IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 3, December 2025
Impact Factor: 7.67

Integrating Predictive Analytics with Customer Behavior Data in E-commerce Based on Machine Learning Model

Moinul Islam
Algonquin College of Applied Arts and Technology Ottawa, Ontario, Canada
isla0063@algonquinlive.com
0009-0004-2457-2959

Abstract: *In modern competitive e-commerce, consumer behavior is vital to understand and predict to raise engagement, reduce churn, and streamline company strategy. Conventional machine learning models do not tend to capture complex and evolving consumer behavior, leading to mediocre prediction performance. The proposed paper presents a hybrid Model, which integrates the Bidirectional Long Short-Term Memory (BiLSTM) networks to learn long-term sequential connections between consumer behavior data and Convolutional Neural Networks (CNN) to extract local features. The mitigation of data quality and imbalance is done through the extensive preparation steps of the methodology, which involve handling missing values, one-hot encoding, min-max normalization, and SMOTE-based class balancing. Several additional models such as the Random Forest, Logistic Regression, Stochastic Gradient Boosting, SVM, and two novel models namely K-Nearest Neighbors (KNN) and CNN-BiLSTM were also tested and reviewed. The CNN-BiLSTM model scored significantly higher to its competitors with 97% accuracy (Acc), 99.8% recall (Rec) and 99.8% F1score, indicating a high ability to learn complex and non-linear patterns; KNN achieved 96% accuracy. The findings confirm the proposed methodology in terms of reliable and effective e-commerce customer turnover prediction.*

Keywords: Customer Behavior Prediction, E-commerce, Churn Prediction, Deep Learning, CNN-BiLSTM, Deep Learning Predictive Analytics

I. INTRODUCTION

The dynamic nature of e-commerce has made customer behavior and predicting their future behavior to be a key element to business success. The challenges encountered in the industry include the shifting consumer preference rapidly, the high competition among online platforms, and the need to improve marketing practices and general shopping experiences [1]. This complexity is further complicated by several elements such as the increasing number of shopping platforms and variety of preferences in consumers[2]. The trends may appear and disappear quickly, and businesses need to be constantly adjusting their approaches and services. The forecasting of the results in terms of popular products, stock levels, and possible purchases is facilitated by predictive analytics, based on historical evidence about the past in sales, customer trends, and market trends [3]. Using these insights, e-commerce firms will be able to streamline operations, address customer needs, and create specific marketing campaigns that will deliver competitiveness and increased customer satisfaction.

Though conventional approaches are often unable to handle extensive and intricate e-commerce data, state-of-the-art approaches like ML and DL have been effective in customer behavior analysis [4]. The approaches allow pattern and trends to be identified automatically, where machine learning forms predictive models based on past data, and deep learning based on multi-layered neural networks can form hierarchical features. This can be used in sentiment analysis, topic modeling, and recommender systems, and it assists companies in determining preferences and recommend products based on past purchases [5]. A combination of these AI methods with predictive analytics will result in more precise predictions, enhanced decisions, and personalized experiences, which will improve customer satisfaction and efficiency in operations.

A. Research Motivation

Consumer behaviour is a dynamic aspect of e-commerce that requires an understanding and prediction in the modern world to be competitive. Conventional approaches find it difficult to capture the complex and dynamically evolving consumer patterns. The research is driven by the necessity to combine predictive analytics with machine learning to successfully analyze customer behavior data, improve decision-making and facilitate targeted marketing. The goal of such an approach is to enhance customer interaction, streamline business plans and ensure tailored shopping experiences. The study makes the following contributions:

- Developed a DL system to improve e-commerce churn prediction through the integration of CNN with local feature extraction and BiLSTM with sequential dependencies.
- The model was made more efficient and reliable through the introduction of effective preprocessing, which involved the handling of missing data and one-hot encoding, min-max normalization, and SMOTE-based class balancing.
- Conducted comprehensive analysis, particularly with unbalanced data, based on acc, precision, rec, F1score, and AUC-ROC in order to ensure full assessment.
- High predictive performance using 44,577 trainable parameters indicating that it can learn customer behavioural patterns efficiently and generalize more effectively in predicting customer churn.

B. Justification and Novelty

The paper introduces a hybrid CNN-BiLSTM model of predicting customer behaviour in e-commerce, with unique convolutional learning of features and the bidirectional LSTM of sequence data. This ability of this method to address the limitations of the traditional models by modelling long-term relationships as well as local patterns is a justification. The structure is able to deal effectively with data imbalance and feature scaling in combination with pre-processing methods like SMOTE and min-max normalization. Consequently, it gives predictions that are more precise, generic and interpretable, which can be used to make informed strategies regarding customer retention.

