ISSN 2723-6471


# Deciphering Digital Social Dynamics: A Comparative Study of Logistic Regression and Random Forest in Predicting E-Commerce Customer Behavior

Po Abas Sunarya[1,*], Untung Rahardja[2], Shih-Chih Chen[3], Yung-Ming Li[4], Marviola Hardini[5]

[1,2,5] *University of Raharja, Tangerang 1511, Indonesia*

[3]*National Kaohsiung University of Science and Technology, Kaohsiung 81157, Taiwan*

[4]*National Yang Ming Chiao Tung University, Taipei 30010, Taiwan*

(Received: November 22, 2023; Revised: December 21, 2023; Accepted: January 15, 2024; Available online: January 29, 2024)

**Abstract**

This study compares Logistic Regression and Random Forest in predicting e-commerce customer churn. Utilizing the E-commerce Customer dataset, it navigates the complexities of customer interactions and behaviors, offering a rich context for analysis. The methodology focuses on meticulous data preprocessing to ensure data integrity, setting the stage for applying and evaluating Logistic Regression and Random Forest. Both models were assessed using accuracy, precision, recall, F1-Score, and AUC-ROC. Logistic Regression showed an accuracy of 90%, precision of 91% for class 0 and 82% for class 1, recall of 98% for class 0 and 50% for class 1, F1-Score of 94% for class 0 and 62% for class 1, and AUC-ROC of 0.88. Random Forest, with its ability to handle complex patterns, demonstrated higher overall performance with an accuracy of 95%, precision of 95% for class 0 and 93% for class 1, recall of 99% for class 0 and 74% for class 1, F1-Score of 97% for class 0 and 82% for class 1, and an AUC-ROC of 0.97. This comparative analysis offers insights into each model's strengths and suitability for predicting customer churn. The findings contribute to a deeper understanding of machine learning applications in e-commerce, guiding stakeholders in enhancing customer retention strategies. This research provides a foundation for further exploration into the digital social dynamics that shape customer behavior in the evolving digital marketplace.

*Keywords:* E-commerce Churn Prediction, Machine Learning Algorithms, Logistic Regression, Random Forest, Customer Behavior Analysis, Predictive Modeling

## 1. Introduction

The profound impact of digital technology on societal behavior and interactions is increasingly evident, driven by the interconnectedness of digital networks[1], mobile devices, and vast amounts of data. This digitalization shapes societal transformation, extending beyond technology use to encompass issues like digital inequalities in eHealth and the digital divide, impacting social health inequalities [1]. As such, the societal implications of digital technology are diverse, influencing modern life aspects from the workplace to environmental, political, and economic dynamics [2]. The advent of e-commerce exemplifies these shifts, revolutionizing shopping and customer behavior, fundamentally altering consumer interactions with businesses, and transforming purchasing decisions. This shift in consumer behavior is marked by a greater focus on online behavior and the impact of e-commerce on customer loyalty and perceptions, emphasizing the intricate influence of e-commerce on customer behavior and satisfaction [3].

Understanding digital social dynamics is paramount in this technological era, as digital and social media have reshaped social interactions and communication, redistributed symbolic resources and altered public participation. These dynamics extend to various domains, including education, health care, and the workplace, where the impact of digital technology requires a nuanced understanding of digital media's role in intergenerational communication, digital health literacy, and the workplace [4]. The significance of these dynamics is further highlighted in the context of Society 5.0, where technological advancements are integral to societal transformation [5].

***Corresponding author: Po Abas Sunarya (abas@raharja.info)**
**DOI: https://doi.org/10.47738/jads.v5i1.155**

Data science emerges as a critical tool in unveiling the patterns of digital social dynamics, providing insights into the complex interplay between technology, society, and individuals. By analyzing vast digital datasets, data science enables the identification of trends and correlations in social interactions and customer behavior [6]. Its application spans various domains, including digital scholarship, humanities, and learning media development, where data analysis technology research is vital [7]. Furthermore, data science is essential in assessing e-commerce systems, understanding customer acquisition patterns, and evaluating cross-border e-commerce performance, offering valuable insights for businesses and researchers [8].

E-commerce platforms are invaluable for understanding the intricate dynamics of modern consumer behavior, preferences, and social interactions. This data provides a comprehensive view of customer engagement, aiding in the development of strategies to enhance satisfaction and loyalty. Studies like Lin [9] and Yang [10] emphasize the critical role of trust, satisfaction, and loyalty in e-commerce, highlighting how these platforms shape customer perceptions and drive loyalty.

The significance of e-commerce data extends to analyzing external impacts on consumer behavior, such as the effects of the COVID-19 pandemic researched by Fu [11]. This type of analysis offers insights into how external events shape consumer decisions and behaviors. Additionally, e-commerce platforms are instrumental in predicting and addressing customer churn, with studies like Fan [3] demonstrating how data can be used to develop early warning models for customer retention. The exploration of customer engagement, as discussed in the context of community e-commerce, further underscores the importance of e-commerce data in understanding and enhancing customer interactions.

In broadening the scope, Błoński [12] highlights the strategic use of data and technology in gaining sustainable advantages and understanding consumer misbehavior. The interdisciplinary nature of consumer behavior research is evident in works like Siebert [13], demonstrating the relevance of consumer behavior modeling in diverse fields such as power systems. The transition in consumer spending behaviors, the influence of value perceptions on satisfaction, and the nuances of customer feedback are further explored in studies examining shifts from traditional to modern markets, the impact of values in coffee shops, and the motives behind customer complaints [14].

