create improved judgments constructed according to the data collected by various applications. Researchers can get more out of their huge datasets by applying knowledge discovery in databases (KDD). Data mining refers to the practice of discovering and extracting significant patterns from large amounts of recorded data. It is executed by a number of methods and algorithms. Statistics, machine learning, pattern recognition, AI, and computational skills are just a few of the fields that have included data mining in their practices. Predictions regarding the effects, performances, and other outcomes of films can be made using the collected data. The nearest Neighbor (NN) classification devised by P. E. Hart and T. M. Cover in 1966 and 1967 has been widely used in industry and academia [44]. Its pervasive adoption can be attributed to its simplicity and high-quality output. Ultimately, the search requires a great deal of memory and processing capacity, but after implementing the NN strategy, these resources were increased and conserved. The objective of constructing NN classifiers is to determine whether or not an input is similar to a given training point neighbor. Due to their high degree of precision, they are extensively incorporated in Data Mining and Machine Learning [45]. This algorithm is also helpful for pattern recognition (see references [44,46], and [47]) and machine learning (see references [44,46], and [47]). The KNN classifier provides classification according to the consensus of the k nearest neighbors, which are the K training cases Xi, i = 1...k most similar to X. Working with training datasets may necessitate a substantial quantity of computing. This is due to the fact that for each new data point, it is necessary to compute the distance between the training points and provide a score for this distance. k, the number of neighbors, is the only factor determining the complexity of KNN. As "k" increases, the development of the categorization borders becomes more uniform. Consequently, KNN becomes less intricate as "k" increases. Supervised learning increases the capacity of KNN models to classify data. The closest prototype classifier determines the class labels for observations using the training samples with the most similar means. Additionally, multiple KNN classifiers can be derived depending on the type of distance metric, its weight, and the number of its neighbors (the k value) [48]. Obtaining

information or knowledge from unprocessed data is difficult, but it is possible. Consequently, there is value in employing specialized algorithms for data processing, leading to data science development.

4 KNN Algorithms

In a multi-dimensional feature space, the training set is represented by a vector for each sample and a class descriptor. During the training phase, the feature vectors and class labels of the training samples are only stored. Throughout the classification process, a user-defined constant k is used to identify which label is more mutual among the other k training points in close proximity to an inquiry location. As a distance metric with a continuous variable, the Euclidean distance [47] is commonly employed. On the other hand, the overlap metric can be used in place of or in addition to text classification (or Hamming distance) to quantify discrete variables. Versions of large margins and component analysis are examples of focused methods that can be used to learn the distance measure, which can affectedly increase the correctness of the KNN classifier [49]. The majority rule does not function where there are big class divides. The majority of a new instance's k nearest neighbors will often be other instances of a more common class. One possibility is to incorporate the test site's distance from its closest neighbor as a weight to the classification process. In regression circumstances, a class or value is improved by giving more credence to the points that are closer to the test point among the set of k nearest points. Abstraction is another tool for dealing with asymmetry in data representation. Regardless of how prevalent a node type was in the original training data, each node in a self-organizing map (SOM) represents the hub of a group of similar nodes. The KNN can resort to the SOM as a last resort [50]. In general, increasing k's value lessens the influence of noise on classification, albeit at the cost of fewer sharply delineated classes. However, the best k will be different for every dataset. Numerous strategies exist to help find the best value for k. When the predicted class is identical to that of the closest training sample, the nearest neighbor technique is applied (i.e., when $k = 1$). In this study, the main parameters that degrade a K-NN algorithm's performance such as noise, irrelevant features, and uneven feature sizes were detected and minimized to improve its accuracy. Selecting and scaling elements of the algorithm to improve its classification has been adopted as part of the evolution of this algorithm. In this work, evolutionary algorithms have been used to determine the optimal scaling of the features and their characteristics. Empirical determination of the best value of k is common practice, and the bootstrap method is often employed for this purpose [51,52]. Feature x is placed in the same class as its nearest neighbor in the simplest implementation of the closest neighbor classifier. Even when the size of the training data set approaches infinity, the one nearest neighbor classifier guarantees an error rate no larger than twice the Bayes error rate (the lowest conceivable error rate given the data distribution). This work intended to classify the dataset of science fiction films using these criteria, determine its classification structure, and show its interconnections with the other items within the datasets. This approach not only will aid in predicting a film's likelihood of success but also it will spot subpar works, giving directors more time to assess and enhance their efforts. Using similarities between features, the KNN is a supervised machine learning algorithm. It can be used for problems involving either classification or regression. Because the training data is preserved and applied to the test point classification process, KNN is considered a classic non-parametric approach [53]. The KNN algorithm assigns a test point to a class based on its nearest neighbors after measuring the distance between it and the training points.

This study will use the most popular distance expression [54].

Minkowski distance:

$$d(x_i, y_i) = \sqrt[p]{|x_i - y_i|^p} \tag{1}$$

Euclidean distance:

$$d(x_i, y_i) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{2}$$

Cosine distance:

$$d(x_i, y_i) = 1 - \frac{x_i y_i}{\sqrt{(x_i x_i)(y_i y_i)}} \tag{3}$$

Cubic Minkowski:

$$d(x_i, y_i) = \sqrt[3]{\sum_{i=1}^{n} (x_i - y_i)^3} \tag{4}$$

where x_i and x_j are the test point and training point respectively.

Fine KNN, Medium KNN, Cosine KNN, Cubic KNN, and Weighted KNN were all used in this study. Table 4 displays the range of neighbor types, the distance metric formula, and the distance weight. We used a variety of KNN algorithms, including Fine KNN, Medium KNN, Cosine KNN, Cubic KNN, and Weighted KNN. The neighbor count, distance metric formulation, and neighbor weight for each grouping are displayed in Table 4. Medium KNN has ten neighbors, cosine KNN has twenty, cube KNN has thirty, and weighted KNN has twenty. Fine KNN has one neighbor, coarse KNN has one hundred, and weighted KNN has ten. The Euclidean distance is used by Fine KNN, Medium KNN, Coarse KNN, and Weighted KNN, whereas the cosine distance is used by cosine KNN and the cubic Minkowski distance is used by cubic KNN. In contrast to Cosine KNN, Cubic KNN, and Weighted KNN, which all utilize the equal distance (no distance weight), Fine, Medium, and Coarse KNN all employ the squared inverse for distance weight.