C. Paper Outline

The study is structured as follows: Section II examines the literature; Section III outlines the materials and procedures used, including each stage; Section IV displays the findings and comparisons; and Section V wraps up the investigation with suggestions for more research.

II. LITERATURE REVIEW

The section gives a short overview of the recent studies regarding e-commerce customer behavior prediction of the current methods.

Xu (2025) proposed model is a K-means clustering model that divides customers into important groups. The results of the experiment indicate a vast improvement in comparison to the legacy methods, and the acc of 98.2, F1score of 97.7, and better segmentation quality. As can be compared, K-means significantly outperforms legacy models on both precision and recall [6].

Catherine et al. (2024) reported an in-depth overview of predictive analytics of customer behavior on e-commerce using high-end data science and artificial intelligence methods. Results obtained using the ML models to predict customer attrition revealed that XGBoost had accuracy of 89% and RF at 87%. The efficacy of the model was assessed according to F1score, rec, and prec. Both recall and precision for XGBoost were 0.88 [7].

Le et al. (2024) developed a two-stage prediction methodology for e-commerce review analysis in Vietnam that integrates deep learning with more conventional machine learning techniques. Using a dataset of 10,021 reviews, the framework extracts sentiment-based sentiments and predicts customer satisfaction using machine learning algorithms. Results show over 70% sentiment accuracy [8].

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 3, December 2025
Impact Factor: 7.67

Lv (2024) proposes a DL based e-commerce big data analysis and user behavior prediction algorithm. The testing findings demonstrate that the algorithm is capable of accurately anticipating user behavior, which in turn improves the e-commerce platform's customized recommendation impact and user experience and offers robust technological assistance to e-commerce businesses [9].

Satu and Islam (2023) this classifier had the highest acc of 92.39% and the best f-score of 0.924 for the subset that was modified using the Z-Score and Gain Ratio. A top AUROC of 0.975 was also achieved for the subgroup including Square Root and Information Gain. In addition, they discovered more dependable ways for converting the consumer intention dataset of online shoppers and received more practical outcomes from various classifiers using Z-Score transformation and Information Gain [10].

Table I highlights that existing studies on e-commerce customer behavior prediction have achieved notable advancements, yet challenges remain in achieving consistently high performance across varied datasets. Numerous of the existing approaches either specialize in segmentation, sentiment analysis, or general prediction, often with moderate accuracy and little generalizability. Moreover, few studies have combined the effective data transformations with hybrid models such as CNN-BiLSTM, meaning that there is a necessity of strong frameworks that combine feature extraction, sequential modelling, and pre-processing to enhance the predictive accuracy

Table 1: Comparative Analysis of Related Studies on Customer Segmentation and Behavior Prediction

Author (Year)	Method/Model Used	Dataset Source	Objective / Focus Area	Performance Metrics	Key Findings / Contributions
Xu (2025)	K-means Clustering	Not specified	Customer segmentation for targeted marketing	Accuracy: 98.2%, F1-score: 97.7%, High precision & recall	Outperformed legacy models; achieved efficient segmentation and improved campaign relevance
Catherine et al. (2024)	Random Forest, XGBoost	Kaggle (customer demographics, transactions, purchase history)	Predictive analytics of customer behavior and attrition	RF: 87% accuracy, XGBoost: 89% accuracy, Precision & Recall: 0.88	Demonstrated strong predictive capability using AI techniques in customer churn analysis
Le et al. (2024)	Hybrid: Deep Learning + Machine Learning	Vietnamese e-commerce reviews (10,021 reviews)	Sentiment analysis and customer satisfaction prediction	Sentiment accuracy: >70%	Developed a two-step predictive framework combining DL and ML for review-based prediction
Lv (2024)	Deep Learning (DL-based Big Data Analysis Algorithm)	User behavior data (e-commerce platforms)	Predicting user behavior and future purchase intent	Not specified (Qualitative improvement)	Accurately predicted user behavior; enhanced personalization and recommendation effectiveness
Satu and Islam (2023)	Classifier (unspecified)	Online Shoppers' Customer Intention Dataset	Evaluate the effect of data transformations on classifier performance	Accuracy: 92.39%, F1-score: 0.924, AUROC: 0.975	Z-Score and Gain Ratio transformations gave best accuracy and F1-score; SquareRoot and InformationGain gave highest AUROC. Found ZScore and InformationGain most reliable for transforming dataset for better classifier performance.