In summary, e-commerce platforms are a critical source of data, offering detailed insights into customer behavior, preferences, and social interactions. This data is essential for businesses to understand and adapt to the digital lifestyle and decision-making processes of modern consumers. By harnessing these insights, businesses can enhance customer experiences, foster loyalty, and drive growth in an increasingly digital marketplace.

Predicting customer churn and understanding the drivers of customer loyalty and engagement in a digital context present multifaceted challenge that require sophisticated analytical tools and a deep understanding of consumer behavior. The task is complex due to the vast and varied data generated from multiple customer touchpoints. Advanced analytical techniques, such as machine learning algorithms and data mining, are increasingly employed to identify at-risk customers and develop targeted retention strategies [15]. However, the challenge lies in the complexity of customer data, which includes call behaviors, purchase history, and demographic information, necessitating comprehensive models tailored to specific industries like telecommunications, banking, and e-commerce [16].

Another critical challenge is deciphering the underlying factors contributing to churn. Businesses must move beyond predicting churn to understand the reasons behind it, such as service quality, customer satisfaction, and feedback. This requires a deep dive into data to uncover the drivers of churn and devise proactive strategies to mitigate it [17].

In the realm of customer loyalty and engagement, the digital landscape has transformed how businesses interact with consumers. Trustworthiness, positive experiences, and perceptions of relationship quality are key drivers of engagement, leading to loyalty [18]. To address these challenges, organizations must leverage sophisticated analytical tools and algorithms capable of handling the complexity and volume of digital data. These tools should not only predict behaviors but also provide insights into the emotional and psychological drivers of customer loyalty and engagement. As businesses strive to enhance customer experiences and drive loyalty in the digital era, understanding and leveraging these multifaceted factors are essential for success.

Machine learning algorithms like Logistic Regression and Random Forest are becoming integral in the e-commerce sector for analyzing customer behavior and predicting churn. These sophisticated tools delve into complex customer

data, enabling businesses to discern patterns and trends that inform decision-making. For instance, Logistic Regression, a statistical model, helps estimate the likelihood of customers discontinuing their services, providing a clear understanding of various factors influencing customer decisions. Random Forest, an ensemble learning method that employs multiple decision trees, is particularly effective in improving prediction accuracy and preventing overfitting, making it a valuable tool for identifying at-risk customers [19], [20], [21].

The effectiveness of these algorithms extends beyond simple churn prediction. They offer actionable insights that enable businesses to tailor marketing strategies, improve customer service, and enhance overall customer satisfaction and loyalty. This capability is crucial in today's digital landscape, where customer retention and loyalty are pivotal for business success. Moreover, the application of ensemble learning approaches, such as Adaboost and stacking techniques, has been shown to enhance the predictive power of these algorithms, making them more robust and reliable in various settings, including telecommunications, banking, and e-commerce [22].

In conclusion, the integration of machine learning algorithms like Logistic Regression and Random Forest in e-commerce provides businesses with a competitive edge by allowing them to better understand and predict customer behavior. These algorithms are instrumental in extracting detailed insights from customer data, leading to more informed and effective strategies for customer retention and engagement. As e-commerce continues to evolve, the role of these advanced analytical tools in shaping business strategies and enhancing customer experiences becomes increasingly significant.

While numerous studies have showcased the effectiveness of various machine learning models in predicting customer behavior in e-commerce, there remains a significant research gap in comprehensively comparing these models' effectiveness. Ahmad [23], [24], [25] experimented with algorithms like Decision Tree, Random Forest, GBM, and XGBOOST specifically in the telecom sector, offering insights into their predictive power in customer churn. However, the direct applicability and comparative effectiveness of these models in an e-commerce context remain less explored. Similarly, Pondel [15] developed a deep learning model tailored for e-commerce churn prediction, yet how this model stacks up against other machine learning techniques in the same domain is not fully addressed [26], [27]. These studies, while valuable, highlight the need for more focused comparative research to discern which models offer the best insights and predictive accuracy specifically for e-commerce. Gan [28] emphasizes the effectiveness of the XGBoost algorithm in predicting e-commerce customer loss, yet without a comparative framework, it's challenging to determine its relative performance against other models. This lack of comparative studies creates a significant research gap, leaving e-commerce stakeholders without a clear understanding [29] of which model or combination of models would best suit their specific needs.

To address this research gap, there's a pressing need for systematic comparative studies that evaluate the predictive power and practical utility of various machine learning models in an e-commerce setting. Such research should not only compare traditional models but also explore the potential of advanced techniques and hybrid models. Understanding the relative effectiveness of these models will significantly aid e-commerce stakeholders in selecting the most appropriate and effective tools for analyzing customer behavior and predicting churn, ultimately leading to better-informed strategies and improved customer engagement. As the e-commerce landscape continues to evolve, filling this research gap becomes increasingly critical for maintaining competitive advantage and fostering sustainable growth.

This study specifically focuses on a comparative analysis of Logistic Regression and Random Forest algorithms in predicting customer churn in an e-commerce setting. The central research question driving this investigation is: How do Logistic Regression and Random Forest algorithms compare in terms of accuracy, precision, and recall in predicting customer churn, and what insights do they provide about the social dynamics of customers in the digital era. To address this question, the study is structured around several key objectives. The first objective is to understand customer behavior in e-commerce by analyzing patterns and characteristics such as tenure, device usage, payment methods, and shopping preferences within the e-commerce dataset. This foundational analysis sets the stage for applying and evaluating the chosen machine learning models.