Table 4: KNN algorithms

Algorithm	Number of neighbors	Distance metric	Distance weight	Standardize data
Fine KNN	1	Euclidean	Equal	True
Medium KNN	10	Euclidean	Equal	True
Coarse KNN	100	Euclidean	Equal	True
Cosine KNN	10	Cosine	Equal	True
Cubic KNN	10	Cubic Minkowski	Equal	True
Weighted KNN	10	Euclidean	Squared inverse	True

5 Results and Discussion

The algorithm was put into action by gathering the data from the IMDb website [55]. The downloaded dataset contains more than a million individual records. Movie title information can be found in the files "title.akas.tsv" and "title.basics.tsv." The 'title.crew.tsv' and 'title.principals.tsv' files include information about the film's directors, actors, and actresses. These files mostly only contain empty space. The 'titles.ratings.tsv' file contains the most critical information, such as the average rating and total votes cast.

Form-fitting and preprocessing: The rating prediction system considers the film's genre, length, budget, the fame of its actors and crew, and the film's aspect ratio. Several sources of ".tsv" data are

used to compile IMDb's database; these files must be translated to ".csv" format and then merged into a single file. In these proceedings, only English-language films are used. Several methods were used to examine the data for inappropriate symbols. For lacking numbers, the median is used, whereas for missing categories, the most common is substituted. In order for models to engage with data, it is necessary to transform categorical and ordinal variables into numerical features [56]. Two subsets, the training dataset and the testing dataset, were created from the processed data. Once the model has been trained on the training data, its performance is measured on the testing data. Eighty percent of the information is used as a training set, while twenty percent is used as a test set. This data-distribution-ratio was chosen in accordance with the 80/20 Pareto principle. When building a model, training is used to find the optimal parameters based on the training data. There are a number of free variables and one dependent one (the rating). Several supervised learning algorithms are used to disentangle objective variables and features. Regression trees, linear regression, and RF are used to train the model. Nonparametric supervised learning uses heuristics like the "simple regression tree" to make predictions. It takes the input and processes it on its own, based on the fundamental model. One-tree models are still completely reliable, though. The supervised learning and estimation technique referred to as "RF" is used to fit several trees of decision to different parts of many datasets. RFs use averaging to increase prediction accuracy and regulate overfitting. Predicting the value of a dependent variable using the attributes of that variable is the goal of supervised learning techniques like linear regression. Using the best-fit feature and the rating relationship, a machine-learning predictor can make an accurate prediction. Accuracy, prediction speed, and training time were not the only metrics compared; the confusion matrix and area under the receiver operating characteristic curve (AUC-ROC) were also studied. The confusion matrix is an N-by-N matrix used by statisticians to evaluate the performance of a classification system. N here stands for the total number of categories, which is 4. In this scenario, we see the expected classes along the x-axis and the real ones along the y-axis. The confusion matrix was used to determine the true positive and false positive rates for each class. The confusion matrix [57] can be used to compute accuracy, precision, and recall, three performance metrics. The receiver operator characteristic (ROC) is a curve that shows how often a test is true positive against how often it is false positive. The Area Under the Curve (AUC) is a statistical measure used to assess the precision with which a classification is made. Figs. 3–8 show respectively the AUC-ROC curves for the fine KNN, medium KNN, coarse KNN, cosine KNN, cubic KNN, and Weighted KNN algorithms. Results from multiple classifiers are summarized here.

5.1 Fine KNN

Accuracy (%): 93.0

Prediction Speed (obs/sec): ∼31000

Training Time (Sec): 0.6328

Class	True positive rate	False positive rate	AUC
Average	96%	4%	0.93
Below average	93%	7%	0.95
Flop	100%	0%	1.0
Hit	77%	23%	0.88

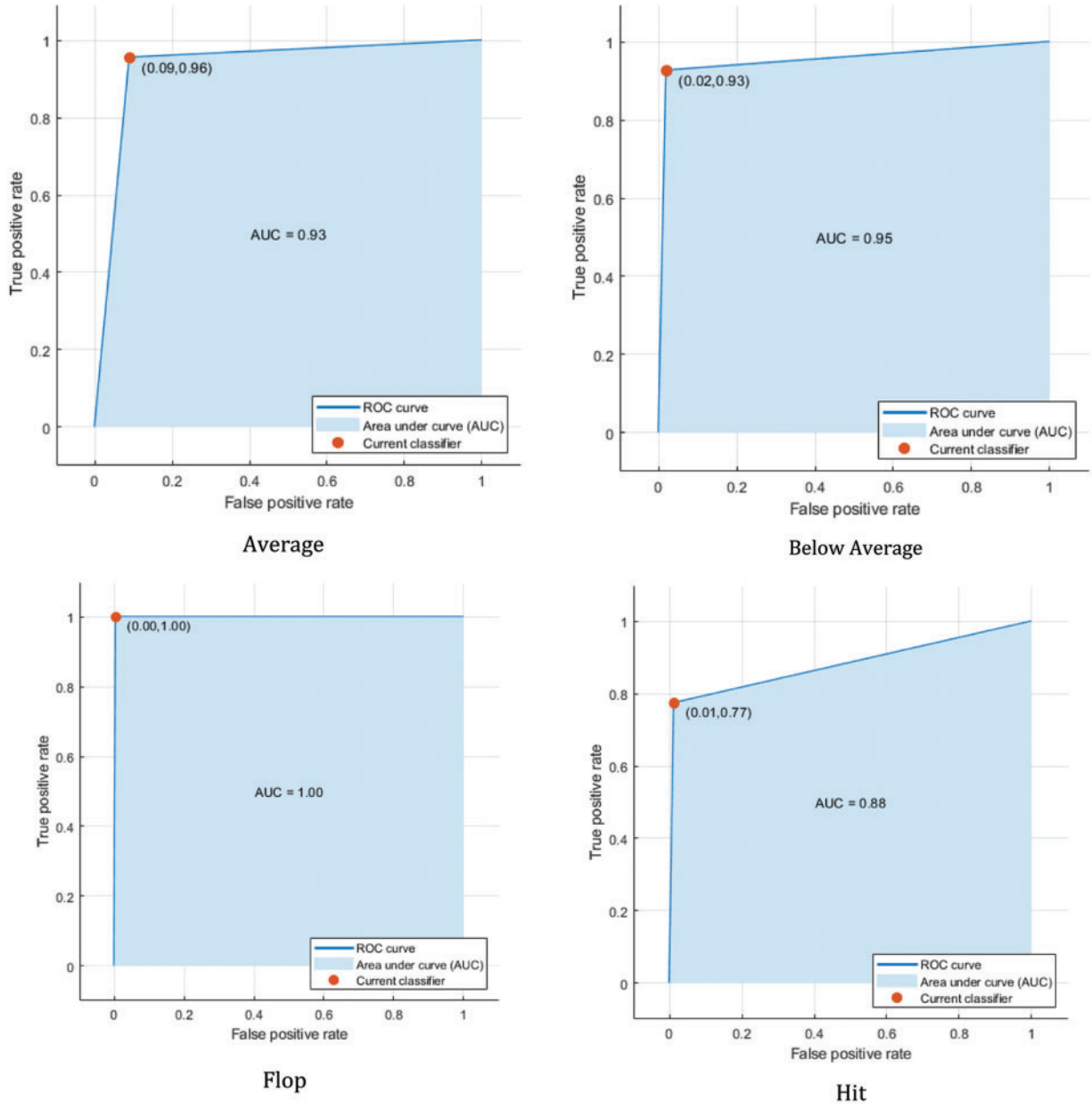

Figure 3: Fine KNN ROC

5.2 Medium KNN

Accuracy (%): 90.3

Prediction Speed (obs/sec): ~24000

Training Time (Sec): 0.21764

Class	True positive rate	False positive rate	AUC
Average	98%	2%	0.98
Below average	89%	11%	0.99
Flop	73%	27%	0.99
Hit	48%	52%	0.97