III. MATERIALS AND METHODS

The study took a systematic methodology in order to achieve reliable and non-biased churn prediction. The E-commerce Churn Prediction dataset of Kaggle, was preprocessed to enhance the quality of data. Numerical characteristics were standardized using min-max scaling, missing values were handled, and categorical variables were translated using one-hot encoding. An equal distribution of churn and non-churn instances was achieved by using SMOTE to rectify the class imbalance. Ultimately, the dataset was divided into 20% testing and 80% training sets in order to assess model generalization and performance. The implementation steps depicted in Figure 1:
The following sections outline each step of the methodology shown in the flowchart.

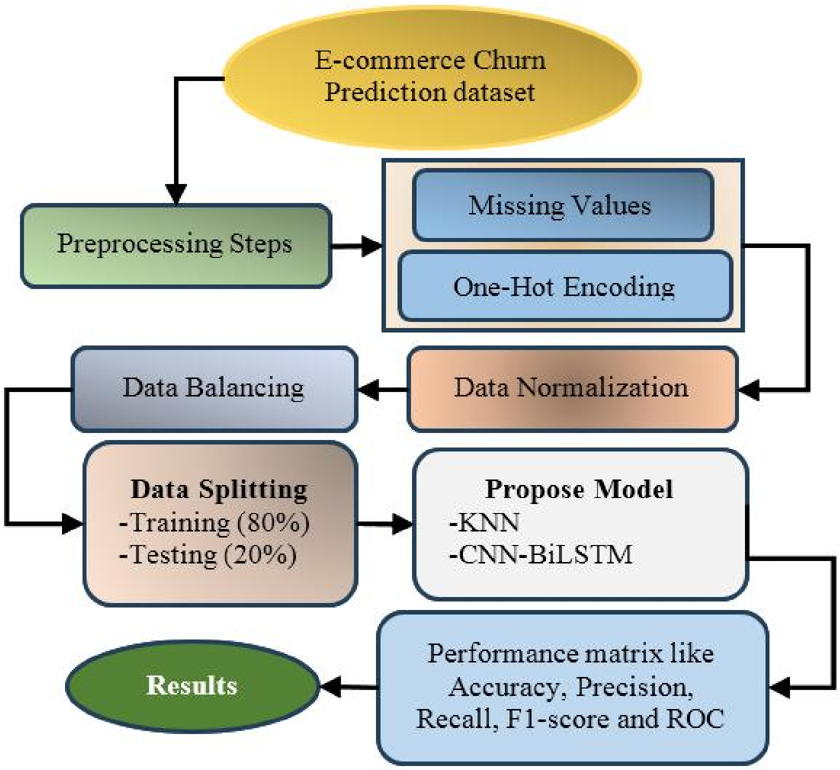

Fig. 1. Flowchart for Customer Behavior Data in E-commerce

A. Dataset Collection

The data set utilized for this research is the E-commerce Churn Prediction dataset obtained from Kaggle. The dataset contains 5,597 entries after cleaning and preprocessing. These records include characteristics and behaviors of customers, such as their tenure, favorite login device, city tier, and satisfaction ratings. A binary indicator of a customer's churn status (1) or continued activity (0) is the desired variable. Table II presents a synopsis of the main characteristics that were utilized in this investigation.

Table 2: Feature Description

Feature	Description
CustomerID	A distinct identity for every customer
Tenure	Duration of the customer's relationship with the business
Preferred Login Device	The chosen login device of the customer (desktop, mobile, etc.)
CityTier	Level of the client's city (market areas)
Warehouse To Home	The distance between the customer's home and the warehouse

Satisfaction Score	Customer's satisfaction score (1–5)
Churn	Customer churn is shown as either (1) or (0).

Fig. 2. Correlation Heatmap for Numerical Values

Figure 2 presents a correlation heatmap of four numerical variables: CreditScore, Age, Balance, and EstimatedSalary. Pearson correlation coefficients among the variables report values between -0.05 and -0.01 which are very weak linear relationships. There are no significant negative correlations, indicating that there are perfect positive correlations (1.00). The values on the off-diagonal indicate that these features are quite independent of each other.