The next two objectives are the core of the study, focusing on the implementation and evaluation of Logistic Regression and Random Forest algorithms, respectively. This comparative analysis aims to uncover the strengths and limitations

 ISSN 2723-6471


of each model in the context of e-commerce customer churn prediction. Finally, the study aims to synthesize these findings to interpret the broader implications for e-commerce stakeholders. Together, these objectives provide a comprehensive roadmap for understanding the comparative effectiveness of Logistic Regression and Random Forest in predicting customer churn and deriving valuable insights for e-commerce strategies.

## 2. Method

This section provides an overview of the entire methodological framework, detailing the systematic steps taken from data collection to the final analysis. To visually encapsulate our methodological journey from data collection through to the final analysis, Figure 1 presents a conceptual diagram of the methodological approach. This illustration provides a clear, step-by-step guide to the processes and analytical techniques employed, setting the stage for a deeper understanding of the methodology detailed in the following sections.

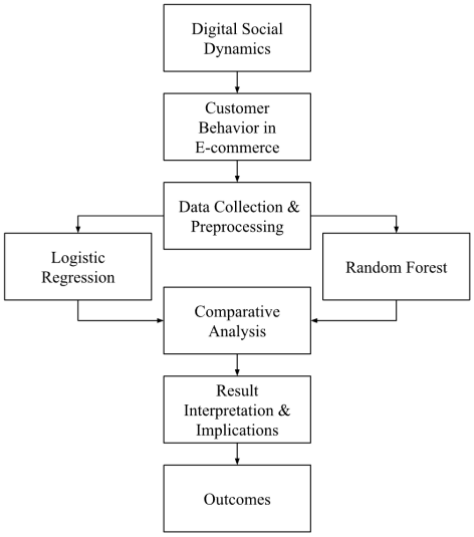


**Figure 1.** Conceptual Framework

### 2.1. Data Collection

In this section, we focus on the "E-commerce Customer" dataset, a pivotal element in our analysis. This dataset has been sourced from Kaggle. Specifically, this dataset is designed to understand and predict customer churn in e-commerce settings, making it highly relevant to our study's objectives.

The dataset comprises a variety of features capturing customer behavior and interaction with the e-commerce platform. It includes both numerical and categorical data types, each offering unique insights into the customer's relationship with the service. Numerical variables include 'Tenure' (the duration of customer engagement with the platform), 'WarehouseToHome' (distance in between warehouse to home of customer), 'HourSpendOnApp' (number of hours spent on the mobile application or website), 'NumberOfDeviceRegistered', 'OrderAmountHikeFromlastYear', 'CouponUsed', 'OrderCount', 'DaySinceLastOrder', and 'CashbackAmount'. These variables provide quantitative measures of customer engagement, preferences, and service utilization, which are crucial for predicting churn.

Categorical data, on the other hand, includes 'PreferredLoginDevice', 'CityTier', 'PreferredPaymentMode', 'Gender', 'PreferedOrderCat', 'SatisfactionScore', 'MaritalStatus', 'Complain', and the target variable 'Churn'. These categorical variables offer insights into the customer's demographics, preferences, and satisfaction levels, which are vital for understanding the factors influencing churn. The 'Churn' variable, specifically, is a binary indicator of whether a customer has left the platform, serving as the primary outcome for our predictive models. The relevance of each variable in this dataset is substantial, as they collectively provide a holistic view of customer behavior and preferences.

 ISSN 2723-6471 

## 2.2. Data Preprocessing

In this section we detail the critical steps undertaken to prepare the dataset for the comparative analysis of Logistic Regression and Random Forest algorithms. Firstly, we tackle the issue of missing value. For numerical variables, missing values are imputed using the median of each respective column. For categorical variables, the most frequent category is used for imputation. This approach ensures the dataset is complete and reflects the underlying distribution of the data without the bias that outliers might introduce. A thorough verification follows to ensure no missing values remain.

Next, we address data transformation and encoding. Numerical variables are standardized using StandardScaler, normalizing them to have a mean of zero and a standard deviation of one. For categorical variables, one-hot encoding is applied, transforming them into a binary format. Lastly, outlier detection and treatment are conducted using the Interquartile Range (IQR) method. Values that fall far from the central tendency, defined as 1.5 times the IQR from the 1st and 3rd quartiles, are treated to minimize their potentially skewing effect on the analysis.

These meticulous preprocessing steps are designed to optimize the dataset for modeling, ensuring it accurately reflects the complex dynamics of customer behavior and is suitable for the rigorous comparative analysis that will follow. With a clean, comprehensive, and appropriately transformed dataset, the study is well-prepared to move into the modeling phase, where the performance of Logistic Regression and Random Forest algorithms will be evaluated and compared.

## 2.3. Algorithm Implementation

Logistic Regression is a widely used statistical method that models the probability of a binary outcome based on one or more predictor variables. It's particularly suitable for this study due to its effectiveness in binary classification tasks, like predicting whether a customer will churn (1) or not (0). The algorithm provides probabilities that a specific event occurs, making it invaluable for understanding the likelihood of churn based on various customer characteristics. To implement Logistic Regression, we will use the 'liblinear' solver, a good choice for small datasets and binary classification, and adjust the Regularization Strength (C) to balance the trade-off between correctly classifying training instances and maintaining a simple model to avoid overfitting. The implementation will be carried out using libraries like Scikit-learn, a robust and widely used Python library for machine learning.

