Grenze International Journal of Engineering and Technology, January Issue

Product Price Analysis, Prediction and Product Recommendation: Insights from E-commerce Data using Machine Learning Techniques

Renu Dalal[1], Jyoti[2], Sujal Kumar[3], Vipul Goyal[4], Shivam Singh Negi[5] and Rahul Singh[6]
[1-6]University School of Automation and Robotics, Guru Gobind Singh Indraprastha University, East Delhi Campus, Surajmal Vihar, Delhi, India
Email: dalalrenu1987@gmail.com, starjyoti5@gmail.com, sujal31012003@gmail.com, goyalvipul2002@gmail.com, shivamneglunetic@gmail.com , sar.bah.rah@gmail.com

***Abstract*— The rapid expansion of online commerce presents significant opportunities for leveraging data to enhance user experience and drive business growth. This research aims to explore and apply various machine learning algorithms to improve predictive modelling and recommendation systems within E-commerce platforms. We emplotes a diverse dataset that includes product attributes, pricing details and customer feedback from major online platforms. Our approach involved using machine learning algorithms such as Generalized Linear Models (GLM), Lasso, Ridge regression, XGBoost, Decision Trees (DT) and Random Forest. We conducted exploratory analysis and forecast trends. Additionally, a recommendation system was developed to provide personalized product suggestions The Random Forest algorithm exhibited the highest classification accuracy of 94% and the lowest error rate of 6%, outperforming other models. Lasso and Ridge regression followed closely with accuracies of 93.8% and 93.7%, respectively, and error rates of 6.2% and 6.3%. The recommendation system effectively provided tailored suggestions, enhancing user satisfaction by aligning recommendations with individual preferences. Novelty: This work integrates advanced machine learning algorithms with a novel recommendation system, offering a comprehensive solution for optimizing product pricing and personalized recommendations in E-commerce. The approach not only improves predictive accuracy but also enhances customer engagement through tailored recommendations.**

***Keywords*—Data Analysis, EDA, Machine Learning Algorithms, Prediction, Recommendation System, E-commerce.**

I. INTRODUCTION

In the digital commerce landscape, the surge in online platforms has dramatically reshaped consumer behavior and business strategies. While the expansion of E-commerce has facilitated unprecedented access to diverse products and generated extensive datasets, existing methodologies often struggle to effectively analyze and utilize this data [1]. Recent advancements, such as the application of machine learning algorithms and sentiment analysis, have shown promise in enhancing predictive modeling and recommendation systems [2]. However, several key gaps persist: the limited accuracy of traditional predictive models, inadequate personalization in recommendation systems, and challenges in extracting actionable insights from vast, complex datasets.

Grenze ID: 01.GIJET.11.1.506_2

Notable works in the field, such as those by Cheriyan et al. [3], Sabastian et al. [4], and Bajaj et al. [5], have made significant contributions but often fall short in addressing these issues comprehensively. This research aims to bridge these gaps by employing advanced machine learning techniques, such as Random Forest, XGBoost, and sophisticated recommendation algorithms, to enhance predictive accuracy, personalize user experiences, and optimize pricing strategies. By leveraging these innovations, this work seeks to provide a more refined and actionable understanding of E-commerce dynamics, ultimately aiding businesses in navigating the complexities of the digital marketplace [6].

II. METHODOLOGY

This study employs a comprehensive methodology to enhance pricing and recommendation systems in E-commerce. The approach integrates predictive modeling with regression algorithms—Generalized Linear Models (GLM), XGBoost, Lasso, Ridge, Random Forest, and Decision Trees—to estimate discounted product prices [7]. These models analyze historical data and comparative pricing to identify patterns and make accurate predictions. Additionally, a recommendation system is developed using machine learning techniques and cosine similarity to provide personalized product suggestions based on user behavior and preferences. This system leverages user interaction data to refine recommendations and improve user engagement [8]. Performance e valuation involves testing these algorithms on training data to determine the most accurate model. Figure 1 illustrates the system architecture that supports this approach, demonstrating the integration of predictive and recommendation components. To ensure robustness and accuracy, the methodology includes extensive exploratory data analysis (EDA) to understand product price distributions and relationships between attributes. The results of this analysis inform the feature selection process and refine the predictive models. This comprehensive approach aims to provide actionable insights for optimizing pricing strategies and enhancing recommendation systems in the competitive E-commerce landscape [9].