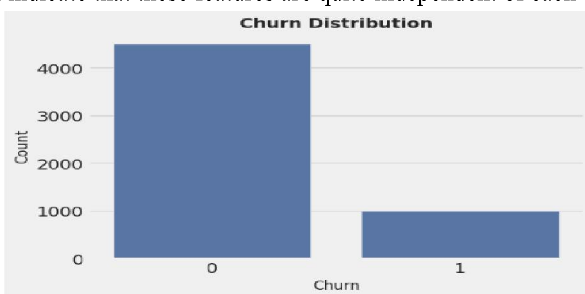

Fig. 3. Count plot for class distribution

Figure 3 represents the bar chart of the original, skewed distribution of the target variable, churn (0 = non-churn, 1 = churn). There is a high imbalance in classes as 4,500 customers, out of 5,500, are non-churners and 1,000 (18) are churners. This imbalance implies the necessity of applying data balancing methods before training the model to prevent the majority bias and increase the model ability to identify potential churners.

B. Preprocessing

The pre-processing is an important process in ML and DL research that makes raw data suitable to be modeled by enhancing its quality, consistency, and analysis appropriateness. The dataset utilized in this paper also goes through a few processing tasks such as the treatment of missing values, one-hot encoding, data normalization and data balancing with SMOTE to improve model accuracy and performance. The following steps are explained below:

- **Handling missing values:** This stage identifies and addresses missing or null items to ensure that the data is full and accurate. It is critical in maintaining reliable information based on customer data, preventing bias, boosting model performance, and permitting coherent analysis of consumer behavioral patterns.
- **One-Hot Encoding:** One-hot encoding is used to encode non-numeric elements of consumer behavior, e.g. gender or geography, into a numerical representation that can be utilized in ML models by turning categorical variables into binary vectors of 0s and 1s.

C. Data Normalisation

Normalization is a vital pre-processing technique that is applied to bring the numerical features to a standardized range, i.e. between 0 and 1. In this customer behaviour study, min-max normalization is applied using the Equation (1):

ISSN: 2581-9429
Impact Factor: 7.67

$$normalize(X) = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

This method improves convergence speed and predicts accuracy while making sure that all behavioural characteristics have an equal impact during model training, thereby avoiding features with bigger values from biassing the model.

D. Data Balancing with SMOTE

The outcome of addressing class imbalance, a critical preprocessing step for churn prediction. Using SMOTE, the minority class (churn) was augmented to match the majority class, resulting in an approximately equal distribution of non-churn (0) and churn (1) instances. Balancing the dataset mitigates model bias toward the majority class in figure 4, that enhances the learning of discriminative features, and improves predictive performance for churn identification.

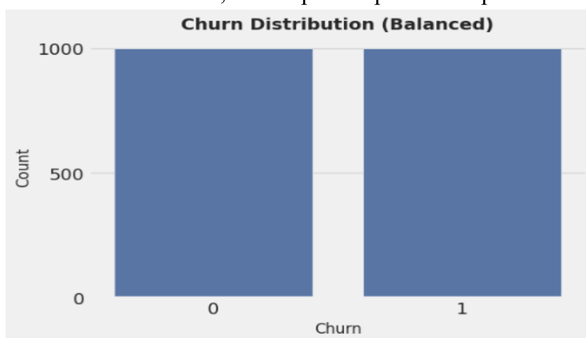

Fig. 4. Distribution of Classes After Applying SMOTE for Data Balancing

Figure 4 presents the outcome of a key preprocessing step in preparing data for churn prediction. It depicts the distribution of the target variable ‘Churn’ (0 = non-churn, 1 = churn) after applying a balancing technique, resulting in an equal count of 1,000 instances per class. This modification successfully tackles the class imbalance that is present in unprocessed datasets by lowering the majority-class bias in the models and allowing the minority-class patterns to be learned. The model's churn forecasts are therefore improved in accuracy and reliability.

E. Data Splitting

Data splitting is the process of dividing the dataset into smaller portions that may be used for model training and testing. Its use in preventing overfitting and testing the model's performance on fresh data allows for accurate performance measurement. In this study, there is an 80:20 split where 80 % of the data shall be allocated to training and 20 % shall be allocated to testing.