Random Forest is an ensemble learning method that operates by constructing multiple decision trees during training and outputting the class that is the mode of the classes (classification) of the individual trees. It's particularly relevant to this study due to its high accuracy, ability to handle large datasets with higher dimensionality, and its effectiveness in addressing overfitting issues often present in decision trees. The Random Forest algorithm will be implemented with a specified number of trees (n_estimators), depth (max_depth), and the number of features considered when looking for the best split (max_features). These hyperparameters will be tuned to optimize the model's predictive performance. Similar to Logistic Regression, we will utilize the Scikit-learn library for implementation.

Both models will be trained and tested on a split dataset. The data will be divided into a training set, which is used to train the models, and a testing set, which is used to evaluate their performance. This split ensures that the model's performance is evaluated on unseen data, reflecting its potential performance in real-world scenarios. The split ratio is 70% for training and 30% for testing.

## 2.4. Comparative Analysis and Performance Metrics

The comparative analysis will be conducted through a series of steps designed to assess and contrast the performance of Logistic Regression and Random Forest in predicting e-commerce customer churn. Both models will be trained and tested on the same dataset to ensure a fair comparison. To determine if differences in performance between the two algorithms are statistically significant, we will employ statistical tests such as the paired t-test on their performance metrics. This test will help us ascertain whether the observed differences in performance are due to the models' inherent capabilities or just random chance.

Several performance metrics will be used to evaluate and compare the models, including accuracy, precision, recall, F1-Score, and the Area Under the Receiver Operating Characteristic curve (AUC-ROC). Accuracy measures the proportion of total correct predictions, providing an overall effectiveness of the model. Precision and recall are particularly important in the context of churn prediction, where the cost of false positives and false negatives can be

 ISSN 2723-6471


high. Precision measures the proportion of actual positives among predicted positives, while recall measures the proportion of actual positives that were correctly identified. The F1-Score is the harmonic mean of precision and recall, providing a single measure of the model's precision and recall balance. Finally, the AUC-ROC score is a performance measurement for classification problems at various threshold settings, indicating the model's ability to distinguish between the classes.

These metrics are appropriate for comparing Logistic Regression and Random Forest as they provide a comprehensive view of each model's performance across different dimensions. Accuracy offers a baseline comparison, while precision, recall, and F1-Score provide deeper insights into the models' performance concerning the churned customers, which is critical for business decisions. AUC-ROC provides an aggregate measure of performance across all classification thresholds, reflecting the models' ability to rank predictions rather than their absolute values.

## 3. Result and Discussion

### 3.1. Data Collection

The dataset central to this study is titled "E-commerce Customer" from Kaggle. This dataset contains 5630 entries, each with 20 attributes related to customer behavior and demographic information in an e-commerce setting. The attributes range from customer ID to various metrics indicative of their shopping behavior and satisfaction.

The dataset comprises features vectors belonging to 12,330 sessions, each representing a unique customer's interaction with an e-commerce website over a 1-year period. The variables include both numerical and categorical data, ranging from the tenure of the customer within the organization to more intricate details such as the number of devices registered, satisfaction scores, and the nature of their transactions.

Another noteworthy aspect of this dataset is its direct relevance to modern e-commerce dynamics. As the dataset encompasses a wide array of behavioral metrics over a substantial period, it offers a rich, multidimensional perspective on customer churn. This characteristic is particularly beneficial for applying and comparing complex machine learning models like Logistic Regression and Random Forest, as it provides a sufficient depth and variety of data to train and test the models effectively.

### 3.2. Exploratory Data Analysis

In the exploratory data analysis (EDA) of the dataset, various key findings have been unveiled. CustomerID serves as a unique identifier, spanning from 50,001 to 57,630, while the Churn variable reveals that approximately 16.8% of customers have churned. Tenure data indicates that the average customer tenure is about 10.19 months, with a broad range encompassing both new and long-standing customers. The distribution of CityTier suggests that most customers come from city tiers 1 and 3, with fewer from tier 2. Additionally, WarehouseToHome, which averages approximately 15.64 units, showcases diverse delivery distances. HourSpendOnApp demonstrates that customers spend an average of around 2.9 hours on the app, while NumberOfDeviceRegistered reveals an average of about 3.69 devices registered per customer. SatisfactionScore ranges from 1 to 5, signifying varying levels of customer satisfaction, and NumberOfAddress shows that customers have between 1 to 22 addresses registered, with an average of about 4.2. OrderAmountHikeFromlastYear indicates an average increase of 15.7% in the order amount from the previous year, potentially reflecting changes in purchasing behavior or pricing strategies. Moreover, CouponUsed reveals that customers, on average, use 1.75 coupons, with a wide range from 0 to 16. The number of orders placed in the last month, as depicted by OrderCount, varies widely with an average of about 3, and CashbackAmount indicates that the average cashback received is about 177 units, with a maximum of 325.

Count plots presented in Figure 2 below are constructed to analyze categorical data. These plots displayed the varied preferences and demographics among customers, with some categories being more prevalent than others, reflecting common trends or preferences.

