Computers, Materials & Continua
DOI: 10.32604/cmc.2022.025858
Article

Tech Science Press

XGBRS Framework Integrated with Word2Vec Sentiment Analysis for Augmented Drug Recommendation

Shweta Paliwal[1], Amit Kumar Mishra[2,*], Ram Krishn Mishra[3], Nishad Nawaz[4] and M. Senthilkumar[5]

[1]MIET Meerut, Meerut, 250005, India
[2]School of Computing, DIT University, Dehradun, 248009, India
[3]Department of Computer Science, BITS Pilani, Dubai Campus, Dubai, 345055, United Arab Emirates
[4]Department of Business Management, College of Business Administration, Kingdom University,
Riffa, 40434, Kingdom of Bahrain
[5]School of Information Technology and Engineering, Vellore Institute of Technology, Vellore, 632014, India
*Corresponding Author: Amit Kumar Mishra Email: aec.amit@gmail.com
Received: 07 December 2021; Accepted: 16 February 2022

Abstract: Machine Learning is revolutionizing the era day by day and the scope is no more limited to computer science as the advancements are evident in the field of healthcare. Disease diagnosis, personalized medicine, and Recommendation system (RS) are among the promising applications that are using Machine Learning (ML) at a higher level. A recommendation system helps inefficient decision-making and suggests personalized recommendations accordingly. Today people share their experiences through reviews and hence designing of recommendation system based on users' sentiments is a challenge. The recommendation system has gained significant attention in different fields but considering healthcare, little is being done from the perspective of drugs, disease, and medical recommendations. This study is engrossed in designing a recommendation system that is based on the fusion of sentiment analysis and radiant boosting. The polarity of the sentiments is analyzed through user reviews and the processed data is fed into the Extreme Gradient Boosting (XGBOOST) framework to generate the drug recommendation. To establish the applicability of the concept a comparative study is performed between the proposed approach and the existing approaches.

Keywords: Recommendation system; Word2Vec; XGBOOST; sentiment analysis; natural language processing (NLP); machine learning

1 Introduction

With each day passing technology is transforming in different astonishing aspects; the ability to predict future outcomes based on past experiences has completely changed the outlook to view our surroundings. We are talking about none other than one of the most prominent emerging technologies of the era-Machine Learning and Deep Learning. These technologies have completely revolutionized the crucial field of healthcare by delivering excellent results. They are proficient in doing the chores that are performed by humans in a cost-effective and time-saving manner and hence became a part of our health ecosystem. In the current scenario Internet of medical things and Artificial Intelligence is already helping individuals through their virtual assistance, monitoring and detecting probable life-threatening diseases at earlier stages. Current disease diagnosis and treatment are entirely based on the doctor's knowledge and experiences. There is enormous availability of the data records related to patients, diseases, and concerned treatments but there are no systems that are capable of analyzing this data as well identifying patterns and associations between diseases and effective treatments. Thus, we are in a need to provide doctors with systems that can make predictions and share medical knowledge at earlier stages. The vast amount of clinical data has given rise to the need for recommendation systems. The recommendation system in healthcare is designed to generate a balance between the continuous generation of data and real-time response for the treatment. The integrated development of information technology with medical science has generated several research studies on different fields of healthcare including disease diagnosis, disease treatment, and drug discoveries. The arrival of Healthcare 4.0 has already accelerated the speed of digitization in the field of healthcare. Industry 4.0 is automating the service and production industries incredibly. The prominent technologies of Industry 4.0 such as Big data, Internet of Things (IoT), Cloud, and Fog Computing have transfigured the healthcare industry and directed its entire infrastructure towards Healthcare 4.0 [1]. The advent of Healthcare 4.0 is already observed in the e-health industry through the terminology of Smart Health. Smart Health is the acquisition of information and communication technologies but the scientific class of literature states different meanings for it ranging from "support to medical health through smartphones" or "sensor-based equipment for medical practice" but at the same time, one should not confuse the terms smart health and Healthcare 4.0. Healthcare 4.0 is based on three eminent technologies and is on its way to completely transforming the viewpoint towards e-health. The term Healthcare 4.0 comprises digital data foundations. Machine Learning, Deep Learning, and Natural Language Processing are responsible for learning digital knowledge and predictions; the second foundation contributes to digital knowledge representation and the third foundation is IoT which describes data collection through smart media, medical imaging, and smartphones. The field of medical science is significantly affected by the way clinical data is analyzed and stored through big data. Data mining and analysis help in the identification of illness along with the genetics and lifestyle that leads to certain prolonged diseases [2]. The paper further gives information on the available literature related to recommendation systems in Section 2, followed by the proposed methodology in Section 3, the evaluation of results in Section 4, comparative analysis in Section 5, and discussion and future scope in Section 6. The major findings of the study are as follows:

- Analysis of user reviews on the particular drug through identification of sentiment polarity.
- Feeding the processed reviews into the base classifiers and the proposed method.
- Recommendation of top drugs based on the integration of word2vec and XGBOOST.
- Comparison of the proposed methodology with the existing recommendation systems termed as a case study.

During the first phase, comparisons have been made between several base classifiers, and performance measures of accuracy, precision, recall, and f1-score are observed. The proposed methodology Extreme Gradient Boosting Recommendation System (XGBRS) suggested top drugs for a certain condition based on the sentiment of the user reviews with an accuracy measure of 0.96, precision measure as 0.92 & 0.89, recall measure as 0.91 & 0.89, and f1-score 0.94 & 0.89 for positive and negative class, apart from this each classifier is evaluated on two more metrics; mean average precision and coverage for clarity of the concept. To establish the applicability of the method, a comparative analysis has been drawn from the existing recommendation systems that are based on the framework of machine learning. The flow of the research paper is depicted in Fig. 1.

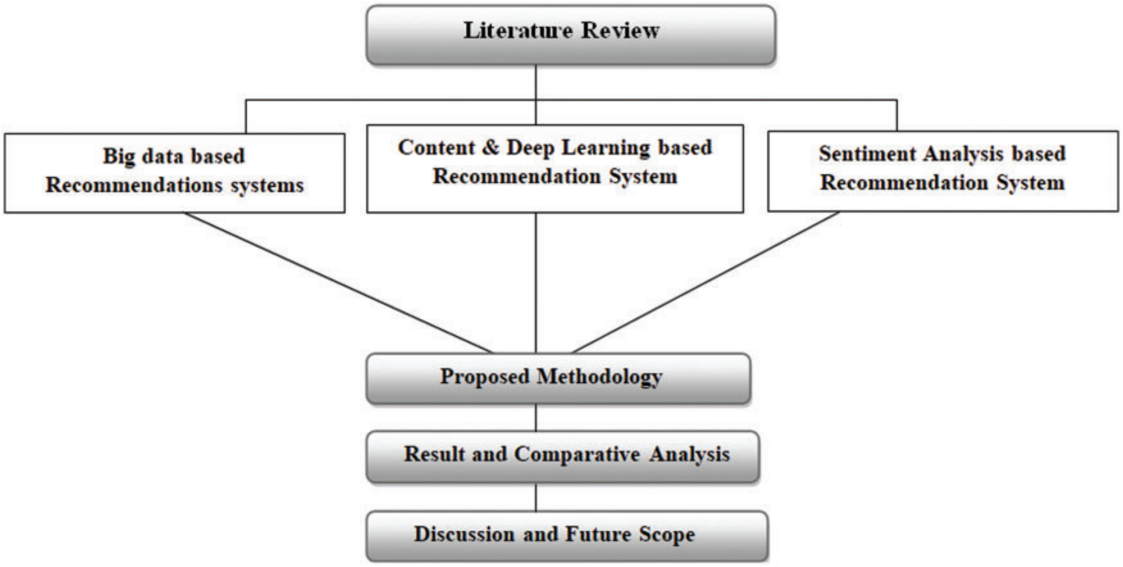

Literature Review
Big data based Recommendations systems
Content & Deep Learning based Recommendation System
Sentiment Analysis based Recommendation System
Proposed Methodology
Result and Comparative Analysis
Discussion and Future Scope

