Inferential Analysis of Amazon's Top 50 Bestselling Books

Rishika[1, a)], Sneha Singh[2, b)], Nonita Sharma[3, c)], Monika Mangla[4, d)]

[1]*Electronics and Communication, Indira Gandhi Delhi Technical University for Women*

Delhi, India

[2]*Electronics and Communication, Indira Gandhi Delhi Technical University for Women*

Delhi, India

[3]*Information Technology Department, Indira Gandhi Delhi Technical University for Women*

Delhi, India

[4]*Department of Information Technology, Dwarkadas J Sanghvi College of Engineering*

Mumbai, India

[a)] *rishikasuhag04@gmail.com*
[b)] sneha949759@gmail.com
[c)] nonitasharma@igdtuw.ac.in
[d)] monika.mangla@djsce.ac.in

Abstract. People today lack the time to go through the synopsis and prologue of every book to determine if they genuinely wanted to read it or not. This may cause missing reading of some really good books for a reader. Hence, books' rating (based on choice of other readers) is made available to the readers so that they can decide which books to read necessarily. This review or rating of the book is generated based on the genre and author of the book. In this study, authors have implemented machine learning models like linear regression and logistic regression for the same. Metrics such as precision-recall curve and AUC-ROC curve are used to determine the rating of the book using different datasets from the data frame. The experimental evaluation is done in the Google Collaboratory platform where authors aim to evaluate the books' reviews using a numerical dataset. The results obtained during the experimental evaluation are encouraging and hence advocate the implementation of such models at large.

INTRODUCTION

The idea of books as a gateway towards information and wisdom is prevalent. Whether they're nonfiction or fiction, they are always valuable in some way [1]. While fiction texts can occasionally be used as a coping mechanism to help readers tune out of the world around them and enter a new fantasy world known as books; non-fiction books take us into the past, present, or future so that we can relive those events. They are based on facts, real-life events, or scenarios.

We can learn the basic idea about the book through its synopsis to save time. This is based on book reviews that critics and readers have written about the books. They help us to figure out whether it fits our reading choice or not. Current research firstly consists of a comparative analysis using histograms, line plots, and lmplots to analyze the rating distribution and the correlation of rating with time and price [2]. Machine learning models like linear regression and logistic regression have been used along with precision-recall and AUC-ROC curve to predict the reviews based on various attributes. These models are predicting the review/rating of the book, a dependent value [3][4].

The two Machine learning models used are- Linear Regression and Logistic Regression. Both the models lie under supervised learning techniques i.e., both of them use a labeled dataset to make predictions and find the accuracy and precision. Linear regression solves regression problems whereas logistic regression solves the classification problems. The models are being used for comparative analysis of the dataset. This analysis tells us which of the two models provides us more accurate and precise predictions.

Now during survey of literature in the related field, it is observed that numerous researchers have directly implemented different machine learning algorithms. The efficiency of the different machine learning algorithms is determined during comparative analysis through various performance metrics Viz. Accuracy, precision, AUC, and many more. However, in the current work authors strongly believe in performing empirical analysis of the data through different statistics and visualization tools ahead of employing machine learning algorithms. Empirical analysis of the data gives the authors a better choice to select the machine learning algorithm among different available.

The goal of this research is to identify and classify the various fictional and non-fictional books based on its reviews and ratings by different readers. Authors are also looking for devising some ML model in order to achieve the highest accuracy of all. There are several datasets considered and it is experimented that how efficiently different ML algorithms behave for different datasets. Further, 10-fold cross-validation test, a Machine Learning Technique is used in Google Collaborator to evaluate the efficiency and effectiveness of different machine learning algorithms.

Current research work target to associate rating with books is organized into various sections. Here, section I lays the brief foundation for the need of the research. Similar work done by different researchers is presented in section II. Current methodology proposed by authors in presented in section III and the results are discussed in section IV. Finally concluding remarks are presented in section V.

RELATED WORK

In this section, authors aim to present the similar work undertaken by different researchers. As per the work by authors in [5], it is a good initiative to share users' experience with other users so that others are able to make an educated choice among various options. This review can be given in form of description or in form of some rating (quantitative indicator). Authors in [5] have performed a study to unleash the association among review and rating of a product. For the same, the descriptive review about the product is processed to predict the rating. Authors have used the data from Amazon.com for this evaluation.