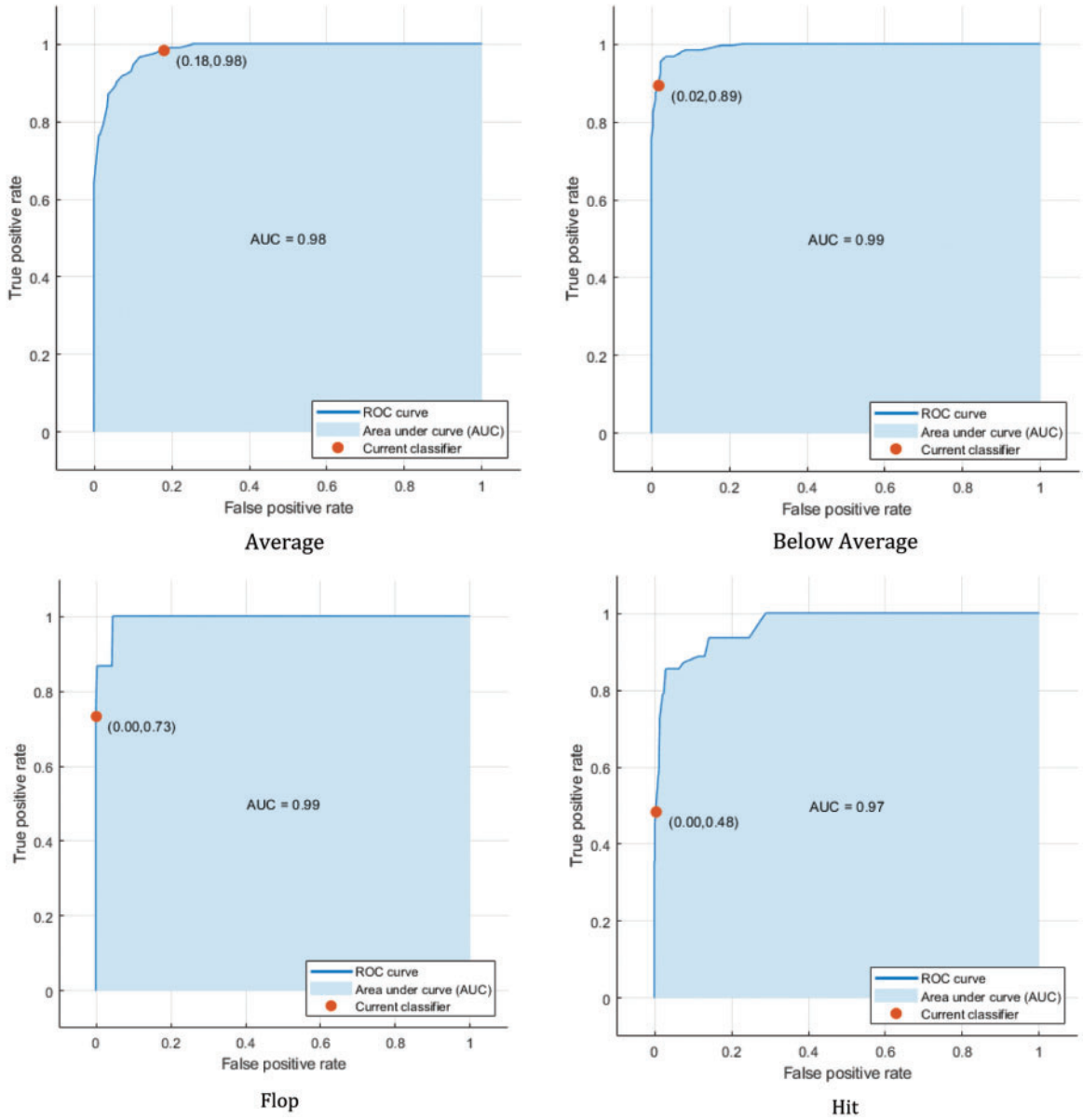

Figure 4: Medium KNN ROC

5.3 Coarse KNN

Accuracy (%): 82.5

Prediction Speed (obs/sec): ~19000

Training Time (Sec): 0.26899

Class	True positive rate	False positive rate	AUC
Average	100%	0%	0.94
Below average	81%	19%	0.97
Flop	0%	100%	0.58
Hit	0%	100%	0.87