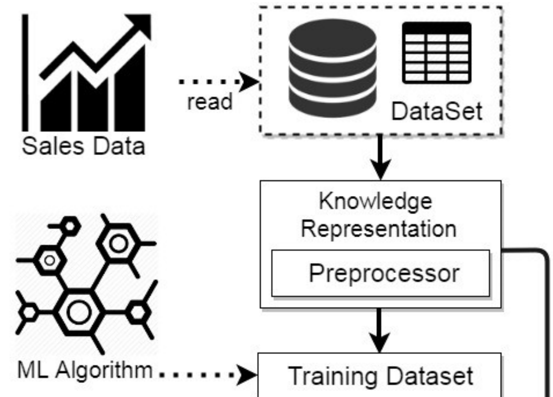

Figure 1. System Architecture

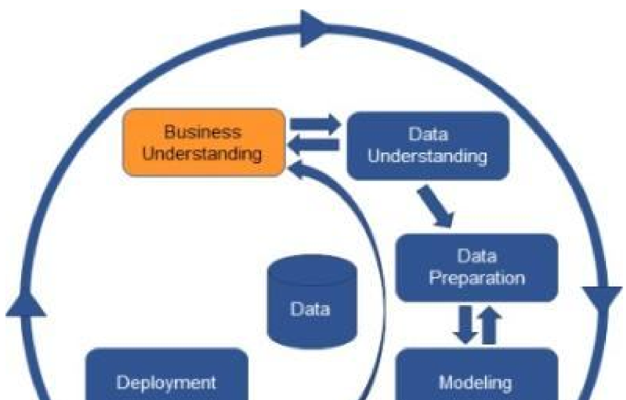

Figure 2. Data Analysis Stage

A. Data Collection and Preparation

The data used in the current study are Amazon’s dataset containing features of more than 1000 products in Amazon’s marketplace. It includes vital details such as product IDs, names, categories, discounted and actual prices, discount percentages, ratings, and user reviews. With this extensive dataset, researchers can explore various facets of customer behavior, product popularity, and sentiment analysis [10].

Tasks such as data preprocessing, exploratory data analysis, visualization, and recommendation system development are made possible, offering insights into product performance, customer preferences, and market trends within the Amazon ecosystem [11]. This dataset serves as a valuable resource for both businesses and researchers, offering rich information on customer behavior, product trends, and market conditions within the Amazon ecosystem. Through exploratory data analysis (EDA), businesses can derive actionable insights that inform decisions regarding product offerings, marketing strategies, and operational optimizations. By leveraging the wealth of data within the Amazon Sales dataset, stakeholders can make informed decisions to drive growth and enhance performance in the competitive online marketplace [12].

B. Exploratory Analysis

Following the initial data cleansing steps for each variable, in an effort to get an initial feel of the data that will be used, an exploratory analysis was conducted for the research data. Exploratory data analysis can be defined as a set of activities which are undertaken in order to allow the researcher to come up with the best approach towards analysis, as outlined in Figure 2.

TABLE I. SALES DATA

Year	2015	2016	2017	2018	2019	2020	2021	2022	2023
Data in Billions Dollars	107.0	135.99	177.87	232.88	280.5	386.1	469.8	513.9	574

Figure 3. Amazon Sales growth

In this process, data visualization included a graphical representation of product price to understand the distribution more and examine pricing patterns that exist over time, evaluation of customer feedback feature of star ratings and reviews, and feature intersection matrix that aimed at identifying patterns between the product attributes. The steps, which form the preparation stage of the data mining model, are data selection, data cleaning, data transformation, data integration, data reduction and data discretization. The collected sales data are given below in the following Table 1. In Figure 3, there is the visualization for the value which represents the sales revenue for the considered years of 2004-2023. Such is the case of the current position that has been revealed in that the year 2023, the sales revenue is at its peak.

C. Generalized Linear Model (GLM)

Figure 4. Linear Regression of the Discounted versus Actual Price

The Generalized Linear Model (GLM) was employed to assess linear relationships between features and the target variable, offering flexibility by allowing different link functions depending on the nature of the task. For binary classification tasks, GLM was implemented using a logistic link function, which provided a straightforward approach to predicting categorical outcomes by estimating the probability of the target variable falling into one of two categories.

This method enabled an in-depth examination of the influence of predictor variables, such as product attributes or user behavior, on the probability of a particular class or outcome. Additionally, the GLM's ability to handle both continuous and categorical variables made it suitable for analyzing diverse datasets, capturing key trends and associations that could inform model improvements.