Figure 1: Paper's flowchart

2 Literature Review

The advent of healthcare 4.0 has opened up numerous ways in which recommendation systems can be designed for assisting healthcare. A recommendation system is a concept where predictions are made based on the past availability of data. The origin of the recommender system is traced back to the 1990s but the current advancement in Artificial Intelligence, Big Data, Cloud and Fog Computing, IoT, and NLP has made it possible to design systems that have high efficiency and accuracy for personalized healthcare. There are 4 basic categories of recommender systems; content-based, context-based, collaborative filtering, and hybrid recommender system [3]. Context is intelligence that differentiates the present circumstances from user-based systems. These systems are based on a set of contextual features that differs from situation to situation. The research studies have classified them further into categories namely: Location-based, Social based, Time based, Emotion-based, Activity-based. The definition of context is not particularly due to the versatility of the term in different fields. In a generalized manner, it can be said as the collection of facts for a particular surrounding. The context approaches are classified as pre-filtering approaches, post-filtering approaches, and contextual modeling approaches. The collection of contextual data plays an important role in the design process of recommendation systems. The cited procedures currently used so far are; tuple-based models, hierarchical models, graph-based models, and logic-based models [4]. The research gaps identified till today suggest that there are highly significant differences in the performance measures and the computational complexity that has created the need for better design of these systems. Content-based Filtering is the technique where a track of user's interest is kept and is further used to suggest the recommendations. This filtering technique is based upon the Term Frequency (TF) and Inverse Document Frequency (IDF) measure, where TF states the maximum number of times a word has occurred in the document and IDF states the maximum times a word has occurred in the corpus of documents. Collaborative filtering is one of the techniques that are currently being employed in the designing of the recommendation system. The concept states that if a user has an interest in a particular data item the same user may have a similar type of interest in other data items as well. The process begins with the pre-processing stage to make a user-item matrix, enumerate the same users and then generate the recommendation for the querying user. Studies and research has been carried out to recognize the diseases more accurately and utilize the knowledge, thus recommendation systems come under an active research area [5]. An ontology approach-based recommendation system integrated with fuzzy rules is used to identify the products of the user's interest. This RS is capable of identifying the content and determining the opinion of the review. The context of the review is extracted through the neuro-fuzzy rules [6]. A context-based recommendation system is designed for the insertion of objects using an unsupervised learning technique. The RS is built upon the modeling of joint probability distribution and the Gaussian model [7]. Geo-tagged data RS is designed for improving the context of social media, where the context of different Geo-tagged data is analyzed from the fields of tourism, traffic management, health monitoring, etc… The proposed work is an integration of 2 methods; context awareness and semantic analysis [8]. A nursing-based diagnosis system is proposed. The system utilizes a prefix-tree structure common in itemset mining to construct a ranked list of suggested care plan items based on previously-entered items [9]. A hybrid context-aware-based RS is designed for E-Health using evolutionary algorithm, Merkle-hash tree approach from the cloud is used to store the data and the hybrid RS is more efficient than other systems based on privacy preservation, recommendation, and computational complexities [10].

A content-based recommendation system using a neuro-fuzzy approach is designed to recommend processed items for the users. The researchers developed an Artificial Intelligence-based framework and a web application to give a glimpse of the user interface. This interface is designed to provide a simulation of real users and the proposed work is compared with a deep learning-based method [11]. A semantic health-based RS is designed that is responsible for complementing the health videos. The RS is capable of suggesting health-related websites per video by recommending links related to the videos. The algorithm implemented is capable of filtering out videos with relevant content [12]. A disease diagnosis and treatment recommendation system consist of two parts; the disease symptom analysis module is based on the clustering algorithm and the second part is responsible for treatment suggestions has been proposed [13]. Each disease leaves some adverse effects on the human brain but diabetes and abnormal blood pressure are among the most threatening diseases thus earlier classification of these diseases into suitable categories are required to opt for an effective treatment. The headway in information technology has digitalized the products used in healthcare, wearable sensors, and smartphones are the new trends that help in the collection of clinical data. Hence a novel healthcare system has been proposed to handle a large amount of data and give accurate predictions. The system is an integration of deep learning, data mining, cloud servers, and big data. The framework enhanced the performance of data processing and improved the accuracy of healthcare data classification [14]. A deep learning-based recommendation system for heart disease diagnosis has been proposed which is a combination of multiple kernel learning and neuro-fuzzy inference system. The kernel learning method is responsible for the division of parameters between patients suffering from heart disease and healthy individuals. Obtained results are fed into the neuro-fuzzy system and specificity, sensitivity, and mean squared error is used as performance evaluators [15]. A fuzzy type-2 ontology-based recommendation system came into existence for supporting IOT healthcare and at the same time providing assistance in monitoring the patient's body and suggesting specific drugs and food. The data is gathered with the help of wearable sensors and semantic web rule language rules and fuzzy logics are integrated to automate the recommendation process [16]. The data-driven research with supervised learning and reinforcement learning has been used for designing the medical recommendation system that is assisting medical professionals in making better clinical decisions. A framework has been proposed based on Supervised Reinforcement learning and recurrent neural network to knob complex relationships among multiple medications and diseases. Experimental observations are carried out on a publicly available dataset [17]. IoT-based health monitoring system is designed and integrated with cloud computing for effective recommendations using historic and empirical data. The monitoring system has incorporated Random Forest that delivered an accuracy of 97.26% on the dermatology dataset [18]. Another significant term that is currently trending is *personalized healthcare; a patient's clinical history is taken into consideration for suggesting the desired treatment and medication*. A personalized healthcare recommendation system is designed for diabetic patients. The recommendation is performed based on the patient's bio-cultural ontology. Rule-based knowledge is integrated with the knowledge base. The system is then verified by developing use cases and expert verification [19]. A recommendation system for personal health and well-being has been designed using methods of hybrid learning. The system used K-Nearest Neighbor (KNN) categorization and give suggestions assigned to only the closest neighbor. Restricted Boltzmann Machine (RBM) is combined with Convolution Neural Network for the designing of an intelligent health recommender system. The system is named DEEP RECO and uses collaborative filtering. Root mean square and Mean squared error are considered as the performance measures. Health recommendation systems are gaining significant importance in acquiring supplementary information to support clinical decisions [20].