Further, authors in [6] also have worked in the direction of evaluating the rating of a book based on various attributes. For the same, authors have used the Goodreads dataset to determine the association to determine the association among various attributes of a book viz. book rating, author's rating, and language of the book. For the same, authors have proposed an Artificial Neural Network (ANN) model to predict the rating of books. The ANN model is created in JNN environments and yields an accuracy of 99.78% supporting the efficacy of proposed model.

Now while there are several models to determine the rating of a product, authors in [7] claim that it is not completely correct to rely on online reviews as the reviews can be manipulated. This manipulation of reviews is done by the various bodies to increase the sale of their product. The irony of this system is that it is difficult to differentiate genuine reviews and fake review and hence is left to intuition of the user only. To address this issue, authors in [7] proposed a system using statistical method to identify manipulation in online reviews. Proposed system also aims to determine the reaction of users to manipulated reviews. The system also performs sentiment analysis in the descriptive reviews of the book. It is found that around 10.3% of the products have manipulated reviews. Hence, the work by authors in [7] can be taken as a cautionary note by users ahead of relying solely on online reviews so that they don't get deceived through manipulated reviews by fraudulent firms.

Authors in [8] have also worked in the direction of predicting numeric rating of a product based on descriptive comments of users. For the same, authors used Unigram and n-gram representations of the comments. Authors also discuss the inability of unigram to capture multiword feedback which can be handled by n-grams. Further, authors also propose concept of bag-of-opinions where each opinion is given a numeric rating. Additionally, the paper also discusses a constrained ridge regression algorithm to know scores based on reviews. Experimental evaluation demonstrate that proposed method outperforms the traditional method of review rating.

PROPOSED METHODOLOGY

The proposed methodology comprises of various steps viz. data collection, data pre-processing, employment of ML models, performance evaluation and finally comparative analysis [9]. The step-wise description of proposed methodology is given in Fig. 1.

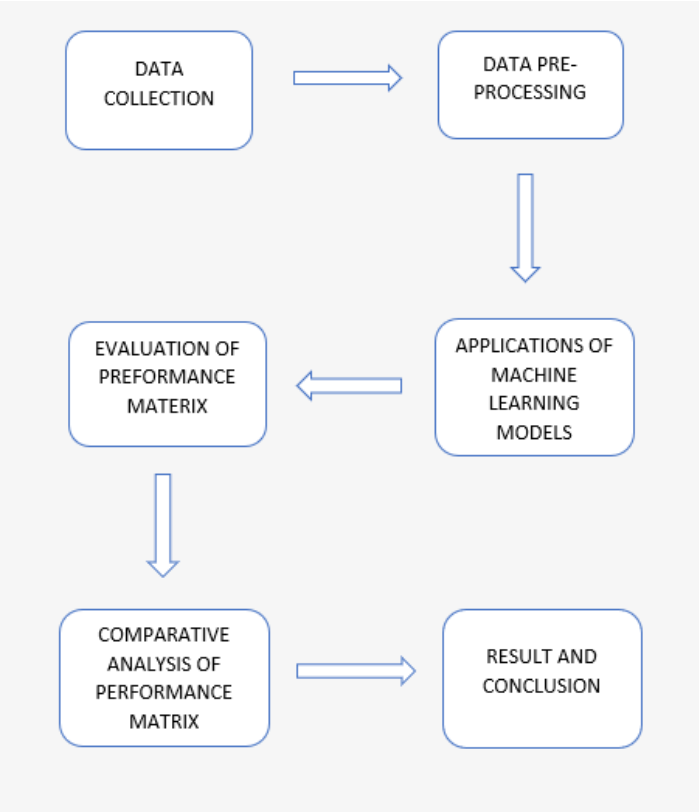

FIGURE 1. *Flowchart showcasing different phases of Machine Learning analysis.*

Data Collection

The dataset is taken from Kaggle [10]. This dataset was obtained from the Amazon website and Goodreads Website. It shows the Top 50 best-selling books on Amazon from 2009-2019. It was scrapped from the amazon website in October 2020. The 550 books in the dataset have been categorized into fiction and non-fiction books. There are 7 columns in the dataset with different attributes namely, Name, Author, User Rating, Reviews, Price, Year and Genre. The data classified is 56% of the non-fiction books and the rest of the 44% is fiction books.

Data Pre-processing

The process of transforming unstructured data into structured data, as well as resizing and deleting unwanted data from a dataset, is referred to as "data pre-processing". The mean value is used to fill in the qualities that the dataset lacks. To ensure proper distribution, the data is then randomly chosen from the dataset.