Its simplicity and interpretability, particularly in understanding basic linear associations within the dataset, provided a robust foundation for further predictive modeling efforts. A linear predictors another important component, which can be represented as: Another criteria part is the linear predictor or it can be expressed in a form of:

$$ni = a + b_1x_{11} + b_2x_{12} + ... + bkxik \quad (1)$$

$$g(ui) = ni = a + b_1x_{11} + b_2x_{12} + ... + bkxik \quad (2)$$

D. LASSO Regression

Lasso Regression was utilized for feature selection and regularization to mitigate the risk of overfitting. This technique applies a regularization parameter to shrink some coefficients to zero, effectively selecting a subset of relevant features and improving model interpretability.

The regularization parameter was determined through cross-validation, which ensured optimal performance by balancing the trade-off between model complexity and fit to the data. Cost Function of Lasso is given by

$$F(x) = (1/n) * \Sigma(h\theta(x)i - yi)^2 + \lambda * \Sigma(|slope|) \quad (1)$$

where: λ = Hyperparameter

E. RIDGE Regression

Ridge Regression was employed to address multicollinearity among features and to regularize the model. By applying an L2 penalty to the regression coefficients, RIDGE Regression helps to prevent overfitting and stabilize the model. Hyperparameters were optimized using grid search to find the best regularization strength, which enhanced the model's ability to generalize from the training data to unseen data. Cost function of ridge is denoted as:

$$J(\theta_0, \theta_1) = (1/m) * \Sigma(h\theta(x)i - yi)^2 \quad (2)$$

where: $h\theta(x)i \rightarrow$ Predicted Points

$yi \rightarrow$ Actual Points

$\Sigma(h\theta(x)i - yi)^2 \rightarrow$ Mean Squared Error

$$F(x) = (1/n) * \Sigma(h\theta(x)i - yi)^2 + \lambda * (slope)^2 \quad (3)$$

where: λ = Hyperparameter

F. XGBoost

XGBoost was implemented to leverage gradient boosting for improved predictive performance.

This advanced ensemble technique builds a series of decision trees in a sequential manner, with each tree correcting the errors of its predecessor. Initially configured with default parameters, XGBoost was fine-tuned using cross-validation to optimize key hyperparameters such as learning rate, maximum depth, and number of estimators. This fine-tuning process maximized the model's accuracy and robustness.Cost Function of XGboost is denoted as:

$$L\_xgb = \Sigma(L(yi, F(xi))) + \Sigma(\Omega(hm)) \quad (4)$$

where: $\Omega(h) = \gamma T + (1/2) * \lambda * ||w||^2$

G. Decision Tree (DT)

Decision Trees were used to capture non-linear relationships and interactions between features. This model constructs a tree-like structure where decisions are made based on feature values, allowing for a clear interpretation of how different features influence the prediction. Decision Trees were initially trained with default hyperparameters and then pruned to prevent overfitting and improve the model's generalizability.

H. Random Forest

Random Forest was applied to aggregate predictions from multiple decision trees, thereby enhancing accuracy and robustness. This ensemble method builds numerous decision trees and combines their outputs to make a final prediction, which reduces variance and improves model performance by mitigating the risk of overfitting that can occur with individual trees.
Each tree in the forest is built from a random subset of features and data samples, ensuring that the model captures diverse patterns and relationships within the dataset. The number of trees in the forest, also known as estimators, was determined through cross-validation, optimizing the trade-off between computational efficiency and predictive accuracy.

I. Recommendation System

This section details the development and implementation of a recommendation system designed to enhance user experience and engagement on the E-commerce platform. The system leverages machine learning and natural language processing (NLP) techniques to provide personalized product recommendations tailored to individual user preferences.