One of the innovative researches that have been performed is designing a recommender system to quit smoking through the usage of motivation messages. The research uses a mobile application that transmits messages through the recommender system that considers users' opinions and user's profiles [21]. A stacked discriminative auto encoder recommendation system is designed. The system provides security and suggests top recommendations based upon the analysis of the user's preferences. The security of the system is ensured with the blowfish algorithm. To provide a structure to the collected data Hadoop framework is used and it has been observed that this system performed better than others [22]. A new algorithm KNN based recommendation system is designed. The system is based on analysis of the patterns of diseases with patterns in the human body, which was then implemented in Healthcare 4.0 for the recommendation of diagnosis and treatment [23]. A social network marketing recommendation system is designed using big data analytics. The database is divided into five sections and is analyzed through the clustering and association rule approach. The study revealed the knowledge on behavior through a rule-based recommendation system and generated recommendations for personalized social network marketing [24]. A product-based recommendation system using ensemble technique is designed that has improved the accuracy of suggesting products that can be frequently brought together in the period of COVID-19. The model has generated the least error rate and has shown significant results as compared to the traditional approaches [25]. Another enhanced recommendation system based on extreme boosting is designed to improve the sentiment analysis in social media posts. The model proved that the proposed method outperformed all the existing classifiers and improved the sentiment classification [26]. Ensemble learning-based hotel recommendation system was designed that was based on the ensemble of Bidirectional Encoder Representation from Transformers (BERT) technique and Random Forest. In this study sentiment analysis of the textual data is performed and it has been observed that the recommendation accuracy was enhanced to a significant level along with a prominent F1-score measure as compared to others [27]. A smart healthcare recommendation system is designed for patients suffering from diabetes. The recommendation system is based on ensemble and deep fusion techniques and has achieved maximum accuracy for the prediction of multidisciplinary diabetes disease and suggests recommendations accordingly [28]. A sentiment analysis-based recommendation system is designed using deep neural networks. The sentiment of stressed and depressed users is analyzed and recurrent neural networks are used for further processing. The proposed method has reached an accuracy of 0.89 and 0.90 for both types of users [29]. The sentiment recommendation system is designed for healthcare. The aim was to satisfy the user's psychological preferences [30]. A fuzzy-based recommendation system is designed using sentiment analysis and ontology. The proposed methodology revealed better performance as compared to the existing product recommendation system [31]. After performing the literature study, it has been observed that although recommendation systems are emerging as a hot research field but still not much attention is given to the development of recommendation systems related to medical science. In today's scenario, enormous data is generated through people's reviews that have significant importance as medical data and thus need to be processed. Moreover, it has also been observed that ML frameworks deliver significant results thus the study is focused on designing an efficient framework built on the structure of sentiment analysis to suggest a drug (medicine) for severe medical conditions.

3 Proposed Methodology

The evolution in the field of computer-based technologies has significantly increased the volume of user-generated data in the forms of text over different platforms. This large volume of data is still not processed completely through the techniques of NLP. In the field of medical science, textual information represents the interaction between the patient and medical professionals along with the treatments suggested by the professionals. They also provide insights into people's emotions and reactions to these real-time situations. In this study, we are describing a method that is based upon the analysis of sentiments of drug review. The proposed architecture is depicted in Fig. 2. The sentiments are classified according to the polarity as positive and negative sentiments. For improving the performance of the classifiers, XGBOOST is implemented to improve the performance of weak classifiers by combining them and thus using a gradient boosting framework. XGBOOST has yielded a higher performance as compared to other algorithms discussed above because of its ability to handle regularization to avoid overfitting and bias of the classifier. The drug recommendation system is predicting the drugs for a certain condition based on user reviews. In the study, the stochastic gradient descent method has been used in XGBOOST. A single sample is selected for each iteration and the goal is to minimize the cost function or loss function. The parameter of learning rate is flexible, hence it is kept smaller as if the learning rate is large in number, there is a possibility that the algorithm might jump across leaving the minimal point. Algorithm 1 describes the stochastic gradient descent method. The basic function is represented by Eq. (1) below. XGBOOST calculates the loss function, then trains the model, adds the model to the ensemble, and then makes the prediction. The training data (A_i) and its associated labels (L_i) the classifier utilizes base classifiers for predicting the outcome (O_i) given by Eq. (2). The algorithm classifies the parameters into three types; general, boosting, and learning task parameters. For the study, the general parameters (booster, no of threads) are set to 1 whereas boosting parameters are tuned ('eta' is tuned at each step to reduce the step size, 'depth of the tree' is tuned between the values (5–8) and high values are neglected to avoid overfitting of the generated trees, 'gamma' which specifies the loss function is also tuned between the values 0.05–0.5).

$$F(X) = \frac{1}{n}\sum_{i=1}^{n} f_i(x) \tag{1}$$

$$O_i = \sum_{n=1}^{N} f_n(A_i) \tag{2}$$

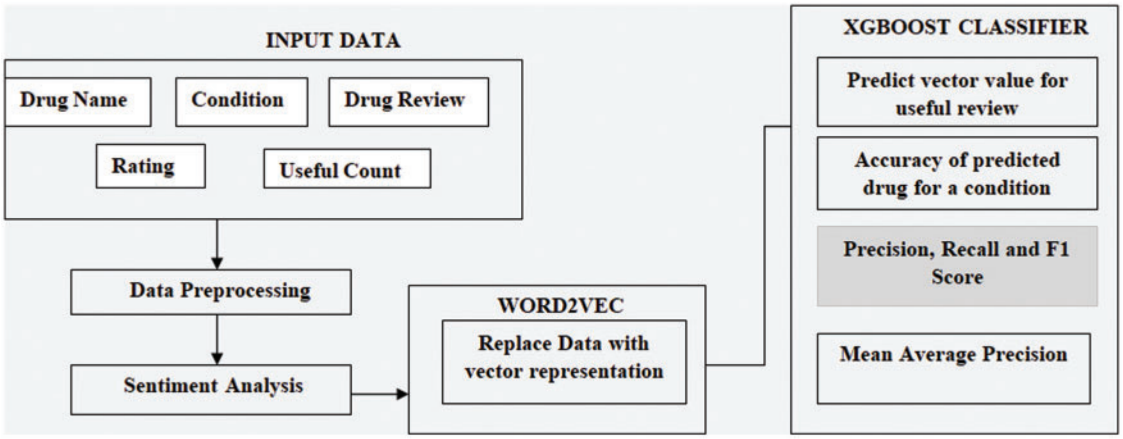

INPUT DATA
Drug Name
Condition
Drug Review
Rating
Useful Count
Data Preprocessing
Sentiment Analysis
WORD2VEC
Replace Data with vector representation
XGBOOST CLASSIFIER
Predict vector value for useful review
Accuracy of predicted drug for a condition
Precision, Recall and F1 Score
Mean Average Precision