Testing and Training Phase

The testing phase provides fresh data to evaluate how well our algorithm performs and behaves when it comes to prediction. As mentioned earlier, the dataset is divided into two parts. The process of cross-validation is used to prevent fitting. The data is divided into ten halves for each iteration of our approach, nine of which are utilized for training while the remaining for testing.

Here, authors are comparing two different algorithms in a numerical dataset to determine which is most accurate and appropriate for the dataset. Algorithms used are Linear regression and Logistic regression. Along with them Precision-Recall curve and AUC-ROC curve are also used to get the highest accuracy and precision.

Linear Regression

It is one of the most prominently used regression and Machine Learning techniques. It has a varying ease of interpreting results. It is used to derive the connection between two or more variables. In simple linear regression, one of the variables is considered Independent while the other is called as Dependent. Linear Regression in a Machine learning environment can also be used to make simple predictions.

Here's how a simple linear regression algorithm works:

? Define the dependent and independent variables.

? Fitting the model: to obtain a line that best fits the data points.

? Then, regression plots and equations are obtained.

Logistic Regression

It is the technique that is used to analyze a dataset which has a dependent variable and one or more independent variables. It predicts the outcome in binary variables i.e., it will have only two outputs. It only predicts the output of the algorithm in categorical variable. Log function is used to predict the probability of events.

A simple logistic regression works in the following ways:

? Collecting and analyzing data.

? Training and testing the data.

? conclude the results and retrieving the accuracy.

Precision-Recall Curve

The precision recall curve is the graphical representation of precision and recall retrieved from the data. It is basically used in the cases where data is heavily imbalanced. The values of precision are placed on the y-axis and recall on the x-axis. This graph helps us highlight how relevant the achieved results are. Therefore, precision-recall curve is very commonly used in the problems that requires some information to be retrieved.

AUC-ROC Curve

The ROC i.e., Receiver Operator Characteristic is a probability curve that separates noise from the signal. The AUC (Area Under Curve) is basically a summary of the ROC probability curve. It measures the ability of classifier to separate the different classes.

Hence, AUC-ROC is one of the most significant evaluation metrics for examining a model’s performance. It helps us visualize how well our machine learning classifier is working. The higher the AUC, the better the model is at predicting.

RESULT AND DISCUSSION

In order to perform the experimental implementation, authors have used i5 processor. The memory for the system is 8 GB RAM and 1.5TB GB. The different algorithms in the study are performed using python libraries like matplotlib, pandas, Dearborn, and numpy etc. The classification for training and testing data is 70:30. The experiment is conducted using Google Collaborator as it provides a suitable platform to use machine learning tasks like pre-processing, regression, classification, and clustering. The results obtained during experiment are mentioned.

The dataset has been analyzed in three different ways i.e., statistical, visual and comparative.

Statistical Analysis

Firstly, the numerical dataset is analyzed using different statistical tools as in the count, mean, standard deviation, minimum and maximum values, along with the quartile range.

TABLE 1. Calculative Analysis

	User Ratings	**Reviews**	**Price**	**Year**
Count	550.0	550.0	550.0	550.0
Mean	4.61	11953.28	13.10	2014
St. deviation	0.22	11731.13	10.84	3.16
Min	3.30	37.0	0.0	2009
25%	4.50	4058.0	7.0	2011
50%	4.70	8580.0	11.0	2014
75%	4.80	17253.25	16.0	2017
Max	4.90	87841.0	105.0	2019

Table 1. exhibits the parameters of the dataset, namely, user ratings, reviews, price and year. These parameters are analyzed on different statistical tools and their values are displayed in the table.

Visual Analysis

The visualization of data basically means transforming the dataset into visual images that represents various graphs and plots. The visual analysis of the data provides us an easier access and understanding of a huge dataset.