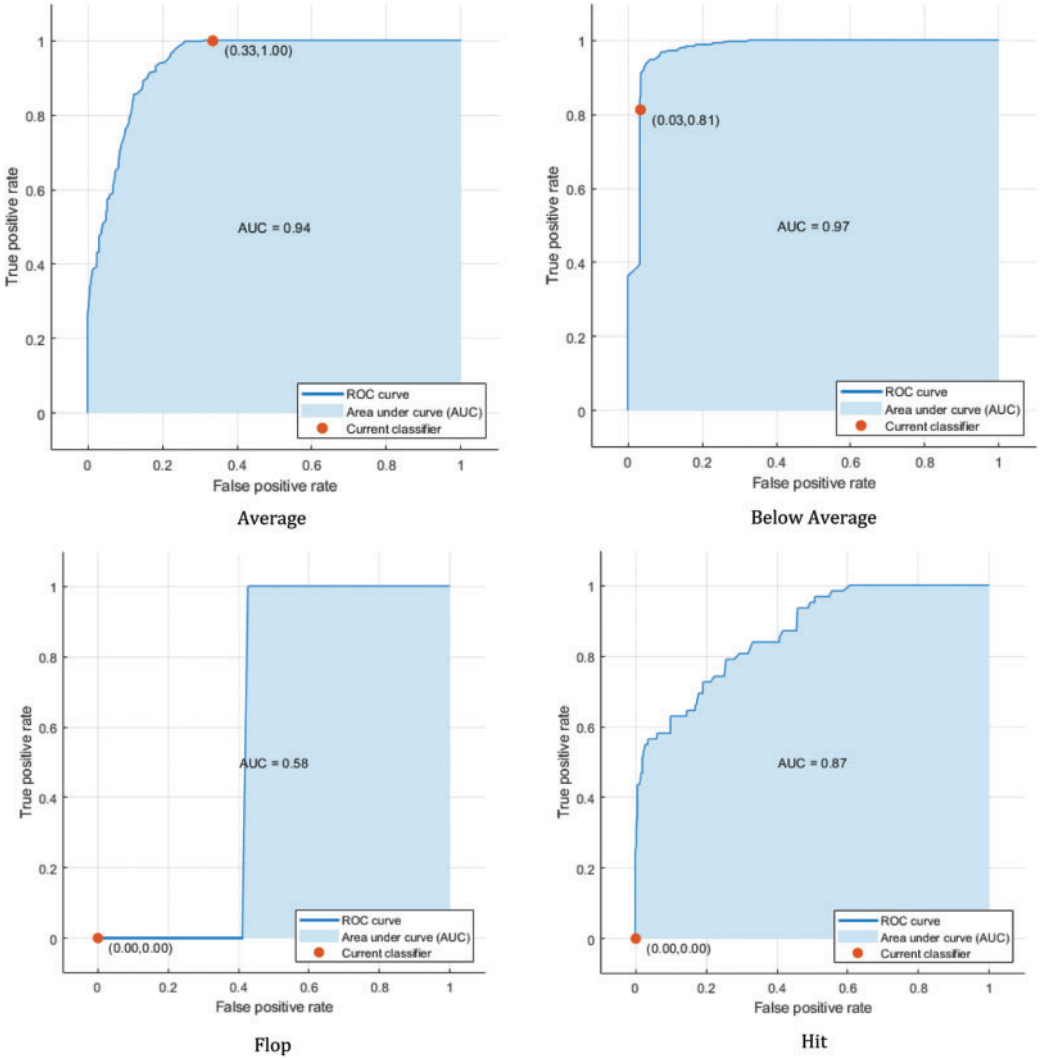

Figure 5: Coarse KNN ROC

5.4 Cosine KNN

Accuracy (%): 86.9

Prediction Speed (obs/sec): ~9400

Training Time (Sec): 0.28923

Class	True positive rate	False positive rate	AUC
Average	94%	6%	0.96
Below average	92%	8%	0.99
Flop	47%	53%	0.98
Hit	32%	68%	0.90

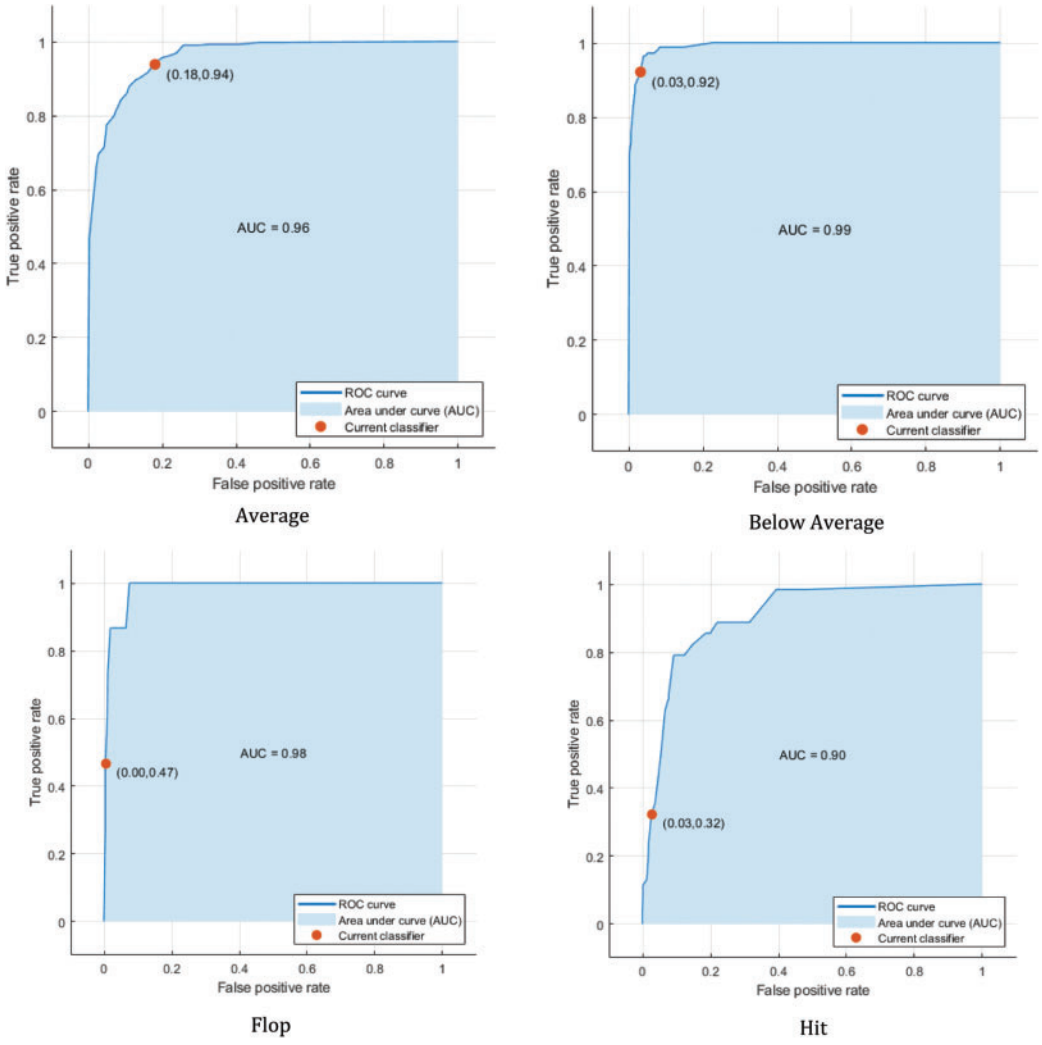

Figure 6: Cosine KNN ROC

5.5 Cubic KNN

Accuracy (%): 89.6

Prediction Speed (obs/sec): ~22000

Training Time (Sec): 0.20158

Class	True positive rate	False positive rate	AUC
Average	98%	2%	0.98
Below average	88%	12%	0.99
Flop	73%	27%	0.99
Hit	45%	55%	0.96