Figure 2: Proposed architecture

In the algorithm, the similarity score is calculated by calculating the statistical measure of gain to find the best split, and the new prediction is calculated using [32]; Let (i) represents each example in our dataset, XGBOOST is based upon the calculation of loss function that is the ultimate objective of the algorithm is to minimize the objective function. Gain = Left similarity + Right similarity-Root similarity, Initial predicted value + learning rate (eta) × output value, and the equations are represented by Eqs. (3) and (4). The first part of the equation is responsible for calculating the residuals of the predicted value represented by (y_i) for each particular leaf, the omega represents the regularization term which is responsible for reducing the insensitivity. Now the initial value is kept as y_0 with a hat, the new prediction is represented through (i-1) prediction plus the output value from the $i^{(th)}$ tree. To approximate the value of loss function for the learner the algorithm uses Taylor's approximation of the second derivative. Eq. (3) presents the objective function that is subjected to minimization, now subjecting the function for the prediction of new learners using Taylor's approximation (second-order derivative) given by Eq. (4). The symbol α signifies the prediction and $(x - \alpha)$ is the new learner.

$$\zeta^{(t)} = \sum_{i=1}^{n} \iota(y_{i,}\widehat{y_i} + f_t\,(x_i)) + \Omega\,(f_t) \tag{3}$$

$$f\,(x) \approx f\,(\alpha) + f'(\alpha)\,(x - \alpha) + \frac{1}{2} f''(\alpha)(x - \alpha)^2 \tag{4}$$

Now calculating the similarity score, XGBOOST multiplies the equation by −1 to transform the parabola over the horizontal axis.

Algorithm 1: STOCHASTIC GRADIENT DESCENT

```
1) Define a function for representing stochastic gradient descent
2) Compute the gradient of the function:
3) Set the gradient parameter as:
      batch_size=1
      learn_rate=0.1
```

(Continued)

Algorithm 1: Continued

```
4) upgrade the gradient function
5) Initialize a random number generator
        If learn_rate<=0
        Raise ValueError
6) calculate the step size for each attribute
        step size=gradient*learn_rate
```

Word2Vec is one of the promising techniques of natural language processing developed by Google. The technique uses shallow neural networks for the creation of embedding and employs neural networks in both of its two methods namely; Skip Gram and Common Bag of Words. The training and testing split for the dataset is kept to be in a ratio of 75% and 25%. Word2Vec used the cosine similarity which means that the angle between the 2 vectors should be closed to one. The skip-gram architecture is used in the study as it predicts the source context for a given center word. It calculates the probability of each word appearing without considering its distance from the center point. It describes two-dimensional word vectors (x,1). The vectorized softmax function is used for modeling the discrete probability distribution. Softmax is a version of the argmax function, the values are scaled so that each probability sum up to 1.0. The logic of stochastic gradient descent is given by algorithm 1 and algorithm 2 describes XGBOOST Classifier.

$$\text{Probability} = \exp(\text{value})/\text{sum v in list} \exp(v)$$

$$\text{Probability} = \exp(1)/(\exp(1) + \exp(3) + \exp(2))$$

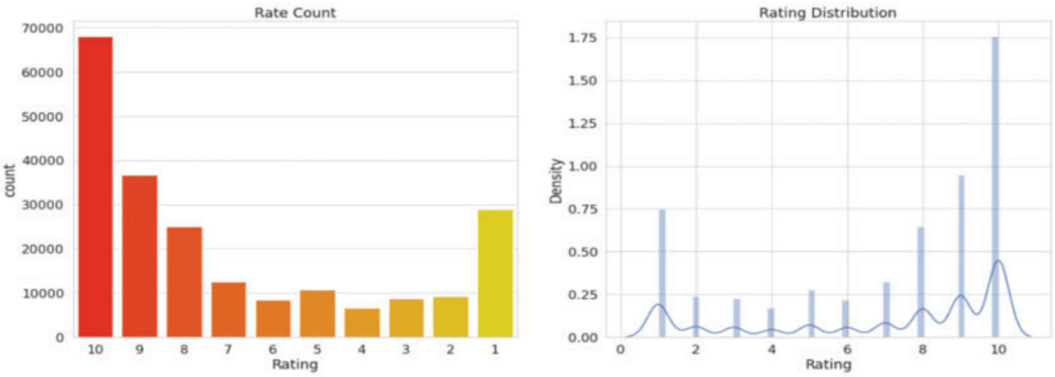
$$\text{Probability} = \exp(1)/(\exp(1) + \exp(3) + \exp(2))$$

Algorithm 2: XGBOOST CLASSIFIER

```
1)  Procedure:
2)  Initiate set of features
3)  For each trained classifier($\boldsymbol{x}_i$)do:
4)  Prediction is done for the input using a set of features
5)  Combine all predictions and make predictions using XGBOOST Classifier
6)  Return result
```

a Data Cleaning and pre-processing

A publicly available data set on the UCI Repository is considered for the study. The dataset contains multiple records related to drug reviews that are important due to various reasons such as drug reviews allow people to get to know about the effects of the drug they are using by other individuals and also it gives insights on the side effects and positive results about a particular drug. Total data points count to 161297 records. The data provides reviews given by the users on a particular drug based on a certain condition and a 10-star rating related to the overall satisfaction of the user. The python libraries ‘pandas’ are used for the process of data cleaning, ‘numpy’ for mathematical functions, ‘matplotlib’ for data visualization, and ‘scikit-learn’ for the process of sentiment analysis. After the step of checking for duplicate values, missing values, and outliers the data is then preprocessed for reducing the dimensionality of data. Tokenization is performed for breaking the textual information into a meaningful list of words and phrases. Secondly, words that are occurring multiple times in reviews such as “happen”, “happening” are converted into a standard format. Moreover, for efficient analysis, conditions with only one drug are removed from the dataset. Figs. 3a and 3b describe the rating distribution in the dataset.



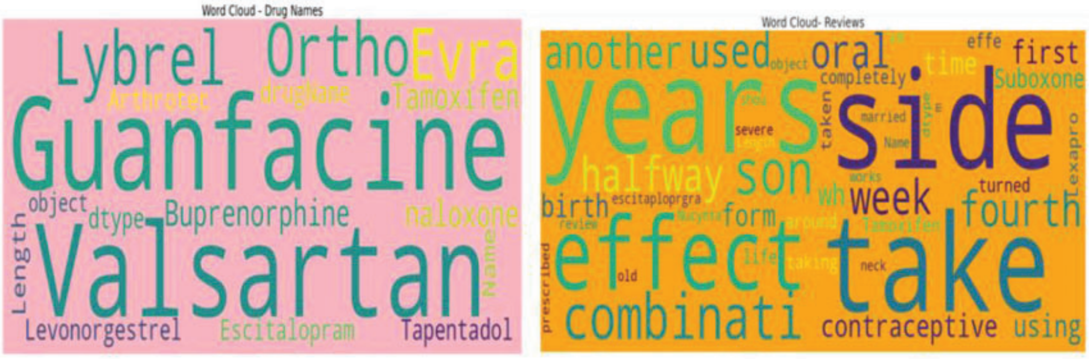
(a) Rating Vs Rate Count

(b) Rating Vs Rating Distribution

Figure 3: Distribution of ratings. (a) Rating *vs.* Rate Count (b) Rating *vs.* Rating Distribution