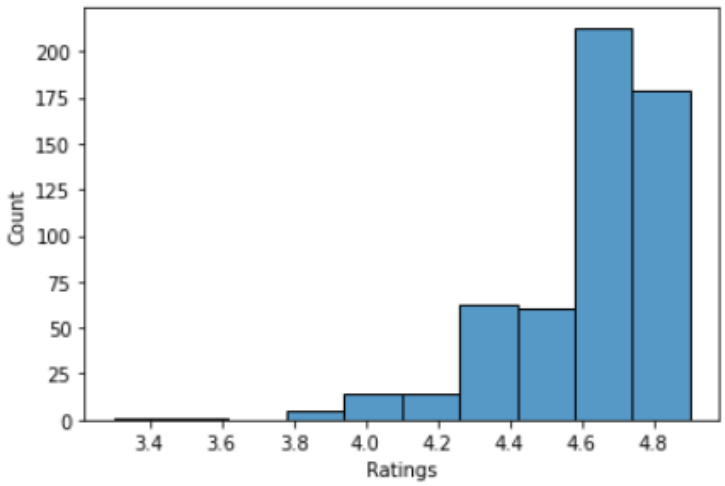

FIGURE 2. *Graphical representation of the ratings of books*

The numerical distribution of the book ratings is displayed in Figure 2. It displays the books with the highest and lowest ratings. Since, users tend to favor books with higher ratings over those with lower ratings, the graph provides them an easier access to that data. The maximum number of books in the dataset have a rating of 4.7 and the minimum is 3.8.

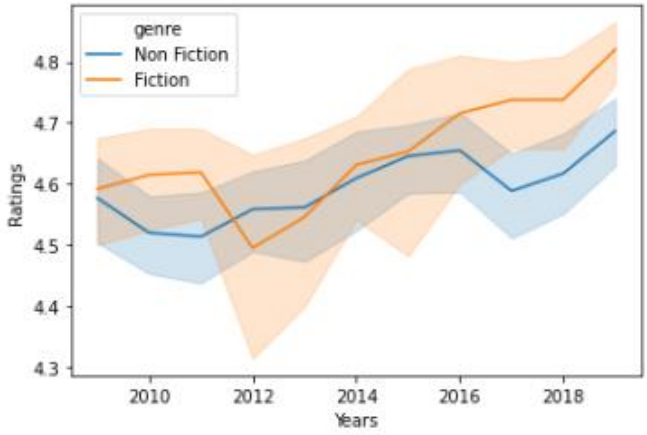

FIGURE 3. *Graphical representation of the ratings of books*

Figure 3. displays the relation of ratings with time (in years). It shows the ratings of two different genres i.e., Fiction and Non-fiction being compared with respect to the years (from 2009-2019). During the years 2014-2019, fiction books have a higher rating as compared to non-fiction books. Also, most of the books have a rating from 4.5 to 4.9

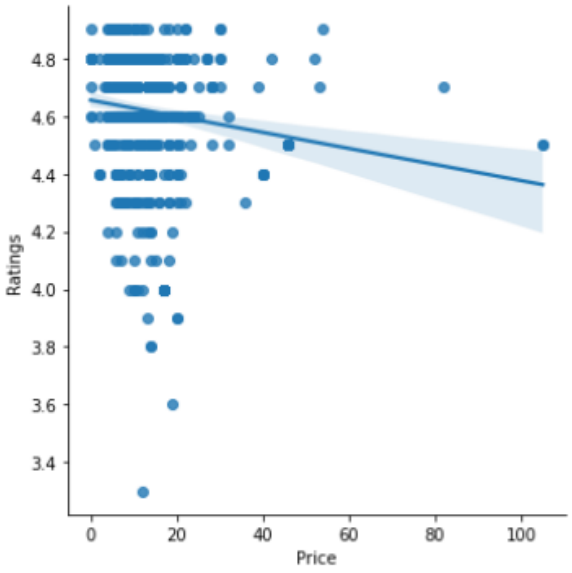

FIGURE 4. *Graphical representation of the ratings with price*

Figure 4. shows a significant relation between the price and ratings of the books. As the price of the books increase, the rating go down for both fiction and non-fiction books, which displays that the costumers usually prefer to read books that are less expensive.