 ISSN 2723-6471


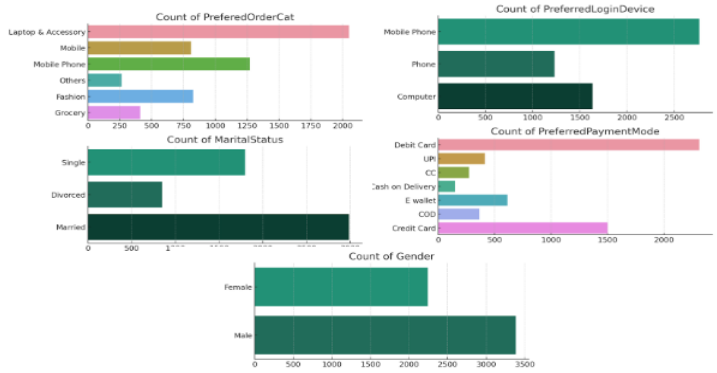


**Figure 2.** Count Plots

Distribution plots constructed in Figure 3 below employed to visualize the distributions of numerical variables such as Tenure, WarehouseToHome, and HourSpendOnApp, revealing a diverse range of customer behaviors and potential segments within the customer base. Some distributions exhibited skewness, hinting at the presence of outliers or specific customer groups with distinct characteristics.

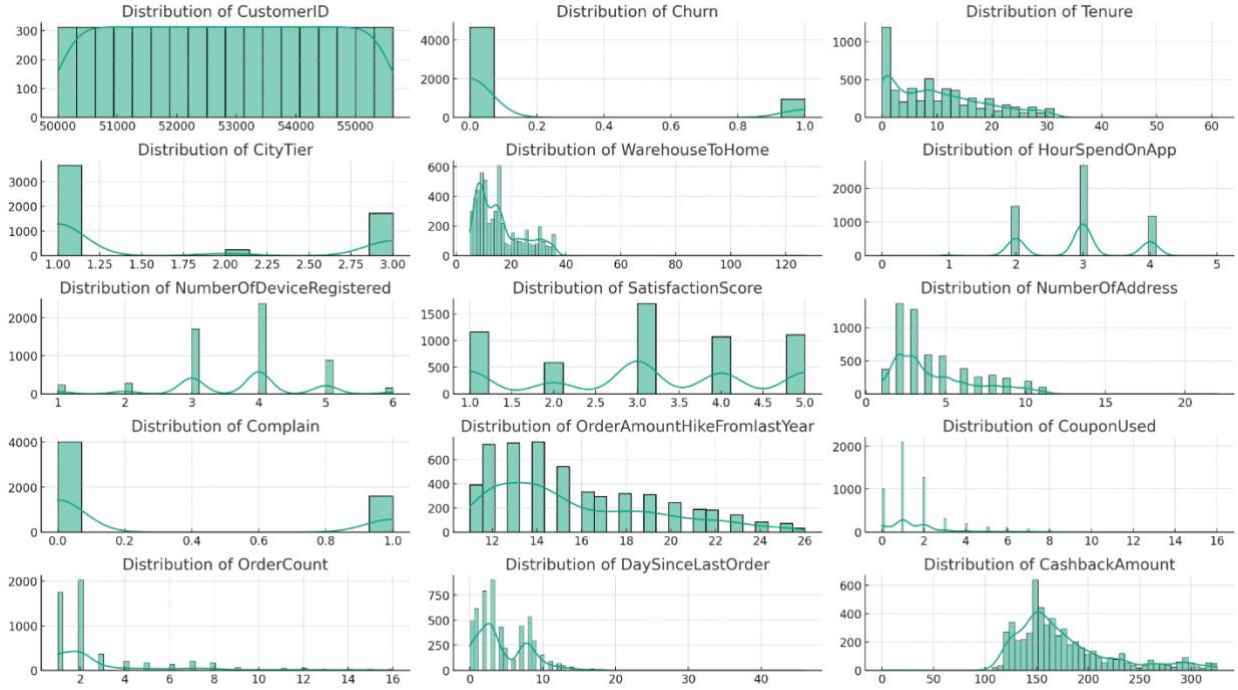


**Figure 3.** Distribution Plots

Additionally, a missing data analysis indicated that several variables, including Tenure, WarehouseToHome, HourSpendOnApp, OrderAmountHikeFromlastYear, CouponUsed, OrderCount, and DaySinceLastOrder, contained missing values. Lastly, a correlation analysis (visualized in Figure 4) was conducted to explore relationships between numerical variables. This analysis highlighted potential correlations, such as the relationship between OrderCount and CouponUsed or the variation in CashbackAmount with OrderAmountHikeFromlastYear. These findings offer valuable insights into customer behavior and help identify potential redundancies within the data.