b Sentiment Analysis

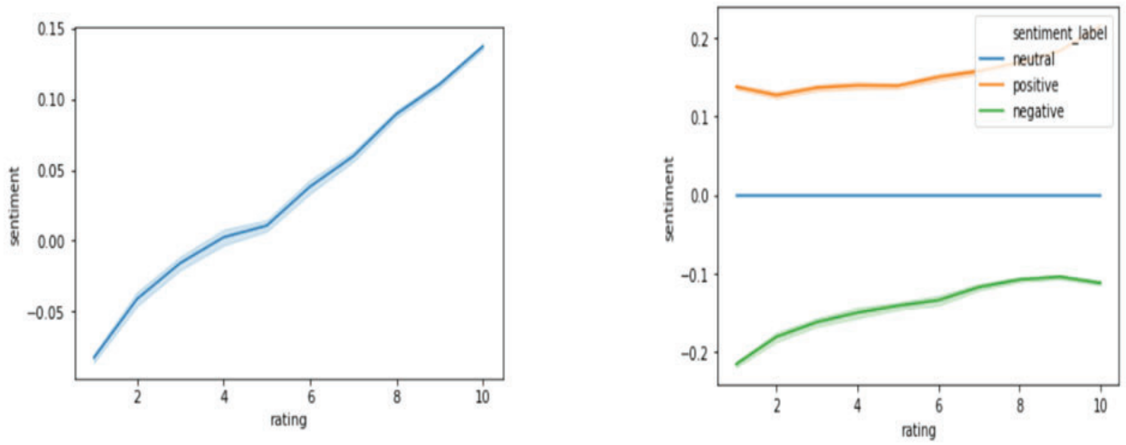
Sentiment's analysis of data on the experience of a drug is challenging research that is gaining attention. Sentiment analysis identifies terms of significance importance from a large volume of data through automation. Sentiment analysis is a subfield of data mining and is built up of multiple processes. In the study, the data is converted into feature vector representation using Word2vec and is followed by classification through machine learning classifiers. Since algorithms of machine learning cannot be directly applied to the textual form of data so we need to convert it into a numerical format. Thus, to perform the classification task we have applied vectorization of words using the Word2Vec technique. Each position is defined as p and center word at that position as ***c*** and the associated context word as ***w***, for identifying context words, a window of size m is defined which means our model will look at words in position $p - m \; to \; p + m$ as the context. After we have all context words at position p, maximize the likelihood of the context words given the center word, by calculating the probability of our model predicting the context words given the center word. Figs. 4 and 5 visualize the word cloud and sentiments polarity.



(a) Drug Names

(b) Reviews

Figure 4: Word Cloud Visuals. (a) Drug Names (b) Reviews



(a) Rating vs Sentiment(Neutral) **(b) Sentiment vs Rating(Positive, Neutral, Negative)**

Figure 5: Sentiment Visualization. (a) Rating *vs.* Sentiment(Neutral) (b) Sentiment *vs.* Rating (Positive, Neutral, Negative)

Algorithm 3: LOGIC FOR FEATURE ENGINEERING

```
Procedure:
Load Dataset as data
Initialize variable v='rating'
If v>=5
    Review sentiment=1
Else
    Review sentiment=0
Data['Review sentiment'].valuecount()
```

Algorithm 4: SENTIMENT POLARITY

```
Procedure:
define sentiment(review)
polarity=[]
for i in review:
    apply Textblob
append polarity
return polarity
```

c Base Classifiers

a) Support Vector Machine **(*SVM*):** SVM follows the ideology of opting for a hyperplane that best separates the data points by their class (either 0 or 1). SVM is implemented through kernels; kernels are a sort of mathematical function that manipulates the data. SVM by default takes the data into numeric form as an input variable, thus if you have a categorical variable you need to convert it into a numerical format. Hence in the study, we have employed a label encoder to convert the categorical variable into the numeric variable given by Eq. (5)

$$A0 + (A1 \times X1) + (A2 \times X2) = 0 \tag{5}$$

Here X1 and X2 are the input variables for determining a slope of a line and A0 is the intercept is searched by the learning algorithm. This equation is responsible for the classification of new data points. SVM supports different kernel functions; hence in the study we have employed two of them; linear kernel and radial kernel. Linear kernel describes the measure of distance between a new data point and support vectors. A linear kernel is used in the study as they provide more accurate classifiers. The next kernel used is the radial kernel which is based on gamma parameters. The parameter is specified by the learning function given by Eq. (6).

$$K\left(y, y\ddot{\imath}\right) = \Sigma\left(y \times y\ddot{\imath}\right) \tag{6}$$

***b Multinomial Naive Bayes* (*MNB*):** It comes under probabilistic learning techniques and uses Bayes theorem for making predictions. MNB incorporates a hyperparameter alpha which is responsible for controlling the model. MNB is employed in the study as it works well for text classification. Multinomial Naive Bayes is a generative classifier in which the parameter alpha is kept at one and the parameter beta is tuned.

***c Random Forest*:** Random Forest initially constructs multiple trees and final prediction is made after combining the prediction from different generated trees. For problems related to classification, random forest calculates the mode of the classes. In the study, the Scikit-Learn library is employed using the Gini importance measure for building the decision trees. Random Forest incorporates several steps out of which few basic important ones are mentioned below:

- *Select (N) number of features from the dataset*
- *For each x in N*
    a) *Calculate entropy and information gain*
    b) *Select the node with the highest information gain*
    c) *Split it out into its sub-nodes*
    d) *Repeat the steps until the terminate condition is reached*
    e) *Repeat the above steps (n) several times for the construction of trees*

d) K Nearest Neighbor: It is a supervised learning technique that is based upon similarity measures. The new data point is sent to the nearest neighbor based on the similarity measure which is generally the distance (Euclidean or Manhattan) between the two data points. We have observed that KNN has not yielded so good results on our dataset and this may be attributed to the large size of data.

***e) Linear Regression*:** Linear Regression models the input variable(x) and output variable(y) through a linear relationship. It uses coefficients to fit the linear model and minimizes the residual sum of squares. Linear Regression is of 2 types; simple and multiple. Simple linear regression searches for a statistic relationship where the main idea is to find a line that best fits the data point.

d Performance Measures

For study following classifiers are incorporated into the study; SVM (both with RBF and Linear Kernel), Multinomial Naive Bayes, Random Forest, K-Nearest Neighbor, Linear Regression and XGBOOST. There are a large number of records available in this dataset and hence due to this the primary objection of reduced time has also been taken into consideration. For the analysis of the predicted sentiments main metrics have been included namely; precision, recall, accuracy and F1-score. Let TP stands for true positive, FP for false positive, TN for true negative, and FN for false negative. The equation for performance measures is given by Eqs. (7)–(10).