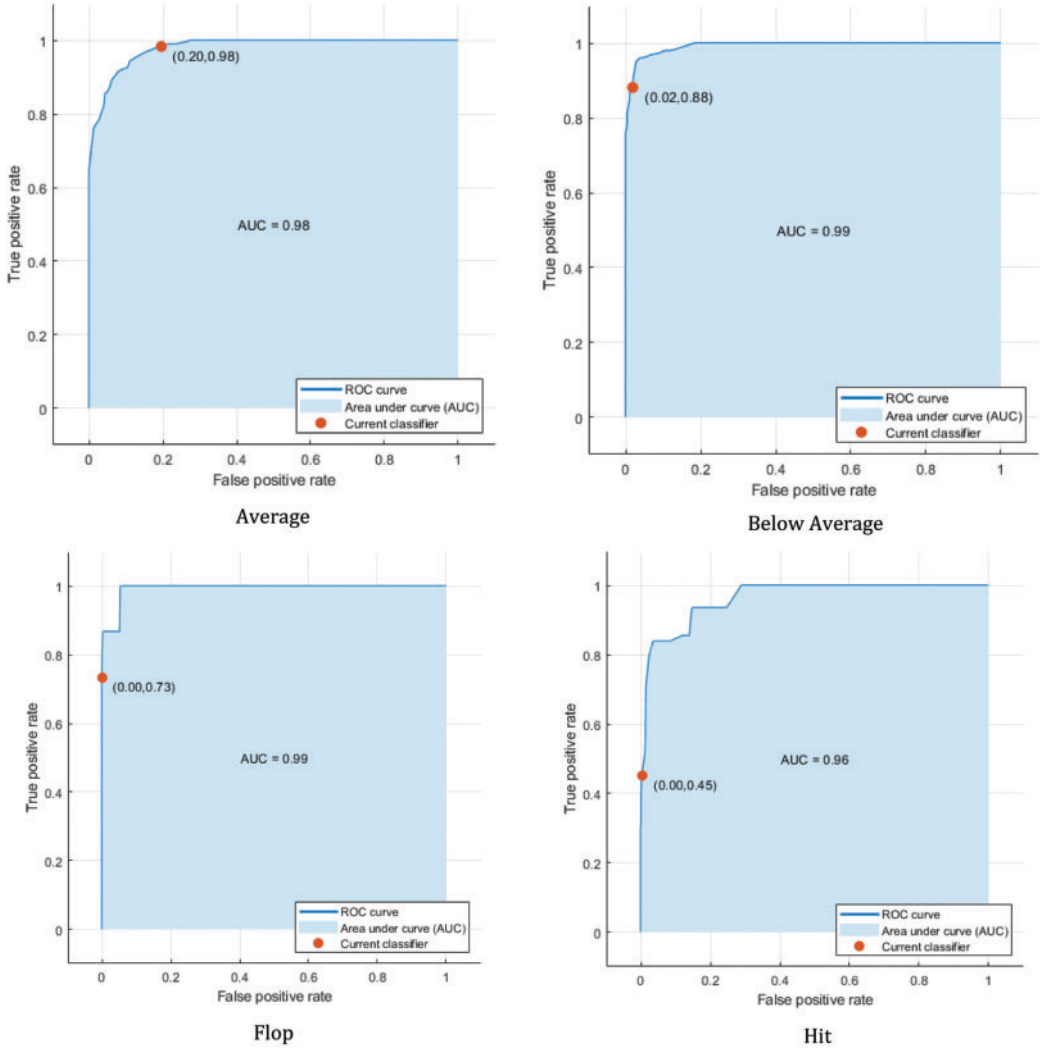

Figure 7: Cubic KNN ROC

5.6 Weighted KNN

Accuracy (%): 92.9

Prediction Speed (obs/sec): ~24000

Training Time (Sec): 0.21903

Class	True positive rate	False positive rate	AUC
Average	97%	3%	0.99
Below average	94%	6%	0.99
Flop	93%	7%	1.0
Hit	60%	40%	0.98