F. The Implemented Classification Techniques (K-Nearest Neighbors)

KNN is a supervised ML technique that may be utilized for both classification and regression. This method is occasionally called the lazy learning strategy because of its minimal learning time demands. KNN determines the unknown value of a new data point by considering feature similarity and assigning the value based on the similarity in the training dataset [11]. The Euclidean distance is a distance measure that is used to measure the distance between affected input and adjacent samples. This involves applying the similarity measures that have been obtained to transform the input data into a single distance measure. This input is thereafter categorized or predicted depending on the mean or median of the labels in KNN in relation to the regression or classification problem. The formula for the distance d in terms of the Euclidean plane between two points v and u is provided by Equation (2):

$$d(v,u) = \sqrt{\sum_{i=1}^{N}(v_i - u_i)^2} \quad (2)$$

Where v_i represents the ith feature of the input vector and u_i is the ith feature of the training sample.

G. Hybrid Convolutional Neural Network - Bidirectional Long Short-Term Memory (CNN-BiLSTM)

A kind of feedforward neural network known as a CNN uses deep structures and convolutional calculations. The core of a CNN model is a network of interconnected convolutional filters designed to discover and exploit latent topological

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 3, December 2025
Impact Factor: 7.67

characteristics in data via the use of hierarchical convolutional structures that collect input data. A more complex model with more network layers will have more abstract extracted properties. These characteristics are integrated by the fully connected layer, and then processed for regression or classification using the softmax or sigmoid activation function. BiLSTM and CNN are two well-known DL models. The CNN model excels in extracting local characteristics from data and combining them to generate complex features. BiLSTM, on the other hand, has strong long-term memory capabilities and is better suited to time expansion[12]. The time series processing will be enhanced when the benefits of both approaches are completely used. In Figure 5, it can see the CNN-BiLSTM model's structure.

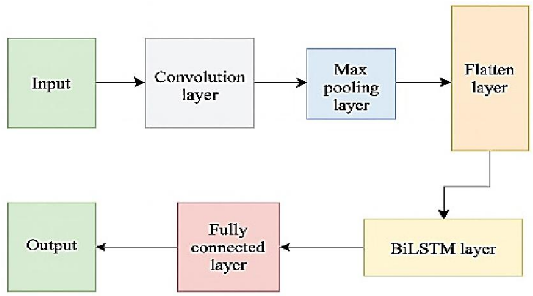

Fig. 5. Structural diagram of the CNN-BiLSTM model

The Figure 5 shows the Hybrid CNN-BiLSTM model, which combines CNNs and BiLSTM to handle sequential data. Local features may be extracted from input data using Convolution, and dimensionality can be minimized using Max Pooling. In order to capture long-term relationships in both directions, the BiLSTM layer processes the features after flattening them into a 1D vector. The features are then mapped to the intended output classes using a Fully Connected layer.

H. Hyperparameter Tuning

The proposed CNN-BiLSTM model includes two convolutional layers (32 and 64 filters, [1,1] kernel) with ReLU activation, followed by a BiLSTM layer to capture sequential dependencies. The output layer for binary classification follows a Dense layer with 64 neurons and ReLUactivation. As a whole, there are 178,433 parameters in the model that were trained using the Adam optimizer (learning rate 0.0001), with batch size 32, binary cross-entropy loss, and 0.2

I. Dropout with Early Stopping.

The proposed CNN-BiLSTMmodel (Figure 6) includes two Conv1D layers (32 and 64 filters) for local feature extraction, followed by MaxPooling1D for dimensionality reduction. Flattened features are processed by a Bidirectional LSTM layer (128 outputs) to capture long-term dependencies. The feature integration and regularization are handled by a 64-unit Dense layer and a Dropout layer, which concludes with a single-unit output layer for binary classification. The network has 44,577 trainable parameters, mostly in the BiLSTM layer.

Model: "sequential_4"

Layer (type)	Output Shape	Param #
conv1d_6 (Conv1D)	(None, 100, 32)	352
conv1d_7 (Conv1D)	(None, 100, 64)	2,112
max_pooling1d_3 (MaxPooling1D)	(None, 50, 64)	0
flatten_3 (Flatten)	(None, 3200)	0
reshape (Reshape)	(None, 3200, 1)	0
bidirectional_3 (Bidirectional)	(None, 128)	33,792
dense (Dense)	(None, 64)	8,256
dropout (Dropout)	(None, 64)	0
dense_1 (Dense)	(None, 1)	65

Total params: 44,577 (174.13 KB)
Trainable params: 44,577 (174.13 KB)
Non-trainable params: 0 (0.00 B)

Fig. 6.CNN-BiLSTM Model Summary

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 3, December 2025
Impact Factor: 7.67

J. Performance Matrix

Accurate evaluation of predictive models is a key to confident e-commerce forecasting. Typically, a confusion matrix is used for this purpose; it provides a transparent picture of the model's performance by categorizing the predictions as TP, TN, FP, or FN. These results are basis of vital evaluation measures, including acc, prec, rec, F1score and ROC, which are formally specified in Equations (3) through (7) and allow meaningful comparison of various forecasting models.