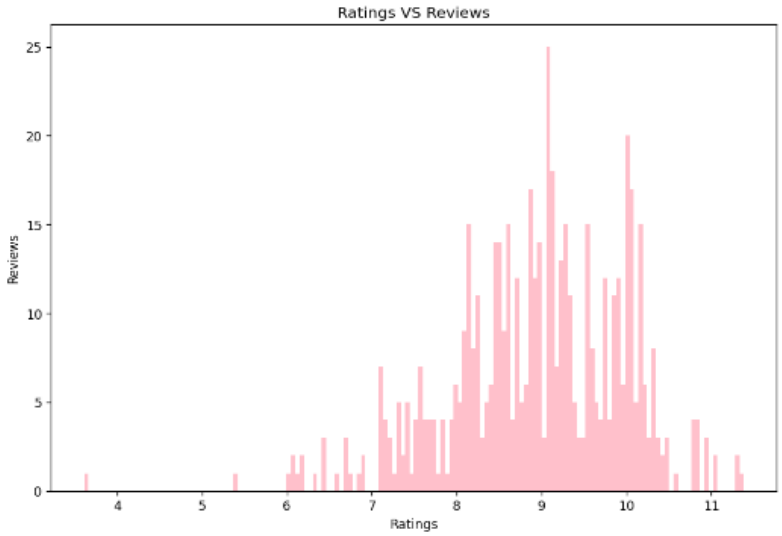

FIGURE 5. *Graphical representation of the ratings with reviews*

The reviews and ratings of the books are compared in fig 5. It displays that the reviews have a non-uniform impact on the user ratings. The books have maximum reviews when the user ratings are in interval of 8-10.

Comparative Analysis

For the comparative analysis of the dataset, two machine learning models namely, linear regression and logistic regression have been used.

Precision-recall curve and AUC-ROC curve are used to predict that which of the two machine learning models provide higher precision and accuracy.

TABLE 2. Value Table for Linear Regression

Algorithm	R Squared Value		Mean Squared Error		Mean Absolute Error	
Linear Regression	y_train	y_test	y_train	y_test	y_train	y_test
	0.99	-0.01	6.18	100.03	1.97	87.73

Table 2. is a value table for linear regression displaying the R squared value along with the mean squared and mean absolute errors, while training and testing the data.

TABLE 3. Precision-recall Table

Algorithm	Precision	Recall	F-1 Score
Logostic Regression	0.272	0.023	0.020

Table 3. is a value table for precision, recall, and F-1 score (accuracy) of the logistic regression model. The graph corresponding to this table is shown in figure 6.

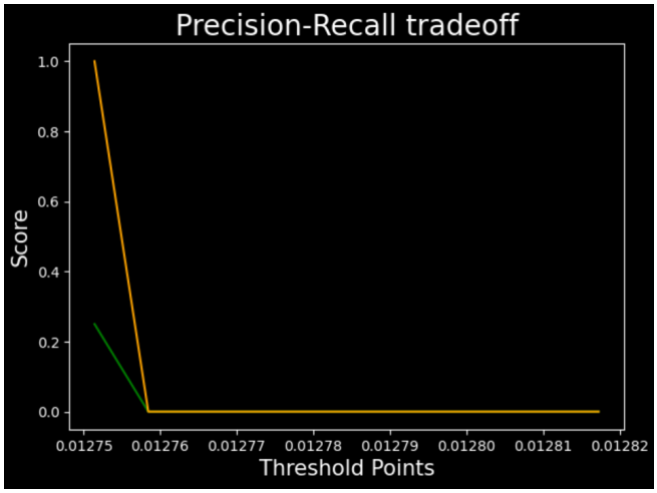

FIGURE 6. Precision-recall Curve

The relationship between recall and precision for each potential cut-off is known as a precision-recall curve. Fig 6. displays the precision-recall curve for the dataset.

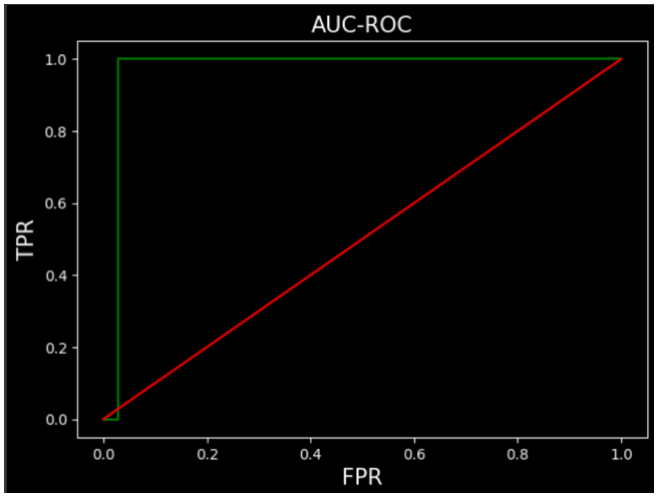

FIGURE 7. AUC-ROC Curve

Figure 7. shows the AUC-ROC curve for the dataset. In this graph, AUC measures a model's capacity to distinguish between classes and outcomes. A graph defined as a ROC curve illustrates how effectively a classification model works at all possible thresholds.