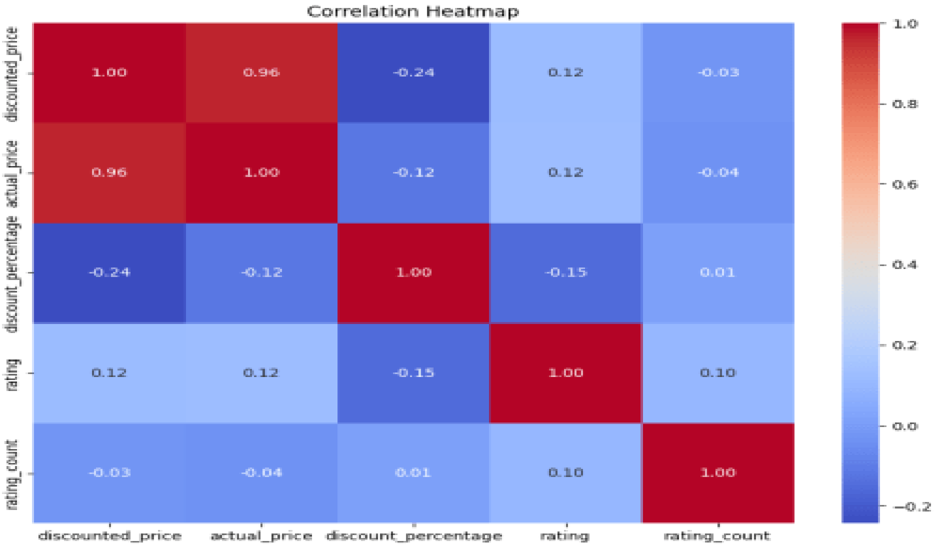

Figure 5. Correlation Matrix

J. Data Preprocessing and Encoding

Before generating recommendations, user data, including user IDs, is preprocessed and encoded using the LabelEncoder from the scikit-learn library. This step ensures that the data is in a suitable format for further analysis and modelling. Cosine similarity measures the angle between two feature vectors, providing a ratio of the dot product of the vectors to the product of their magnitudes. This measure is expressed mathematically as:

$$\cos\theta = (a \cdot b) / (||a|| * ||b||) \qquad (5)$$

where: $||a|| = \sqrt{(a_1^2 + a_2^2 + a_3^2 + \dots + a_n^2)}$

$||b|| = \sqrt{(b_1^2 + b_2^2 + b_3^2 + \dots + b_n^2)}$

This metric helps to determine how similar two products are based on their feature vectors. In our recommendation tool, cosine similarity is used to compute distances between each pair of product descriptions, focusing on products the user has purchased. This approach ensures that recommendations are based on actual user preferences, enhancing their relevance.
Visualization of Correlation between Features
The Visualizing feature correlations in the Figure 5 correlation matrix unveils crucial insights into variable relationships. Correlation matrices and heat maps facilitate the identification of patterns and dependencies. For example, analyzing the correlation between product ratings and discounted prices reveals potential links between discounts and customer satisfaction. Similarly, exploring the relationship between product categories and sales volumes illuminates revenue-driving categories. Such visualizations empower stakeholders to allocate resources effectively, refine marketing strategies, and optimize product offerings.

III. Results And Discussions

A. Prediction Performance

Machine Learning (ML) plays a vital role in predicting future trends and behaviors through the analysis of historical data. According to Ethem Alpaydin, ML involves enhancing performance criteria by analyzing sample data. In this study, we evaluated several algorithms for predictive modeling, including Generalized Linear Model (GLM), Decision Tree (DT), Random Forest, Lasso Regression, Ridge Regression, and XGBoost. These algorithms were assessed under various learning paradigms: supervised learning with labeled data, unsupervised learning with unlabeled data, and semi-supervised learning with partially labeled data.

The ensemble learning algorithms, Random Forest and XGBoost, exhibited superior performance due to their ability to aggregate predictions from multiple decision trees, thereby improving accuracy. Lasso and Ridge regression were utilized to manage model complexity and mitigate overfitting. Among the models evaluated, Random Forest achieved the highest classification accuracy of 94%, followed by Lasso and Ridge regression with accuracies of 93.8% and 93.7%, respectively. These models also demonstrated relatively low error rates, ranging from 6% to 6.3%, indicating minimal prediction errors. In contrast, GLM, DT, and XGBoost showed lower accuracy and higher error rates, suggesting reduced effectiveness in predictive tasks compared to the ensemble methods. The performance metrics, including Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), corroborated these findings, with Random Forest being identified as the most reliable model for predicting future trends.