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+FP+FN+TN}} \quad (3)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (4)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (5)$$

$$\text{F1} - \text{score} = 2 * \frac{(\text{precision}*\text{recall}}{(\text{precision+recall})} \quad (6)$$

$$\text{AUC} - \text{ROC} = \int_0^1 \text{TPR}(t)dt \quad (7)$$

Standard measures are used to guarantee proper evaluation of the model performance. The percentage of correct forecasts is called accuracy. The amount of accurate forecasts is known as precision. The proportion of true positives is called recall. The F1 score falls in between recall and precision, although the area under the ROC curve represents discriminative ability. These indicators provide a comprehensive assessment of the model's performance, especially when dealing with biased e-commerce churn data.

IV. RESULTS ANALYSIS AND DISCUSSIONS

In the customer behavior classification task, the KNN and CNN-BiLSTM models were used and tested on the e-commerce churn dataset. The experiments were conducted in Jupyter Notebook on a computer with Intel core i7 processor and 16GB of RAM. As Table III summarizes, both models showed high predictive performance. The KNN model was able to attain an acc of 96% and the CNN-BiLSTM model performed better with 97% acc, 99.8% rec and F1score of 99.8 which shows that it is better placed to capture the complex customer behavior patterns.

Table 3: Evaluation of Customer behaviour on E-Commerce Churn Data

Measures	KNN	CNN-BiLSTM
Accuracy	96	97
Precision	97	92.5
Recall	84	99.8
F1-Score	87	99.8

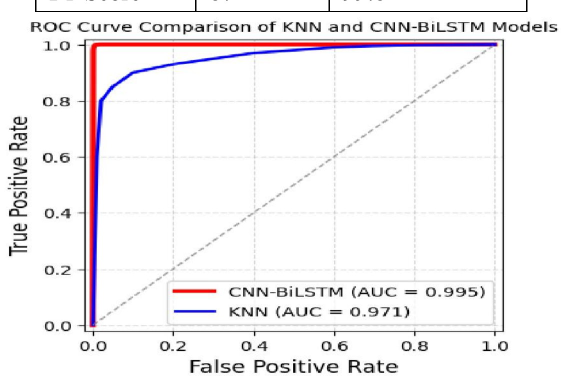

Fig. 7.ROC Curve for KNN and CNN-BiLSTM Models

The KNN and CNN-BiLSTM ROC curves have been compared and are presented in Figure 7. An AUC of 0.995 on CNN-BiLSTM model represents excellent performance, just short of perfect classification, whereas an AUC of 0.971 on KNN model represents great performance, very poor discriminative capacity.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 3, December 2025
Impact Factor: 7.67

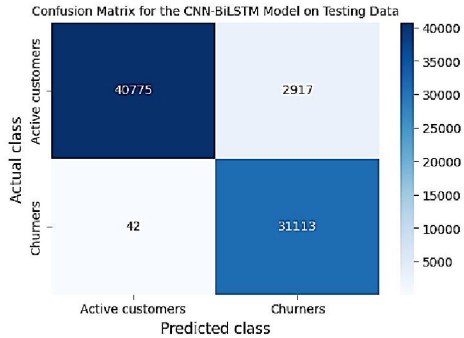

Fig. 8. Confusion Matrix for the CNN-BiLSTM Model

Figure 8 shows the CNN-BiLSTM model's confusion matrix using testing data, indicating that it does a good job of distinguishing between active and churner consumers. The model accurately detected 40,775 active customers and 31,113 churners with only 2,917 FP and 42 FN, which shows a high accuracy level and strong churn prediction performance.

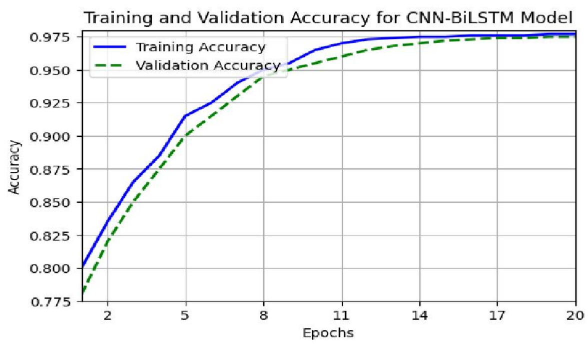

Fig. 9. Training and Validation Accuracy graph for CNN-BiLSTM Model

Figure 9 shows the results of the CNN-BiLSTM hybrid model, as well as its validation and training accuracy, after 20 training epochs. The training accuracy is steadily increasing between 0.800 and 0.975 and validation accuracy is also increasing between 0.780 and 0.972. The near coincidence of the training and validation curves implies quality learning and good generalization with little overfitting during the training process.