- *Precision*: It is stated as the number of true positives divided by the total number of true positives and false positives

$$P = \frac{TP}{TP + FP} \tag{7}$$

- *Recall:* It is stated as the number of true positive over the number of true positive and false negative.

$$R = \frac{TP}{TP + FN} \tag{8}$$

- *F1 Score*: The weighted average between precision and recall is termed ad F1-Score.

$$F1 = 2\frac{P * R}{P + R} \tag{9}$$

- *Accuracy*: It is the percentage of correct predictions for the data given.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

Another important metric that has been taken into consideration is Mean Average Precision which is a performance measure for recommendation systems. So here in the study k number of recommendations has been generated in the following pattern; for example; if the system has generated 5 recommendations in one iteration out of which only three are relevant then the relevant recommendations are denoted by (1) and others are denoted by (0). Hence the order is (10101). Thus, with each classifier, the recommendations are generated and the mean average precision is calculated given by Eq. (11). The study has also tried to explore the parameter coverage for each classifier. Coverage is stated as the percentage of things present in the training data the model suggests on test data. Fig. 6 gives the graphical representation of the model's performance.

$$Mean\ Average\ Precision = \frac{1}{m}\sum_{k=1}^{N} P\left(k\right).rel\left(k\right) \tag{11}$$

After the study it has been noted down that in some cases accuracy and precision can't be considered as an ideal performance measure and hence to establish the concept of the proposed approach more mathematically we have calculated the measure Matthews Correlation Coefficient (MCC). The measure was originally proposed for the chemical structures and back in 2000 is considered among the standard performance measures in the subject of machine learning. The significant difference between the metric F1-score and MCC is that MCC remains invariant even if we shuffle the positive label with a negative label and vice-versa. Thus it calculated the Pearson product-moment correlation coefficient between the actual values and the values that are predicted and are shown in Eq. (12).

$$MCC = \frac{TP.TN - FP.FN}{\sqrt{\left(TP + FP\right).\left(TP + FN\right).\left(TN + FP\right),\left(TN + FN\right)}} \tag{12}$$

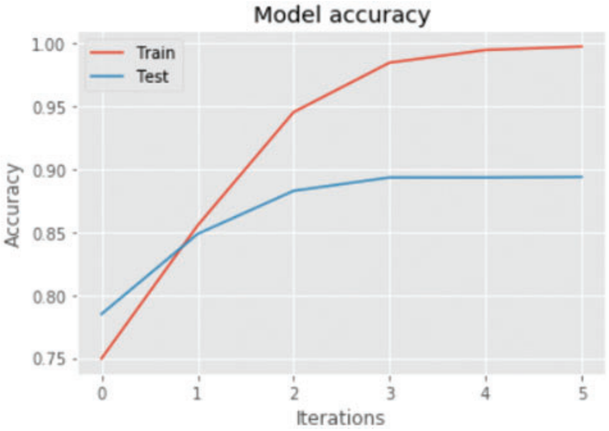


Figure 6: Proposed method-accuracy

4 Result

The best-predicted measures are identified and combined for the production of prediction, these results are then multiplied to generate an overall score for a drug related to a particular condition, the score is generated using a normalized count i.e. higher the count, the better the medicine. The top conditions identified during the study are; ***birth control, depression, anxiety, acne, bipolar disorders***. The reviews are classified as positive and negative sentiments based on the ratings, ratings above five are under the positive sentiments, and ratings below five are under the negative sentiments. The normalization factor is considered because there is a significant deviation between the least and most extreme. The proposed method of the XGBOOST classifier has released the best performance for the measure of accuracy, precision, recall, and F1-score. Moreover, the model is neither overfitted nor underfit in nature and has been cross-validated during the model building process. XGBOOST has inbuilt cross-validation and is capable of handling missing values. Tab. 1 represents the observations drawn for each classifier and Tab. 2 represents the mean average precision and the Matthews correlation coefficient while Tab. 3 represents the top drug for a particular condition.

Table 1: Observations

Algorithm	Sentiment	Precision	Recall	F1 score	Accuracy
Support Vector Machine(L)	Positive	0.85	0.57	0.45	0.80
	Negative	0.71	0.69	0.60	
Support Vector Machine (RBF)	Positive	0.75	0.89	0.71	0.87
	Negative	0.65	0.67	0.55	
Multinomial Naive Bayes	Positive	0.78	0.80	0.82	0.79
	Negative	0.60	0.71	0.71	
Random Forest	Positive	0.77	0.52	0.64	0.87
	Negative	0.62	0.79	0.72	
K- Nearest Neighbor	Positive	0.56	0.63	0.57	0.69
	Negative	0.69	0.50	0.59	
Linear Regression	Positive	0.54	0.65	0.68	0.71

(Continued)

Table 1: Continued

Algorithm	Sentiment	Precision	Recall	F1 score	Accuracy
	Negative	0.47	0.59	0.51	
XGBOOST	**Positive**	**0.92**	**0.91**	**0.94**	**0.96**
	Negative	**0.89**	**0.89**	**0.89**	

Table 2: Mean average precision for each classifier

Classifier	Mean average precision	Matthews correlation coefficient
Support Vector Machine (L)	0.72	0.82
Support Vector Machine (RBF)	0.83	0.82
Multinomial Naive Bayes	0.69	0.79
Random Forest	0.88	0.90
K- Nearest Neighbor	0.54	0.78
Linear Regression	0.71	0.78
XGBOOST	0.92	0.90

Table 3: Glimpse of top drugs for a condition

Condition	Recommended drug
Birth control	*Etonogestrel*
Birth control	Norethindrone
Birth control	Levonorgestrel
Depression	Buplopion
Depression	Sertraline
Depression	Venlafaxine

5 Comparative Analysis

In this section, we have explored some of the studies that are applying the machine learning framework to provide recommendations. The case studies are compared with the proposed method to establish the applicability of the method in recommending drugs for a particular condition after considering the polarity of the sentiments identified through the reviews. Extensive study is carried out and only those studies are considered for comparison which is employing the techniques of machine learning in the recommendation. For each of the studies, the input is the value of the dataset that is fed to the recommendation system, and based on the input provided the system has performed further evaluation.

- **Study 1 (Hossain et al., 2021** [33]**):** The following research has been carried out to design a drug-based recommendation system. Different machine learning classifiers such as; Decision tree, KNN, and Linear support vector machine were used. After the experimental study, they have concluded that linear SVM has yielded better performance in generating the recommendations. The recorded observations for precision, recall, f1-measure were; 0.79%, 0.86%, 0.82% and for accuracy it is 83%. The proposed method based on XGBOOST has yielded an accuracy measure of 0.96 whereas precision is 0.92 and 0.89 for positive and negative classes.

- **Study 2 (Garg, 2021 [34]):** The research is focused on different vectorization techniques for designing a medicine recommendation system, techniques such as word2vec, Bag of Words, and TF-IDF have been used. The experimental study concluded that the TF-IDF-based recommendation system has generated an accuracy of 93% in suggesting medicinal recommendations.
- **Study 3 (Alwateer et al., 2020 [35]):** The research has designed a recommendation system based on the pharmaceutical industry which is an integration of blockchain with machine learning. The Light Gradient Boosting model is used to suggest the top medicinal recommendations to the customers of the pharmaceutical industry. The author conferred that the results were efficient and prominent.