 ISSN 2723-6471


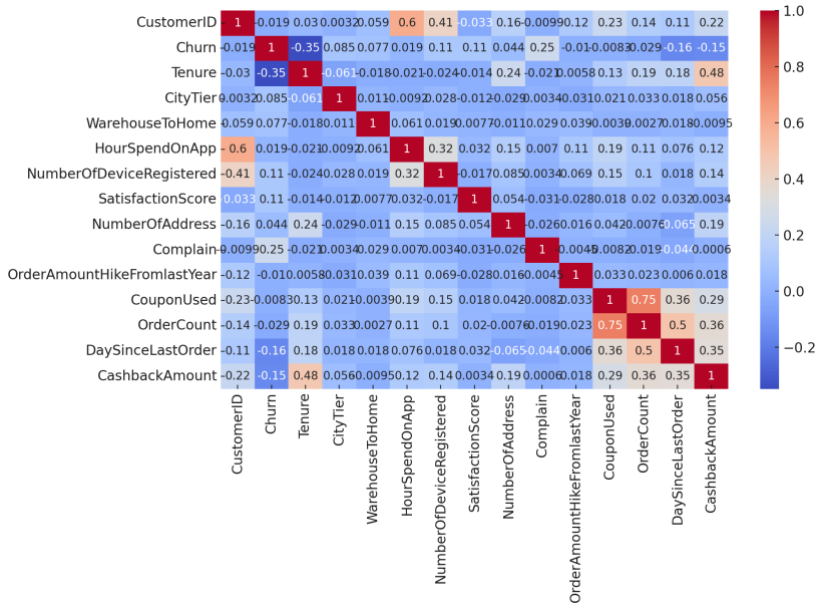


**Figure 4.** Correlation Heatmap

## 3.3. Data Preprocessing

In the data preprocessing phase, several strategies were employed to handle missing data, outliers, and prepare the dataset for analysis. Building on the insights gained during the exploratory data analysis (EDA), these steps ensured the data's quality and suitability for modeling. To address missing data, numerical variables with missing values were imputed using the median of their respective columns, a robust measure that is less influenced by outliers. Meanwhile, categorical variables with missing values were imputed with the mode, representing the most frequent category in each column. Following these imputations, the dataset was thoroughly checked, and it was confirmed that there were no remaining missing values.

Outliers in numerical columns were treated using the Interquartile Range (IQR) method. Values lying beyond 1.5 times the IQR from the 1st and 3rd quartiles were capped to the nearest allowable values within this range. This approach effectively managed extreme outliers while preserving the integrity of the data. In terms of feature engineering, a new feature named 'AppEngagement' was introduced by combining 'HourSpendOnApp' and 'NumberOfDeviceRegistered'. This composite feature aimed to capture overall app engagement, with the assumption that more devices and time spent on the app signify higher engagement.

Further data transformation steps involved standardizing numerical variables using the StandardScaler to normalize their distributions, optimizing their suitability for machine learning models. Categorical variables were one-hot encoded using the OneHotEncoder, converting them into a format compatible with machine learning algorithms. As a result, the transformed dataset now encompasses 36 features, encompassing the original variables and additional features generated through one-hot encoding.

In the post-preprocessing dataset, there are a total of 37 columns. The 'Churn' column, which signifies the churn status of each customer (0 for not churned and 1 for churned), has been retained as the target column, and it occupies the last position in the processed dataset. The remaining 36 columns encompass various predictor variables, including scaled numerical features, one-hot encoded categorical features, and the newly engineered feature 'AppEngagement.' One significant achievement of the preprocessing phase is the complete absence of missing values in the dataset. This indicates that the imputation strategies applied during preprocessing have been successful in ensuring data completeness.

The rationale behind the preprocessing step is: Firstly, handling missing data was imperative to ensure the dataset's completeness, a fundamental requirement for constructing reliable predictive models. By imputing missing values, we ensure that no critical information is omitted from the dataset, thus enhancing the quality and utility of the subsequent modeling process.

Secondly, the treatment of outliers was a crucial step in data preparation. Utilizing the Interquartile Range method effectively managed outliers by capping extreme values, thereby mitigating their potential to disproportionately influence the predictive models. This approach promotes more robust and accurate results by reducing the undue impact of outliers on model performance.

Thirdly, feature engineering introduced the 'AppEngagement' feature, strategically designed to provide additional insights into customer behavior. This newly engineered feature has the potential to enhance the predictive power of the models by incorporating meaningful information that may not have been fully captured by the original dataset.

Lastly, data transformation was essential for aligning the data with the requirements of various machine learning algorithms. Standardizing numerical features and one-hot encoding categorical variables were vital steps in this process. Standardization ensures that numerical features are on a similar scale, preventing any feature from dominating the modeling process due to its scale. One-hot encoding converts categorical variables into a binary format, facilitating their use in machine learning algorithms that require such representation.

With this cleaned and transformed dataset in hand, the next phase of the process will involve Model Building. Logistic Regression and Random Forest models will be constructed, evaluated, and compared using this data. The 'Churn' column will serve as the target variable for predictions, while the other columns will function as predictors to train and assess the performance of these models. This well-prepared dataset sets a solid foundation for the subsequent stages of predictive modeling and analysis.

## 3.4. Model Building Process

The Model Building Process involved the development and evaluation of two distinct machine learning models: Logistic Regression and Random Forest.

For Logistic Regression, the best hyperparameters were determined to be a Regularization Strength (C) of approximately 0.089, and the solver 'liblinear.' This model exhibited strong performance, with an accuracy of around 90%, indicating that approximately 90% of the predictions were correct. Precision for class 0 (Not Churned) was notably high at 91%, while for class 1 (Churned), it was 82%. Recall for class 0 (Not Churned) was 98%, reflecting the model's ability to correctly identify customers who did not churn, while for class 1 (Churned), it was 50%. The F1-Score, a balanced measure of precision and recall, was 94% for class 0 and 62% for class 1. Additionally, the AUC-ROC score was 0.88, indicating the model's capacity to effectively distinguish between the two classes. Challenges and considerations included addressing imbalanced classes, potentially achieved through class weight adjustments, identifying the most predictive features for improved model simplicity and performance, and further hyperparameter tuning, particularly for the regularization strength (C) and solver selection.