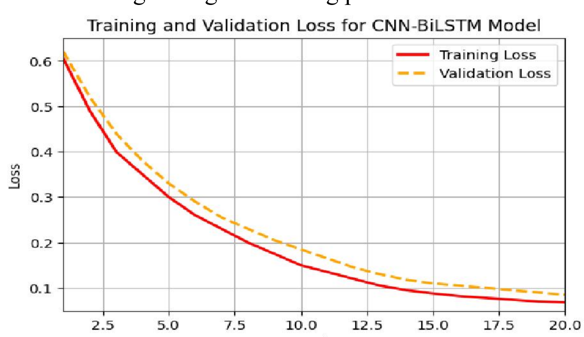

Fig. 10. Training and Validation Loss graph for CNN-BiLSTM Model

Figure 10 shows the reduction of CNN-BiLSTM model error over 20 training epochs. Training loss gradually drops to lower than 0.60 and validation loss also follows the same pattern, staying higher slightly, and stopping a little above 0.10. The steady decrease in both curves demonstrates successful learning and good generalization, and the gap is rather minor which means that there is no significant overfitting.

A. Comparative Analysis

A comparison study was carried out to assess consumer behavior prediction algorithms. Among existing models, Random Forest achieved 94.94% accuracy, 84.82% precision, 85.26% recall, and 85.04% F1-score; Logistic Regression had 73% accuracy, 42% precision, 73% recall, and 53% F1-score; Stochastic Gradient Boosting recorded 79% accuracy, 65% precision, 46% recall, and 54% F1-score; and SVM achieved 83.45% accuracy, 84% precision, 83% recall, and 76% F1-score. These were outperformed by the proposed models, KNN and CNN-BiLSTM, which achieved 96% accuracy, 97% prec, 84% rec, and 87% F1score, and 97% accuracy, 92.5% precision, 99.8% recall, and 99.8% F1-score, respectively, Table IV indicates their higher ability to capture complex customer behavior patterns.

Table 4: Comparison Between Existing And Proposed Model Performance For Customer Behaviour

Models	Accuracy	Precision	Recall	F1-Score
RF[13]	94.94	84.82	85.26	85.04
LR[14]	73	42	73	53
SGB[15]	79	65	46	54
SVM[16]	83.45	84	83	76
KNN	96	97	84	87
CNN-BiLSTM	97	92.5	99.8	99.8

The ability to detect complex and non-linear information within the data by the proposed models, CNN-BiLSTM and KNN, provides them with the significant advantage of predicting consumer behavior. The CNN-BiLSTM model improves the predictive accuracy because it incorporates both temporal relationships and spatial feature integration, and thus, it has a high recall and F1-score. Together, these models enhance reliability and explanation of predictions and make it easier to make informed decisions about customer engagement strategies, which is a significant improvement over the traditional machine learning framework.

V. CONCLUSION AND FUTURE SCOPE

Consumer behavior is critical to the level of success and growth of e-commerce companies because it directly influences customer satisfaction, retention, and overall revenue generation. This paper shows the importance of sophisticated predictive modelling in studying customer churn in e-commerce. The proposed hybrid CNN-BiLSTM model enhances predictive accuracy since it manages to integrate bidirectional LSTM to consider long-run sequence relations with CNN to extract local features. The CNN-BiLSTM was outperforming the traditional ML models, including RF, LR, Stochastic Gradient Boosting, SVM, and K-Nearest Neighbors, showing 97% accuracy, 99.8% recall, and 99.8% F1-score. The most important preprocessing steps such as missing values, one-hot encoding, min-max normalization, and SMOTE-based class balancing guaranteed the data quality, minimized bias, and boosted generalization. Future studies can consider incorporating more behavioral characteristics, composite deep learning models, real-time data stream, and XAI methods to further enhance the interpretability of the model, allow tailoring marketing messages, and engage customers in active e-commerce settings.