6 Discussion and Future Scope

Recognizing a drug that left no side effects is a challenging task in medical healthcare thus recommendation system is becoming an active area of research that is improving with time. In this study, we have enhanced drug recommendations using the XGBOOST technique based on the sentiment analysis of the collected reviews. It has been observed that XGBOOST has delivered significant results as compared to other existing techniques of recommending drugs as far as the author's knowledge and survey. Moreover, we have also evaluated the proposed method on the performance measure mean average precision and coverage for better clarity. These results are achieved by the tuning of the parameters of the algorithm and its nature of regularization. The system recommends the top drug for a certain condition based on the sentiment polarity of the reviews given by the patients. This system will be enhanced more to predict the drugs for the real-time environment. The future objective also includes reduced time complexity and integration of deep learning framework into the proposed XGBOOST classification technique. Also, the focus will be laid upon the extension of the system to multiple class analysis of diseases and suggesting the required diagnosis and drugs based on the severity of the disease. Thus, the proposed system will evolve with time to make more intelligent decisions and could be later on used on to the recommendation of personalized medication. Figs. 7–10 describe the visualization for the performance measure accuracy, mean average precision, recall, precision, and coverage metric for different classifiers.

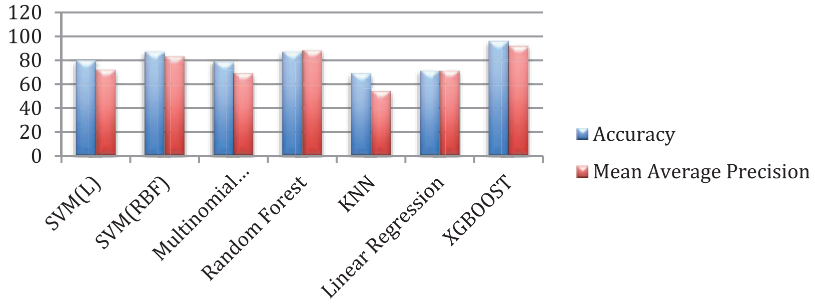


Figure 7: Distribution of accuracy and mean average precision

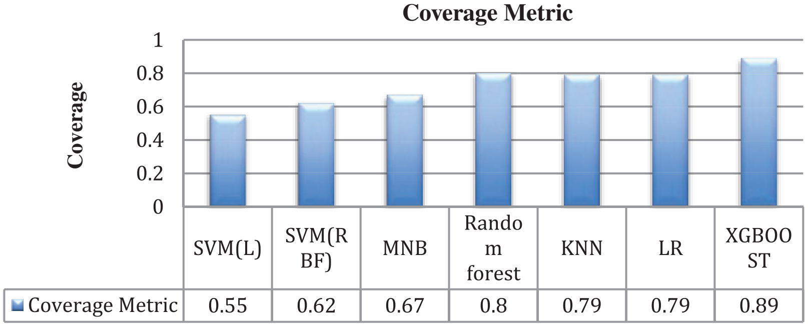


Figure 8: Distribution of coverage metric

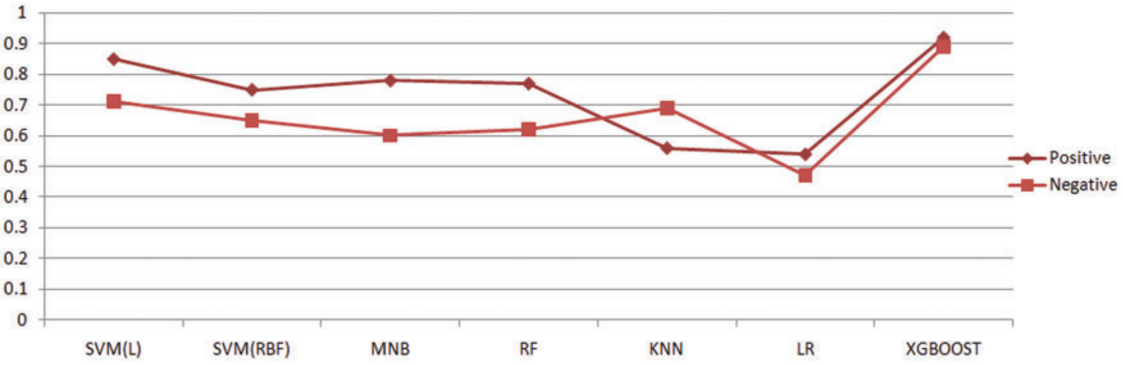


Figure 9: Distribution of precision

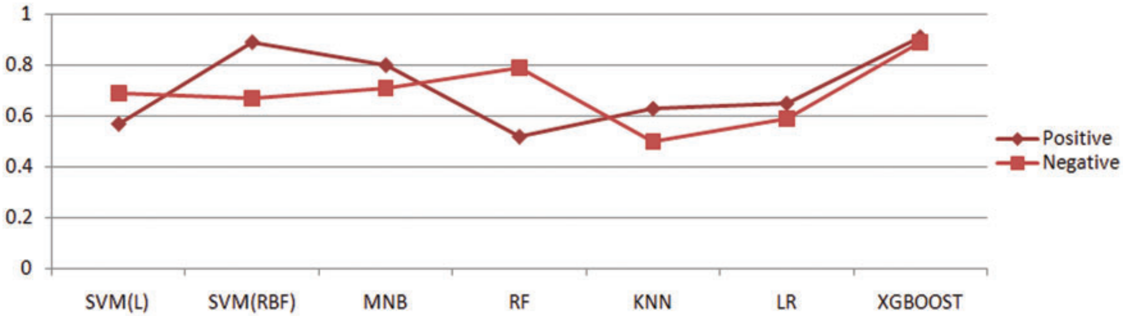


Figure 10: Distribution of recall

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] M. Wehde, "Healthcare 4.0," *IEEE Engineering Management Review*, vol. 47, no. 3, pp. 24–28, 2019.

[2] G. L. Tortorella, F. S. Fogliatto, M. C. Vergara, R. Vassolo and R. Sawhney, "Healthcare 4.0: Trends, challenges and research directions," *Production Planning & Control*, vol. 31, no. 15, pp. 1245–1260, 2020.

[3] W. T. Chu and Y. L. Tsai, "A hybrid recommendation system considering visual information for predicting favorite restaurants," *World Wide Web,* vol. 20, no. 6, pp. 1313–1331, 2017.

[4] P. B. Thorat, R. M. Goudar and S. Barve, "Survey on collaborative filtering, content-based filtering and hybrid recommendation system," *International Journal of Computer Applications*, vol. 110, no. 4, pp. 31–36, 2015.

[5] N. A. Albatayneh, K. I. Ghauth and F. F. Chua, "Utilizing learner's negative ratings in semantic content-based recommender system for e-learning forum," *Journal of Educational Technology & Society*, vol. 21, no. 1, pp. 112–125, 2018.

[6] A. Razia, Sulthana and S. Ramasamy, "Ontology and context based recommendation system using Neuro-Fuzzy classification," *Computers & Electrical Engineering*, vol. 74, no. 2, pp. 498–510, 2019.

[7] S. H. Zhang, Z. P. Zhou, B. Liu, X. Dong and P. Hall, "What and where: A context-based recommendation system for object insertion," *Computational Visual Media*, vol. 6, no. 1, pp. 79–93, 2020.