For Random Forest, the best hyperparameters were determined as follows: the number of trees (n_estimators) was set at 50, the maximum depth of trees (max_depth) was set to 20, and the number of features considered when looking for the best split (max_features) was set to 'auto.' This model demonstrated impressive performance, with an accuracy of approximately 95%, indicating that around 95% of the predictions were correct. Precision was high for both class 0 (Not Churned) at 95% and class 1 (Churned) at 93%. Recall for class 0 (Not Churned) was 99%, highlighting the model's ability to accurately identify customers who did not churn, while for class 1 (Churned), it was 74%. The F1-Score was 97% for class 0 and 82% for class 1. Additionally, the AUC-ROC score was 0.97, signifying a high capacity to distinguish between the two classes. The model demonstrated robust performance, with considerations focused on maintaining balance and interpretability in the model while optimizing its predictive accuracy.

## 3.5. Model Evaluation and Comparison

In comparing Logistic Regression to Random Forest for the task of churn prediction, several key observations had been made:

1) Random Forest generally exhibited superior performance across a range of metrics, including accuracy, precision, recall, F1-Score, and AUC-ROC. It was particularly effective in capturing complex patterns in the data, leading to higher predictive accuracy.

 ISSN 2723-6471


2) Both models are well-suited for binary classification tasks like churn prediction. However, Random Forest's higher complexity and ability to handle non-linear relationships allowed it to capture more nuanced patterns in the data, resulting in improved predictive performance.
3) Logistic Regression offers better interpretability compared to Random Forest. The simplicity of Logistic Regression makes it easier to understand how each predictor variable contributes to the model's predictions. This interpretability can be crucial in a business context where stakeholders need to comprehend and act upon the model's predictions.

Considering the context of research objectives is essential in choosing between Logistic Regression and Random Forest. If the primary goal is to maximize predictive accuracy and uncover intricate data patterns, Random Forest may be the preferred choice. On the other hand, if the ability to explain model predictions to stakeholders is of paramount importance, Logistic Regression's simplicity and interpretability can be more advantageous.

In the visual comparison of the two machine learning algorithms (shown in Figure 5), Logistic Regression and Random Forest, based on various performance metrics, several key insights emerge. Random Forest consistently outperforms Logistic Regression across these metrics in the context of churn prediction. Random Forest demonstrates higher accuracy, indicating a greater number of correct predictions overall. It also excels in precision, which measures the accuracy of positive predictions, and recall, which evaluates the ability to correctly identify actual positive cases.

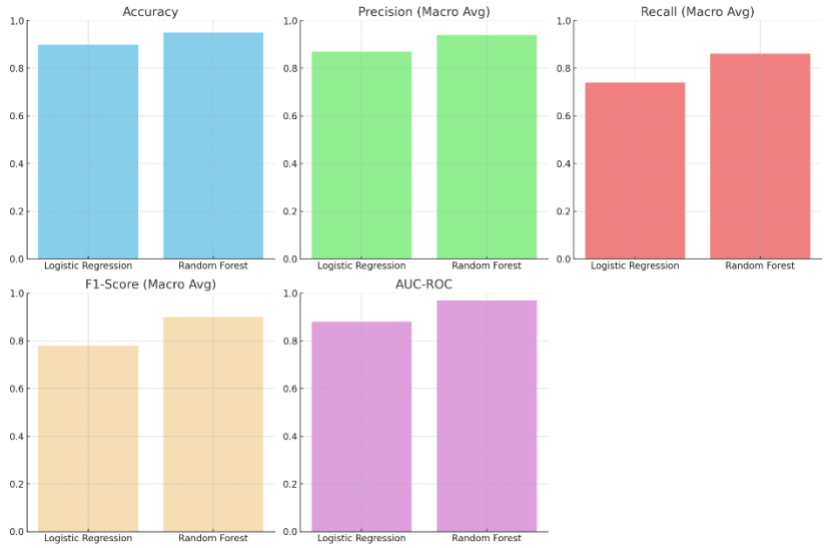


**Figure 5.** Performance Comparison of LR and RF

Additionally, Random Forest achieves a higher F1-Score, which balances precision and recall, indicating a superior ability to manage false positives and false negatives. Moreover, its higher AUC-ROC score suggests that Random Forest is more adept at distinguishing between churned and not churned customers. However, the choice between these models should also consider factors like interpretability, computational cost, and alignment with specific business or research objectives. Therefore, the decision should involve a comprehensive evaluation of these aspects to select the most appropriate algorithm for the task at hand.

## 3.6. In-depth Analysis of Results

The Random Forest model consistently outperformed the Logistic Regression model across multiple performance metrics. This suggests that the dataset contains intricate patterns and interactions among features. The higher performance of Random Forest underscores that customer churn in this context is a complex issue influenced by multiple factors and their interactions, rather than simple, linear relationships.

 ISSN 2723-6471


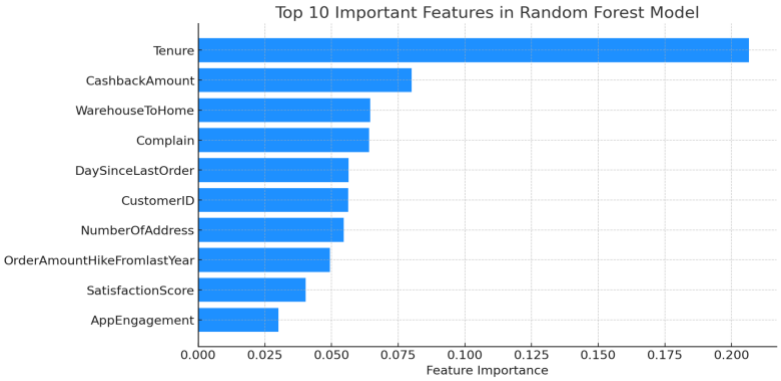


**Figure 6.** Important Features of Random Forest

Regarding the interpretation of important features, the analysis highlights the top 10 most important features as identified by the Random Forest model (shown in Figure 6). These features exert the most significant influence on the model's predictions and likely play a crucial role in determining customer churn. Features ranking high in importance may represent critical aspects of customer behavior and satisfaction. For instance, variables related to customer engagement, such as 'HourSpendOnApp' or 'NumberOfDeviceRegistered,' and transaction history metrics like 'OrderCount' or 'OrderAmountHikeFromlastYear,' can serve as key indicators of churn risk. It's essential to understand how each important feature relates to customer satisfaction and loyalty.