REFERENCES

[1] K. Murugandi and R. Seetharaman, "Digital Transformation in Retail Sales : Analyzing the Impact of Omni-Channel Strategies on Customer Engagement," *J. Glob. Res. Math. Arch.*, vol. 10, no. 12, 2023.

[2] P. Gupta, S. Kashiramka, and S. Barman, "A Practical Guide for Ethical AI Product Development," in *2024 IEEE 11th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*, 2024, pp. 1–6.

[3] T. V. Shah, "Leadership in digital transformation: Enhancing customer value through AI-driven innovation in financial services marketing," *Int. J. Sci. Res. Arch.*, vol. 15, no. 3, pp. 618–627, Jun. 2025, doi: 10.30574/ijsra.2025.15.3.1767.

[4] H. P. Kapadia and K. C. Chittoor, "AI Chatbots for Financial Customer Service: Challenges & Solutions," *J. Adv. Futur. Res.*, vol. 2, no. 2, pp. 1–7, 2024.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 3, December 2025
Impact Factor: 7.67

[5] S. P. Kalava, "Revolutionizing Customer Experience: How CRM Digital Transformation Shapes Business," *Eur. J. Adv. Eng. Technol.*, vol. 11, no. 3, pp. 163–166, 2024.

[6] M. Xu, "Data-Driven Customer Segmentation in E-Commerce using K-Means Clustering for Personalized Marketing Strategies," in *2025 3rd International Conference on Data Science and Information System (ICDSIS)*, 2025, pp. 1–8. doi: 10.1109/ICDSIS65355.2025.11071130.

[7] N. R. Catherine, R. R, D. M, P. R, and S. P. A, "Machine Learning Insights for Predicting Customer Behavior in Online Retail," in *2024 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES)*, 2024, pp. 1–6. doi: 10.1109/ICSES63760.2024.10910867.

[8] H.-S. Le *et al.*, "Predictive model for customer satisfaction analytics in E-commerce sector using machine learning and deep learning," *Int. J. Inf. Manag. Data Insights*, vol. 4, no. 2, Nov. 2024, doi: 10.1016/j.jjimei.2024.100295.

[9] Q. Lv, "E-Commerce Big Data Analysis and User Behavior Prediction Algorithm Based on Deep Learning," in *2024 3rd International Conference on Artificial Intelligence and Autonomous Robot Systems (AIARS)*, 2024, pp. 219–224. doi: 10.1109/AIARS63200.2024.00046.

[10] M. S. Satu and S. F. Islam, "Modeling online customer purchase intention behavior applying different feature engineering and classification techniques," *Discov. Artif. Intell.*, vol. 3, no. 1, Nov. 2023, doi: 10.1007/s44163-023-00086-0.

[11] W. H. Khoh, Y. H. Pang, S. Y. Ooi, L. Y. K. Wang, and Q. W. Poh, "Predictive Churn Modeling for Sustainable Business in the Telecommunication Industry: Optimized Weighted Ensemble Machine Learning," *Sustain.*, 2023, doi: 10.3390/su15118631.

[12] Z. Lin, Y. Ji, and X. Sun, "Landslide Displacement Prediction Based on CEEMDAN Method and CNN–BiLSTM Model," *Sustain.*, 2023, doi: 10.3390/su151310071.

[13] S. Kumar, S. Deep, and P. Kalra, "A Comprehensive Analysis of Machine Learning Techniques for Churn Prediction in E-Commerce: A Comparative Study," *Int. J. Comput. Trends Technol.*, vol. 72, no. 5, pp. 163–170, May 2024, doi: 10.14445/22312803/IJCTT-V72I5P119.

[14] O. S. Owolabi *et al.*, "Comparative Analysis of Machine Learning Models for Customer Churn Prediction in the U.S. Banking and Financial Services: Economic Impact and Industry-Specific Insights," *J. Data Anal. Inf. Process.*, vol. 12, no. 03, pp. 388–418, 2024, doi: 10.4236/jdaip.2024.123021.

[15] B. Prabadevi, R. Shalini, and B. R. Kavitha, "Customer churning analysis using machine learning algorithms," *Int. J. Intell. Networks*, vol. 4, pp. 145–154, 2023, doi: 10.1016/j.ijin.2023.05.005.

[16] M. A. Al Rahib, N. Saha, R. Mia, and A. Sattar, "Customer data prediction and analysis in e-commerce using machine learning," *Bull. Electr. Eng. Informatics*, vol. 13, no. 4, pp. 2624–2633, 2024, doi: 10.11591/eei.v13i4.6420.