[8] Q. Yang, "A novel recommendation system based on semantics and context awareness," *Computing*, vol. 100, no. 8, pp. 809–823, 2018.

[9] L. Duan, W. N. Street and E. Xu, "Healthcare information systems: data mining methods in the creation of a clinical recommender system," *Enterprise Information Systems*, vol. 5, no. 2, pp. 169–181, 2011.

[10] N. Deepa and P. J. S. C. Pandiaraja, "Hybrid context aware recommendation system for E-Health care by merkle hash tree from cloud using evolutionary algorithm," *Soft Computing*, vol. 24, no. 10, pp. 7149–7161, 2020.

[11] A. R. Sulthana, M. Gupta, S. Subramanian and S. Mirza, "Improvising the performance of image-based recommendation system using convolution neural networks and deep learning," *Soft Computing*, vol. 24, no. 19, pp. 14531–14544, 2020.

[12] C. L. S. Bocanegra, J. L. S. Ramos, C. Rizo, A. Civit and L. Fernandez, "HealthRecSys: A semantic content-based recommender system to complement health videos," *BMC Medical Informatics and Decision Making*, vol. 17, no. 1, pp. 1–10, 2017.

[13] J. Chen, K. Li, H. Rong, K. Bilal, N. Yang *et al.,* "A disease diagnosis and treatment recommendation system based on big data mining and cloud computing," *Information Sciences*, vol. 435, no. 12, pp. 124–149, 2018.

[14] F. Ali, S. El. Sappagh, S. M. Riazul Islam, A. Ali, M. Attique *et al.,* "An intelligent healthcare monitoring framework using wearable sensors and social networking data," *Future Generation Computer Systems*, vol. 114, no. 20, pp. 23–43, 2021.

[15] G. Manogaran, R. Varatharajan and M. K. Priyan, "Hybrid recommendation system for heart disease diagnosis based on multiple kernel learning with adaptive neuro-fuzzy inference system," *Multimedia Tools and Applications*, vol. 77, no. 4, pp. 4379–4399, 2018.

[16] F. Ali, S. M. Riazul Islam, D. Kwak, P. Khan, N. Ullah *et al.,* "Type-2 fuzzy ontology-aided recommendation systems for IoT-based healthcare," *Computer Communications*, vol. 119, no. 5, pp. 138–155, 2018.

[17] C. Wu, J. Wang, J. Liu and W. Liu, "Recurrent neural network based recommendation for time heterogeneous feedback," *Knowledge-Based Systems*, vol. 109, no. 9, pp. 90–103, 2016.

[18] P. Kaur, R. Kumar and M. Kumar, "A healthcare monitoring system using random forest and internet of things (IoT)," *Multimedia Tools and Applications*, vol. 78, no. 14, pp. 19905–19916, 2018.

[19] R. C. Chen, Y. H. Huang, C. T. Bau and S. M. Chen, "A recommendation system based on domain ontology and SWRL for anti-diabetic drugs selection," *Expert Systems with Applications*, vol. 39, no. 4, pp. 3995–4006, 2012.

[20] A. K. Sahoo, C. Pradhan, R. K. Barik and H. Dubey, "DeepReco: Deep learning based health recommender system using collaborative filtering," *Computation*, vol. 7, no. 2, pp. 2–25, 2019.

[21] S. Fraile, S. Malwade, D. Spachos, L. Fernandez-Luque, C. T. Su *et al.,* "A recommender system to quit smoking with mobile motivational messages: Study protocol for a randomized controlled trial," *Trials*, vol. 19, no. 1, pp. 1–12, 2018.

[22] T. Selvi, Mahesh and V. Kavitha, "A privacy-aware deep learning framework for health recommendation system on analysis of big data," *The Visual Computer*, vol. 37, pp. 1–19, 2021.

[23] U. A. Bhatti, M. Huang, D. Wu, Y. Zhang, A. Mehmood *et al.,* "Recommendation system using feature extraction and pattern recognition in clinical care systems," *Enterprise Information Systems*, vol. 13, no. 3, pp. 329–351, 2019.

[24] S. H. Liao and C. A. Yang, "Big data analytics of social network marketing and personalized recommendations," *Social Network Analysis and Mining*, vol. 11, no. 1, pp. 1–19, 2021.

[25] Shahbazi, Zeinab, D. Hazra, S. Park and Y. C. Byun, "Toward improving the prediction accuracy of product recommendation system using extreme gradient boosting and encoding approaches," *Symmetry*, vol. 12, no. 9, pp. 1566, 2020.

[26] H. Aziz, R. Hikmat and N. Dimililer, "SentiXGboost: Enhanced sentiment analysis in social media posts with ensemble XGBoost classifier," *Journal of the Chinese Institute of Engineers*, vol. 44, no. 6, pp. 562–572, 2021.

[27] R. Biswarup, A. Garain and R. Sarkar, "An ensemble-based hotel recommender system using sentiment analysis and aspect categorization of hotel reviews," *Applied Soft Computing*, vol. 98, no. 45, pp. 106935, 2021.

[28] I. Baha, M. A. Khan, T. Abbas Khan, S. Abbas, M. S. Daoud *et al.,* "A smart healthcare recommendation system for multidisciplinary diabetes patients with data fusion based on deep ensemble learning," *Computational Intelligence and Neuroscience*, vol. 2021, no. 5, pp. 11, 2021.

[29] Rosa, R. Lopes, G. Maria Schwartz, W. V. Ruggiero and D. Z. Rodríguez, "A knowledge-based recommendation system that includes sentiment analysis and deep learning," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 4, pp. 2124–2135, 2018.

[30] D. Yang, C. Huang and M. Wang, "A social recommender system by combining social network and sentiment similarity: A case study of healthcare," *Journal of Information Science*, vol. 43, no. 5, pp. 635–648, 2017.

[31] R. V. Karthik and S. Ganapathy, "A fuzzy recommendation system for predicting the customers interests using sentiment analysis and ontology in e-commerce," *Applied Soft Computing*, vol. 108, no. 7, pp. 107396, 2021.

[32] S. Rathore and S. Kumar, "A decision tree logic based recommendation system to select software fault prediction techniques," *Computing*, vol. 99, no. 3, pp. 255–285, 2017.

[33] Hossain, M. Deloar, M. S. Azam, M. J. Ali and H. Sabit, "Drugs rating generation and recommendation from sentiment analysis of drug reviews using machine learning," *Emerging Technology in Computing, Communication and Electronics (ETCCE)*, vol. 1, no. 1, pp. 1–6, 2020.

[34] S. Garg, "Drug recommendation system based on sentiment analysis of drug reviews using machine learning," in *11thInt. Conf. on Cloud Computing, Data Science & Engineering (Confluence)*, Noida, India, vol. 11, pp. 175–181, 2021.

[35] M. Alwateer, A. M. Almars, K. N. Areed, M. A. Elhosseini, A. Y. Haikal *et al.,* "Ambient healthcare approach with hybrid whale optimization algorithm and naive bayes classifier," *Sensors*, vol. 21, no. 13, pp. 4579, 2021.