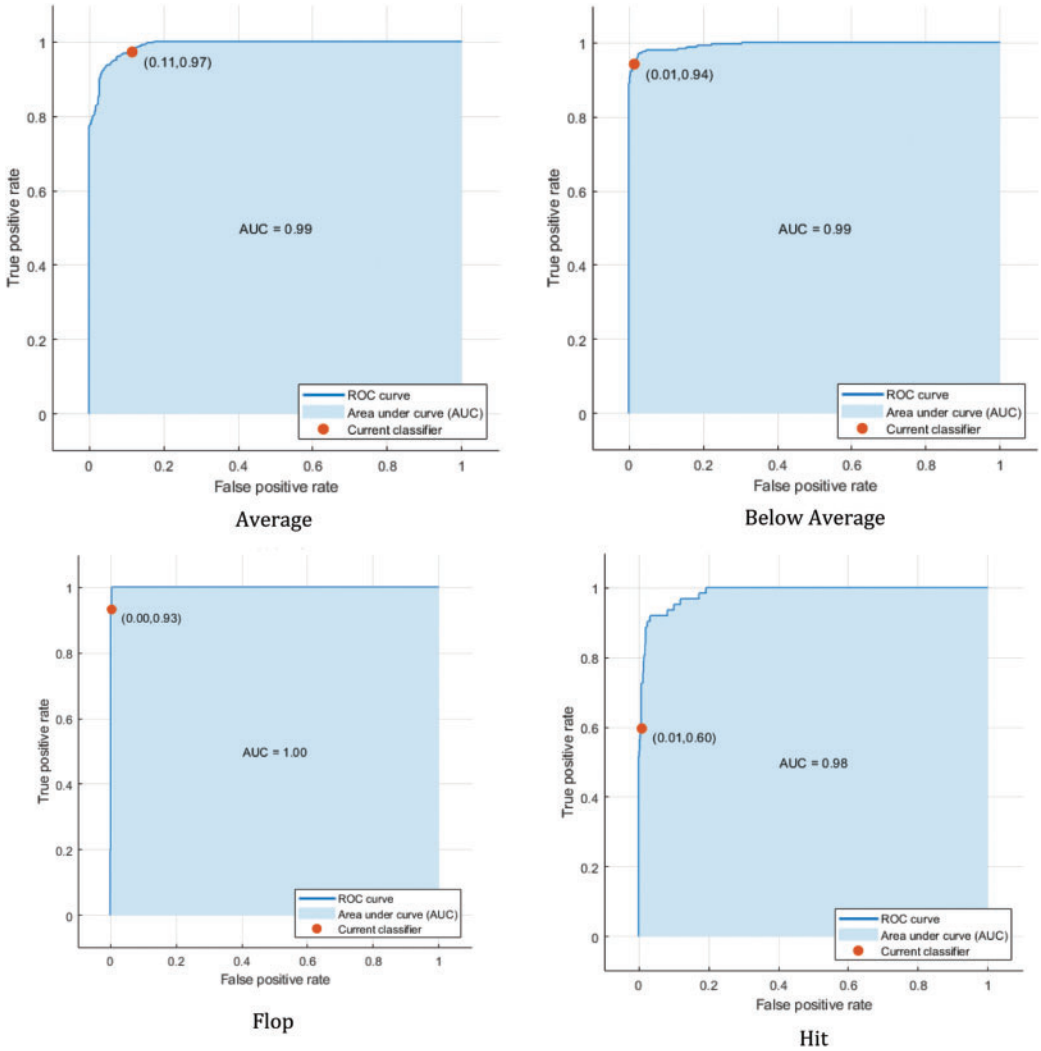

Figure 8: Weighted KNN ROC

This research suggests a different method to alleviate open-space risk by introducing an optimized selection and scaling of KNN algorithms to identify distinguishing characteristics of the feature set. To learn more about k's relevance, a contrastive learning process was conducted with different k values (within a certain range) and evaluated the model's ability to spot out-of-domain intentions under different conditions (while keeping all other hyper-parameters fixed). The results show that the (cosine-based) model's efficacy declines as "k" increases for all datasets. Due to the initial reduction in the fraction of free space, the likelihood that an out-of-domain sample will be incorrectly recognized as an in-domain sample reduces as "k" increases. Over time, as the in-domain semantic space is condensed, a growing share of intra out-of-domain samples will be labeled as in-domain ones, and this distribution will eventually level off. This phenomenon also shows that our method is superior to others in lowering the open-space risk (the danger of having many in-domain intents collide with one another). The highest attainable accuracy, as stated by [43–47], [49,50], was much lower than the figures listed below. The results were compared to those of other investigations, which are reported in [58,59]. In [58], we have a comparison table for a number of machine learning algorithms. The model results for logistic regression did poorly on the dataset, with an accuracy of 59.4%. However, several models have shown success with an accuracy of 87% or higher, including the decision tree, RF, Ada boosting, and gradient boosting. Their recently deployed ensemble model with act-direct has improved accuracy over individual models to the tune of 92.8%. Their research suggests this unique quality is essential. Their model's test results simply could not have improved without the new feature. They suggested an algorithmic approach based on machine learning techniques to determine in advance the likelihood of a film's success or failure [59]. Their research made use of the IMDB dataset. They brought in six algorithms, and the results varied widely: 72.18% accuracy for Nave Bayes, 81.04% accuracy for Decision Tree, 73.76% accuracy for K-Nearest Neighbor, 72.68% accuracy for Support Vector Machine, 73.26% accuracy for Logistic Regression, and 85.2% accuracy for RF. A model for sentiment analysis is proposed in [60]. Of the three models we looked at, the linear SVC model had the highest accuracy at 88.47%; the logistic regression model gave us 84.80%, and the Naive Bayes model gave us 83.8%. The proposed approach is used to rank the best 10 movies in a given year. The accuracy of the hybrid model exceeds that of its individual parts. This indicates that the predicted outcomes closely match the actual outcomes. Different types of KNN algorithms were used in this study, as shown in Fig. 9: Fine KNN with 93.0% accuracy, Medium KNN with 90.3% accuracy, Coarse KNN with 82.5% accuracy, Cosine KNN with 86.9% accuracy, Cubic KNN with 89.6% accuracy, and Weighted KNN with 92.9% accuracy.

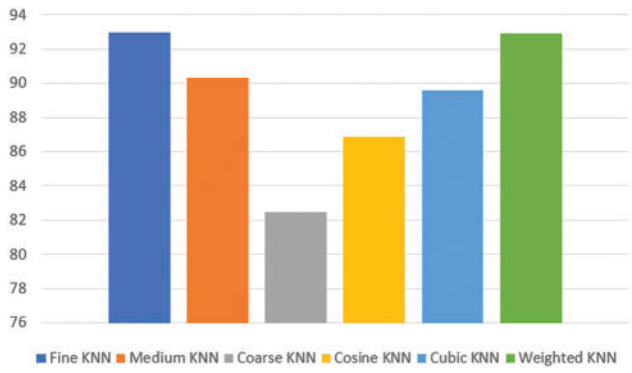

Figure 9: Accuracy for KNN algorithms

6 Conclusion

The purpose of optimization for aspects outside the scope of this article is explained in depth. The limitations and drawbacks of existing methods were discussed, in addition, a simple method for learning discriminative semantic attributes is proposed to enhance the performance of the KNN classifier. The in-domain intents were clustered with their k-nearest neighbors to alleviate experimental and open domain jeopardies and isolate them from other class samples. Numerous experiments on a challenging dataset demonstrate that the proposed classification method produces results without limiting feature distribution. Various KNN algorithms were utilized to predict the successful performance of science fiction films. The movies were ranked as flop, below average, average, and hit using data collected from IMDB. The results indicate that Fine, Weighted, Medium, and Cubic KNNs are more accurate than Cosine and Crude KNNs. All KNN variants have prediction times under 0.7 s. To further refine the prediction model, future work is intended to apply and compare multiple machine learning methods and incorporate new criteria from non-IMDB sources, such as Rotten Tomatoes and Wikipedia.

Acknowledgement: The authors appreciate the dataset's developers for making it accessible online. They also like to express their gratitude to the anonymous reviewers who helped them to make this paper better. Yarmouk University's support is acknowledged.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] "Statista Research Department, Which is your favourite source of entertainment?," 2012. https://www.statista.com/statistics/248352/mostpreferred-sources-of-entertainment-in-india/(2012)

[2] A. Watson, "Film industry in the United States and Canada—Statistics & facts," https://www.statista.com/topics/964/film/ (accessed October. 1, 2022).

[3] U. Nsikak-Abasi and A. Joy Stephen, "Assessing the impact of modern movies on students—A prospective study," *Journal of Culture, Society and Development*, vol. 31, pp. 1–11, 2017.

[4] L. Kirsten, P. Michiel, H. J. S. Ron and C. M. E. E. Rutger, "Effects of smoking cues in movies on immediate smoking behavior," *Nicotine & Tobacco Research*, vol. 12, pp. 913–918, 2010.

[5] V. Nikolaos and M. Loumioti, "Movies as a tool of modern tourist marketing," *Tourismos*, vol. 6, no. 2, pp. 353–362, 2011.

[6] L. Kuen-Yi, T. Fu-Hsing, H. M. Chen and C. Liang-Te, "Effects of a science fiction film on the technological creativity of middle school students," *Eurasia Journal of Mathematics, Science and Technology Education*, vol. 9, no. 2, pp. 191–200, 2013.

[7] D. Lorenčík, M. Tarhaničová and P. Sinčák, "Influence of Sci-Fi films on artificial intelligence and vice-versa," in *2013 IEEE 11th Int. Symp. on Applied Machine Intelligence and Informatics (SAMI)*, Herl'any, Slovakia, IEEE, pp. 27–30, 2013.