CONCLUSION

In this research work, authors have proposed implementation of two Machine learning models- Linear regression and logistic regression to determine the rating of books so that readers can directly access the books as per his interest without wasting his time in browsing through all the books. The similar work by different researchers is also discussed here. The performance of two machine learning models is compared using AUC-ROC and Precision-Recall. The experimental evaluation demonstrates that logistic regression outperforms linear regression in terms of accuracy and precision.

REFERENCES

1. Stien, D. and Beed, P.L., 2004. Bridging the gap between fiction and nonfiction in the literature circle setting. The Reading Teacher, 57(6), pp.510-518.
2. Rajagopalan, G., 2021. Data visualization with python libraries. In A Python Data Analyst's Toolkit (pp. 243-278). Apress, Berkeley, CA.
3. Mangla, M., Akhare, R., Deokar, S. and Mehta, V., 2020. Employing Machine Learning for Multi-perspective Emotional Health Analysis. In Emotion and Information Processing (pp. 199-211). Springer, Cham.
4. Mangla, M., Shinde, S.K., Mehta, V., Sharma, N. and Mohanty, S.N. eds., 2022. Handbook of Research on Machine Learning: Foundations and Applications. CRC Press.
5. Tanawongsuwan, P., 2015, June. Relation between a Book Review Content and Its Rating. In International Conference on Computer Information Systems and Industrial Applications (pp. 853-856). Atlantis Press.
6. Maghari, A.M., Al-Najjar, I.A. and Al-Laqtah, S.J., 2021. Books' Rating Prediction Using Just Neural Network.
7. Hu, N., Bose, I., Koh, N.S. and Liu, L., 2012. Manipulation of online reviews: An analysis of ratings, readability, and sentiments. Decision support systems, 52(3), pp.674-684.
8. Qu, L., Ifrim, G. and Weikum, G., 2010, August. The bag-of-opinions method for review rating prediction from sparse text patterns. In Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010) (pp. 913-921).
9. Sharma, N., Mangla, M., Yadav, S., Goyal, N., Singh, A., Verma, S. and Saber, T., 2021. A sequential ensemble model for photovoltaic power forecasting. Computers & Electrical Engineering, 96, p.107484.
10. https://www.kaggle.com/competitions/goodreads-books-reviews-290312/data
11. Verma, S., Sharma, N., Singh, A., Alharbi, A., Alosaimi, W., Alyami, H., ... & Goyal, N. (2022). An Intelligent Forecasting Model for Disease Prediction Using Stack Ensembling Approach. CMC-COMPUTERS MATERIALS & CONTINUA, 70(3), 6041-6055.
12. Mahajan, A., Sharma, N., Aparicio-Obregon, S., Alyami, H., Alharbi, A., Anand, D., ... & Goyal, N. (2022). A Novel Stacking-Based Deterministic Ensemble Model for Infectious Disease Prediction. Mathematics, 10(10), 1714.
13. Sharma, N., Mangla, M., Yadav, S., Goyal, N., Singh, A., Verma, S., & Saber, T. (2021). A sequential ensemble model for photovoltaic power forecasting. Computers & Electrical Engineering, 96, 107484.

14. Nidhi Goel, Samarjeet Kaur, Deepak Gunjan, S.J. Mahapatra, Investigating the significance of color space for abnormality detection in wireless capsule endoscopy images, Biomedical Signal Processing and Control, Volume 75, 2022, 103624, ISSN 1746-8094, https://doi.org/10.1016/j.bspc.2022.103624
15. Goel, N., Kaur, S., Gunjan, D. et al. Dilated CNN for abnormality detection in wireless capsule endoscopy images. Soft Comput 26, 1231–1247 (2022). https://doi.org/10.1007/s00500-021-06546-y
16. Handa, P., & Goel, N. (2021). Peri-ictal and non-seizure EEG event detection using generated metadata. Expert Systems, e12929. https://doi.org/10.1111/exsy.12929
17. S. Kaur and N. Goel, "A Dilated Convolutional Approach for Inflammatory Lesion Detection Using Multi-Scale Input Feature Fusion (Workshop Paper)," 2020 IEEE Sixth International Conference on Multimedia Big Data (BigMM), New Delhi, India, 2020, pp. 386-393, doi: 10.1109/BigMM50055.2020.00066.
18. Ruchika Bala, Arun Sharma and Nidhi Goel, A lightweight deep learning approach for diabetic retinopathy classification, Artificial Intelligence and Speech Technology. AIST 2021. Communications in Computer and Information Science, vol 1546, pp. 277-287, 12-13 November 2021 Springer, Cham. https://doi.org/10.1007/978-3-030-95711-7_25