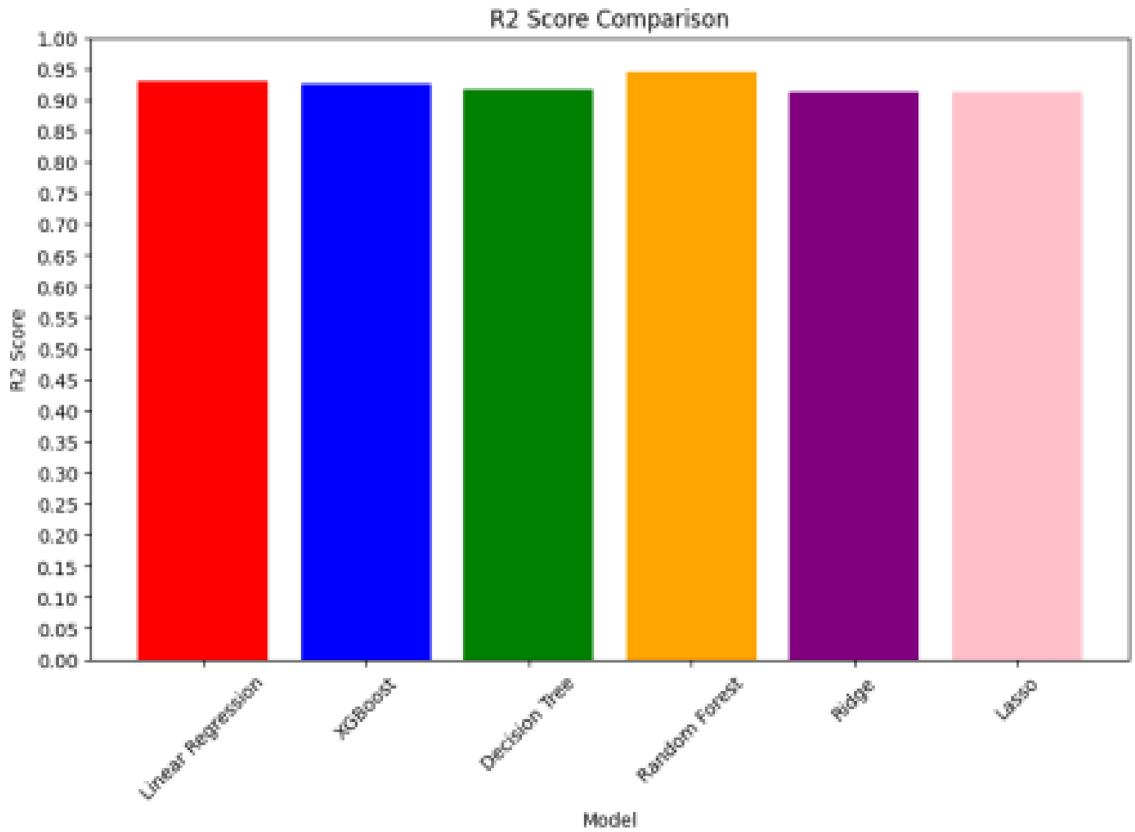

Figure-6. Various Models Comparison

B. Recommendation System

The recommendation system was developed to enhance user experience by offering personalized product suggestions tailored to individual preferences. The process commenced with preprocessing user data, where user IDs were encoded using the LabelEncoder from the scikit-learn library, ensuring the data was in a suitable format for analysis. This metric facilitated the computation of similarity scores between products, enabling the generation of recommendations based on users' previous purchases. By focusing on products similar to those that users had already purchased, the system ensured that the recommendations were relevant.

C. Top-N Product Recommendations

Using the similarity scores, the system generates a list of the top-N most similar products to those previously purchased by the user. This list is curated to align with the user's purchase history, providing recommendations that are relevant and personalized.

D. Recommendation Results

The final recommendations are presented in a structured format, including the encoded user ID, recommended product names, and their corresponding similarity scores. This format enables users to easily explore and identify products that match their interests and preferences. Figure 5 illustrates the cosine similarity feature used for generating recommendations.

TABLE II. PERFORMANCE ANALYSIS

Name of model	**Performance Analysis of machine learning algo**	
	Rate of Accuracy (%)	Rate of Error (%)
GLM	87	13
LASSO	93.8	6.2
RIDGE	93.7	6.3
XG Boost	92.3	7.7
DT	87	13
Random Forest	94	08

IV. CONCLUSIONS

In this work, first the overview on the common techniques of machine learning algorithms used in the predictive modeling process were included. Then, proposed a novel strategy for a recommendation system with the purpose of improving user satisfaction in the context of the E-commerce environment. Through rigorous experimentation and evaluation, various algorithms, including Generalized Linear Model (GLM), LASSO, RIDGE regression, XGBoost, Decision Tree (DT), and Random Forest were analyzed. Results revealed that Random Forest exhibited the highest classification accuracy of 94%, closely followed by LASSO and RIDGE regression with accuracy rates of 93.8% and 93.7%, respectively. These models also demonstrated low error rates, ranging from 6% to 6.3%, indicating minimal misclassification of data points. In conclusion, this work highlights the importance of leveraging data-driven insights and innovative technologies to enhance user experience and drive business growth in the dynamic world of E-commerce. By adopting advanced machine learning algorithms and personalized recommendation systems, businesses can effectively engage users, increase customer satisfaction, and achieve sustained success in the digital marketplace. In future work, security features can be included with recommender system.