[8] T. Lowe, K. Brown, S. Dessai, M. de França Doria, K. Haynes *et al.,* "Does tomorrow ever come? Disaster narrative and public perceptions of climate change," *Public Understanding of Science*, vol. 15, no. 4, pp. 435–457, 2006.

[9] J. L. Stimpert, J. Laux, C. Marino and G. Gleason, "Factors influencing motion picture success: Empirical review and update," *Journal of Business & Economics Research (JBER)*, vol. 6, no. 11, pp. 39–51, 2008.

[10] E. Hafeez and S. Hassan, "Why Do Cash Cow Movies Work?" *Journal of Media & Communication*, vol. 2, no. 1, pp. 82–94, 2021.

[11] P. D. Turney, "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews," arXiv preprint cs/0212032, 2002. https://doi.org/10.48550/arXiv.cs/0212032
[12] B. Pang, L. Lee and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," arXiv preprint cs/0205070, 2002. https://doi.org/10.3115/1118693.1118704
[13] G. Mishne and N. S. Glance, "Predicting movie sales from blogger sentiment," in *AAAI Spring Symp.: Computational Approaches to Analyzing Weblogs*, CA, USA, pp. 155–158, 2006.
[14] A. Kennedy and D. Inkpen, "Sentiment classification of movie reviews using contextual valence shifters," *Computational Intelligence*, vol. 22, no. 2, pp. 110–125, 2006. https://doi.org/10.1111/j.1467-8640.2006.00277.x
[15] W. Zhang and S. Skiena, "Improving movie gross prediction through news analysis," in *2009 IEEE/WIC/ACM Int. Joint Conf. on Web Intelligence and Intelligent Agent Technology*, IEEE, vol. 1, pp. 301–304, 2009. https://doi.org/10.1109/WI-IAT.2009.53
[16] S. I. Wang and C. D. Manning, "Baselines and bigrams: Simple, good sentiment and topic classification," in *Proc. of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 90–94, 2012.
[17] M.Á. García-Cumbreras, A. Montejo-Ráez and M. C. Díaz-Galiano, "Pessimists and optimists: Improving collaborative filtering through sentiment analysis," *Expert Systems with Applications*, vol. 40, no. 17, pp. 6758–6765, 2013. https://doi.org/10.1016/j.eswa.2013.06.049
[18] G. Mesnil, T. Mikolov, M. A. Ranzato and Y. Bengio, "Ensemble of generative and discriminative techniques for sentiment analysis of movie reviews," arXiv preprint arXiv:1412.5335, 2014. https://doi.org/10.48550/arXiv.1412.5335
[19] P. Nagamma, H. R. Pruthvi, K. K. Nisha and N. H. Shwetha, "An improved sentiment analysis of online movie reviews based on clustering for box-office predictionm," in *Int. Conf. on Computing, Communication & Automation*, Greater Noida, India, IEEE, pp. 933–937, 2015. https://doi.org/10.1109/CCAA.2015.7148530
[20] A. Yenter and A. Verma, "Deep CNN-LSTM with combined kernels from multiple branches for IMDb review sentiment analysis," in *2017 IEEE 8th Annual Ubiquitous Computing, Electronics and Mobile Communication Conf. (UEMCON)*, New York, NY, USA, IEEE, pp. 540–546, 2017. https://doi.org/10.1109/UEMCON.2017.8249013
[21] A. U. Rehman, A. K. Malik, B. Raza and W. Ali, "A hybrid CNN-LSTM model for improving accuracy of movie reviews sentiment analysis," *Multimedia Tools and Applications*, vol. 78, no. 18, pp. 26597–26613, 2019. https://doi.org/10.1007/s11042-019-07788-7
[22] G. Charu, G. Chawla, K. Rawlley, K. Bisht and M. Sharma, "Senti_ALSTM: Sentiment analysis of movie reviews using attentionbased-LSTM," in *Proc. of 3rd Int. Conf. on Computing Informatics and Networks*, Singapore, Springer, pp. 211–219, 2021. https://doi.org/10.1007/978-981-15-9712-1_18
[23] G. Verma and H. Verma, "Predicting bollywood movies success using machine learning technique," in *2019 Amity Int. Conf. on Artificial Intelligence (AICAI)*, Dubai, United Arab Emirates, IEEE, pp. 102–105, 2019. https://doi.org/10.1109/AICAI.2019.8701239
[24] Z. Shaukat, A. A. Zulfiqar, C. Xiao, M. Azeem and T. Mahmood, "Sentiment analysis on IMDB using lexicon and neural networks," *SN Applied Sciences*, vol. 2, no. 2, pp. 1–10, 2020. https://doi.org/10.1007/s42452-019-1926-x
[25] S. Pirunthavi, P. R. Vithusia, K. Abishankar, E. M. U. W. J. B. Ekanayake and M. Yanusha, "Movie success and rating prediction using data mining algorithm," 2020. https://doi.org/10.13140/RG.2.2.18052.86402
[26] K. Pradeep, C. R. T. Rosmin, S. S. Durom and G. S. Anisha, "Decision tree algorithms for accurate prediction of movie rating," in *2020 Fourth Int. Conf. on Computing Methodologies and Communication (ICCMC)*, Erode, India, IEEE, pp. 853–858, 2020. https://doi.org/10.1109/ICCMC48092.2020.ICCMC-000158
[27] J. D. Bodapati, N. Veeranjaneyulu and S. Shaik, "Sentiment analysis from movie reviews using LSTMs," *Ingenierie des Systemes d'Information*, vol. 24, no. 1, pp. 125–129, 2019. https://doi.org/10.18280/isi.240119

[28] S. Tiwari, S. Kumar, V. Jethwani, D. Kumar and V. Dadhich, "PNTRS: Personalized news and tweet recommendation system," *Journal of Cases on Information Technology (JCIT)*, vol. 24, no. 3, pp. 1–19, 2022. https://doi.org/10.4018/JCIT.20220701.oa9

[29] M. T. Lash and Z. Kang, "Early predictions of movie success: The who, what, and when of profitability," *Journal of Management Information Systems*, vol. 33, no. 3, pp. 874–903, 2016.

[30] V. R. Nithin, M. Pranav, P. B. Sarath Babu and A. Lijiya, "Predicting movie success based on IMDb data," *International Journal of Data Mining Techniques and Applications*, vol. 3, pp. 365–368, 2014.

[31] M. H. Latif and A. Hammad, "Prediction of movies popularity using machine learning techniques," *International Journal of Computer Science and Network Security (IJCSNS)*, vol. 16, no. 8, pp. 127, 2016.

[32] K. I. Asad and Md S. Rahman, "Movie popularity classification based on inherent movie attributes using C4. 5, PART and correlation coefficient," in *2012 Int. Conf. on Informatics, Electronics & Vision (ICIEV)*, Dhaka, Bangladesh, IEEE, pp. 747–752, 2012.