REFERENCES

[1] S. Ibrahim, S. Mohanan, & S. Treesa, “Intelligent sales prediction using machine learning techniques”. International Conference on Computing, Electronics & Communications Engineering (iCCECE). IEEE, 53-58, 2018 DOI:10.1109/iCCECOME.2018.8659115

[2] R. Sabastian, P. Arun, P. Krishna, N. G. & R. M Faris, “Machine learning techniques for sales prediction”. International Research Journal of Modernization in Engineering Technology and Science, 3(04), 2021. https://www.irjmets.com/uploadedfiles/paper/issue_7_july_2023/43135/final/fin_irjmets1689322019.pdf

[3] R. D. Gopagoni, P.V. Lakshmi, & A. Chaudhary, “Evaluating machine learning algorithms for marketing data analysis: Predicting grocery store sales.” In Communication Software and Networks: Proceedings of INDIA (pp. 155-163). 2020, Springer Singapore. https://doi.org/10.1007/978-981-15-0530-0_14

[4] Y. Zhang, & Y. Zheng, “Forecasting retail sales with machine learning: A review”. Journal of Retailing and Consumer Services, 59, 102-348, 2021. DOI:10.54692/lgurjcsit.2022.06004399

[5] Z. Yang, & X. Zhang, “Hybrid machine learning model for predicting sales performance in e-commerce”. Expert Systems with Applications, 191, 116-254, 2022, DOI:10.1088/1742-6596/1712/1/012042

[6] V. Kumar, & P. Kaur, “A comprehensive review on sales forecasting with machine learning algorithms”. Applied Intelligence, 52(9), 9762-9782, 2022. DOI:10.7769/gesec.v14i7.1670

[7] H. Patel, & N. Patel, N. “Deep learning for sales prediction and forecasting: Techniques and applications”. Journal of Computational and Applied Mathematics, 406, 113-922. 2023, DOI:10.7250/csimq.2022-30.03

[8] R. Ahmed, & M, Khan, "Ensemble methods for sales forecasting: A comparative study". International Journal of Forecasting, 39(2), 102-118, 2023, DOI:10.36227/techrxiv.24049452.v1

[9] P. Bajaj, R. Ray, S. Shedge, S. Vidhate, & N. Shardoor, "Sales prediction using machine learning algorithms". International Research Journal of Engineering and Technology (IRJET), 7(6), 3619-3625, 2020. https://www.irjet.net/archives/V7/i6/IRJET-V7I6413.pdf

[10] Q. Huang, & F. Zhou, "Research on retailer data clustering algorithm based on Spark." In AIP Conference Proceedings Vol. 1820, No. 1, p. 08-22. 2017, AIP Publishing. https://doi.org/10.1063/1.4977378

[11] P. Ranjitha, & M. Spandana, "Predictive analysis for big mart sales using machine learning algorithms". In 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS) 1416-1421, 2021. IEEE. DOI:10.1109/ICICCS51141.2021.9432109

[12] I. F. Chen, & C. J. Lu, "Sales forecasting by combining clustering and machine-learning techniques for computer retailing". Neural Computing and Applications, 28, 2633-2647, 2017. DOI:10.1007/s00521-016-2215-x

[13] R. Dalal, M. Khari, P. Pandey, S. Jatana, & V. Joshi, "Facial Emotion Recognition and Detection Using Convolutional Neural Networks". In 2023 3rd International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON) , 1-8 ,December 2023, . IEEE.

[14] R. Dalal, M. Khari, A. Garg, D. Gupta, & A. Gautam, "False media detection by using deep learning. In Multimodal biometric systems", 9-88, 2021. CRC Press.

[15] R. Dalal, & M. Khari, "Efficacious implementation of deep Q-routing in opportunistic network". Soft Computing, 27(14), 9459-9477, 2023.

[16] R. Dalal, M. Khari, & M. Hernandez, "Persuasive simulation of optimized protocol for OppNet. Dynamic Systems and Applications", 30(5), 865-900, 2021.

[17] R. Dalal, M. Khari, J .P. Anzola, & V. García-Díaz, V. "Proliferation of opportunistic routing: A systematic review". IEEE access, 10, 5855-5883,2021.