[33] M. Saraee, W. Sean and J. Eccleston, "A data mining approach to analysis and prediction of movie ratings," *WIT Transactions on Information and Communication Technologies*, vol. 33, pp. 343–352, 2004. https://doi.org/10.2495/DATA040331

[34] M. Galvão and R. Henriques, "Forecasting movie box office profitability," *Journal of Information Systems Engineering & Management*, vol. 3, no. 3, pp. 1–9, 2018.

[35] D. Gaikar, R. Solanki, H. Shinde, P. Phapale and I. Pandey, "Movie success prediction using popularity factor from social media," *International Research Journal of Engineering and Technology*, vol. 6, no. 4, pp. 5184–5190, 2019.

[36] M. T. Lash, S. Fu, S. Wang and K. Zhao, "Early prediction of movie success: What, who, and when," in *Proc. of the 2015 Int. Conf. on Social Computing, Behavioral-Cultural Modeling, and Prediction*, N. Agarwal, K. Xu, N. Osgood (Eds.), Washington, DC: Springer, pp. 345–349, 2015.

[37] J. Ahmad, P. Duraisamy, A. Yousef and B. Buckles, "Movie success prediction using data mining," in *2017 8th Int. Conf. on Computing, Communication and Networking Technologies (ICCCNT)*, Delhi, India, pp. 1–4, 2017. https://doi.org/10.1109/ICCCNT.2017.8204173

[38] D. Rijul and A. Raj, "Movie success prediction using machine learning algorithms and their comparison," in *2018 First Int. Conf. on Secure Cyber Computing and Communication (ICSCCC)*, Jalandhar, India, IEEE, pp. 385–390, 2018.

[39] P. Chakraborty, M. Z. Rahman and S. Rahman, "Movie success prediction using historical and current data mining," *International Journal of Computer Applications*, vol. 178, no. 47, pp. 1–5, 2019.

[40] P. Sivakumar, V. Rajeswaren, K. Abishankar, J. Ekanayake and Y. Mehendran, "Movie success and rating prediction using data mining algorithms," *Journal of Information Systems & Information Technology (JISIT)*, vol. 5, no. 2, pp. 72–80, 2020.

[41] M. Agarwal, S. Venugopal, R. Kashyap and R. Bharathi, "A comprehensive study on various statistical techniques for prediction of movie success," arXiv preprint arXiv: 2112.00395, 2021.

[42] V. Gupta, N. Jain, H. Garg, S. Jhunthra, S. Mohan *et al.,* "Predicting attributes based movie success through ensemble machine learning," *Multimedia Tools and Applications*, vol. 82, pp. 9597–9626, 2022. https://doi.org/10.1007/s11042-021-11553-0

[43] W. R. Bristi, Z. Zaman and N. Sultana, "Predicting IMDb rating of movies by machine learning techniques," in *2019 10th Int. Conf. on Computing, Communication and Networking Technologies (ICCCNT)*, Kanpur, India, IEEE, pp. 1–5, 2019. https://doi.org/10.1109/ICCCNT45670.2019.8944604

[44] A. N. Papadopoulos and Y. Manolopoulos, *Nearest Neighbor Search: A Database Perspective*. New York, USA: Springer Science & Business Media, 2006.

[45] I. Kononenko and K. Matjaz, *Machine Learning and Data Mining*. Chichester, West Sussex, UK: Horwood Publishing, 2007.

[46] G. Shakhnarovich, D. Trevor and I. Piotr, "Nearest-neighbor methods in learning and vision," *IEEE Trans. Neural Networks*, vol. 19, no. 2, pp. 377, 2008.

[47] X. Wu, V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang *et al.,* "Top 10 algorithms in data mining," *Knowledge and Information Systems*, vol. 14, no. 1, pp. 1–37, 2008.

[48] R. Heese, J. Schmid, M. Walczak and M. Bortz, "Calibrated simplex-mapping classification," *PLoS One*, vol. 18, no. 1, pp. e0279876, 2023.
[49] P. A. Jaskowiak and R. J. G. B. Campello, "Comparing correlation coefficients as dissimilarity measures for cancer classification in gene expression data," in *Proc. of the Brazilian Symp. on Bioinformatics*, Brazil, Brasília, 2011.
[50] G. Burgot, F. Auffret and J. -L. Burgot, "Determination of acetaminophen by thermometric titrimetry," *Analytica chimica acta*, vol. 343, no. 1–2, pp. 125–128, 1997.
[51] F. Nigsch, A. Bender, B. van Buuren, J. Tissen, E. Nigsch *et al.,* "Melting point prediction employing k-nearest neighbor algorithms and genetic parameter optimization," *Journal of Chemical Information and Modeling*, vol. 46, no. 6, pp. 2412–2422, 2006.
[52] P. Hall, U. P. Byeong and R. J. Samworth, "Choice of neighbor order in nearest-neighbor classification," *The Annals of Statistics*, vol. 36, no. 5, pp. 2135–2152, 2008.
[53] C. Zhang, P. Zhong, M. Liu, Q. Song, Z. Liang *et al.,* "Hybrid metric K-Nearest neighbor algorithm and applications," *Mathematical Problems in Engineering*, vol. 2022, Article ID 8212546, pp. 1–15, 2022.
[54] M. Jawthari and V. Stoffová, "Predicting students' academic performance using a modified kNN algorithm," *Pollack Periodica*, vol. 16, no. 3, pp. 20–26, 2021.
[55] IMDb dataset. imdb.com. https://www.imdb.com/interfaces/. Published (1996), (accessed October 1, 2022).
[56] A. S. Al Fahoum and T. A. Ghobon, "Performance predictions of Sci-Fi films via machine learning," *Applied Sciences*, vol. 13, no. 7, pp. 4312, 2023.
[57] U. Shahadat, I. Haque, H. Lu, M. A. Moni and E. Gide, "Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction," *Scientific Reports*, vol. 12, no. 1, pp. 1–11, 2022.
[58] G. Verma, H. Verma and S. K. Dixit, "A hybrid ensemble machine learning model to predict success of Bollywood movies," *World Review of Entrepreneurship, Management and Sustainable Development*, vol. 17, no. 2–3, pp. 343–357, 2021.
[59] S. Sadashiv, S. Sween and S. Sankruth, "Movie success prediction using machine learning," *International Research Journal of Modernization in Engineering Technology and Science*, vol. 3, pp. 2021–2024, 2021.
[60] J. Tripathi, S. Tiwari, A. Saini and S. Kumari, "Prediction of movie success based on machine learning and twitter sentiment analysis using internet movie database data," *Indones. J. Electr. Eng. Comput. Sci*, vol. 2023, no. 29, pp. 1750–1757, 2023.