Features related to customer interaction with the digital platform, such as 'HourSpendOnApp' or 'AppEngagement,' provide insights into the correlation between digital behavior and churn. High engagement may generally reduce churn risk, but excessive engagement without corresponding satisfaction could lead to higher churn rates.

## 3.7. Discussion of Findings

In the broader context of understanding social dynamics in the digital era, the results of this study hold significant implications. The superior performance of the Random Forest model compared to Logistic Regression suggests that customer churn prediction in this context is influenced by intricate patterns and interactions among various features. This aligns with the evolving landscape of digital interactions, where customer behaviors are shaped by multifaceted factors rather than simple, linear relationships. Notably, features related to customer engagement and behavior, such as 'HourSpendOnApp' and 'NumberOfDeviceRegistered,' hold paramount importance in predicting churn. These metrics provide valuable insights into customer satisfaction and loyalty, highlighting the significance of digital engagement metrics in the digital era.

Furthermore, the prominence of features related to transactional history, such as 'OrderCount' and 'OrderAmountHikeFromlastYear,' underscores the importance of the customer's shopping experience and historical interactions. In an era where alternatives are readily available, customers' past behaviors significantly influence their future decisions, making variations in spending, order counts, or satisfaction scores indicative of broader trends in customer expectations and market dynamics.

Additionally, the relevance of features associated with customer preferences, like 'PreferredPaymentMode' or 'PreferedOrderCat,' emphasizes the role of personalization in the digital era. Understanding and catering to individual preferences emerge as crucial factors in retaining customers in a competitive digital landscape.

The implications of this study are significant for e-commerce stakeholders and digital strategists. Understanding which features significantly influence churn allows businesses to tailor their customer experience more effectively. For instance, if time spent on the app or engagement metrics are significant predictors, strategies could include enhancing the app's user interface, personalizing content, or providing targeted incentives to increase engagement. Additionally, businesses can proactively engage with at-risk customers through special offers, personalized communication, or loyalty programs, potentially reducing churn rates. Insights into transactional behavior and preferences can guide dynamic pricing strategies and personalized offers, leading to more effective sales tactics and customer satisfaction. Furthermore, the study emphasizes the importance of regularly collecting and acting upon customer feedback to improve satisfaction and retention.

## 4. Conclusion

The results and discussion of this study yield several key takeaways. First, it becomes evident that the Random Forest model outperforms Logistic Regression when it comes to predicting customer churn. This finding underscores the complexity and non-linear nature of the factors that influence churn in the e-commerce context. Furthermore, the study identifies important predictors, such as engagement metrics, transactional history, and customer satisfaction scores, which play a significant role in influencing churn. This highlights the multifaceted nature of customer behavior and emphasizes the importance of taking a comprehensive approach to understand and address customer needs.

Moreover, the study sheds light on the complex social dynamics that characterize customer interactions in the digital era. It reveals that various factors interplay to influence customer decisions, making it essential for businesses to gain a deep understanding of these dynamics to retain customers and enhance satisfaction in a highly competitive environment. The practical implications of these findings are significant for e-commerce stakeholders and digital strategists. Leveraging predictive analytics and tailoring strategies to align with customer needs and behaviors can substantially improve customer retention and engagement.

Future work in this area could expand data sources to include social media sentiment, direct customer feedback, or more granular transactional data for a more comprehensive view of customer behavior. Employing more advanced or alternative modeling techniques, such as neural networks or ensemble methods, may improve predictive performance and insights.

## 5. Declarations

### 5.1. Author Contributions

Conceptualization: P.A.S., U.R., S.C.C., Y.-M.L., and M.H.; Methodology: U.R.; Software: P.A.S.; Validation: P.A.S. and U.R.; Formal Analysis: P.A.S. and U.R.; Investigation: S.C.C.; Resources: Y.-M.L.; Data Curation: S.C.C.; Writing Original Draft Preparation: S.C.C. and M.H.; Writing Review and Editing: S.C.C. and M.H.; Visualization: M.H.; All authors have read and agreed to the published version of the manuscript.

### 5.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

### 5.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

### 5.4. Institutional Review Board Statement

Not applicable.

### 5.5. Informed Consent Statement

Not applicable.

### 5.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## References

[1] M. Leontowitsch, F. Wolf, and F. Oswald, “Digital (in)equalities and user emancipation: Examining the potential of Adorno’s maxim of Mündigkeit for critical intergenerational learning,” *Frontiers in Sociology*, vol. 7, no. 1, pp. 1–10, 2022. doi:10.3389/fsoc.2022.983034.

[2] H. S. Sætra and E. Fosch-Villaronga, “Healthcare digitalisation and the changing nature of work and Society,” *Healthcare*, vol. 9, no. 8, pp. 1007–1017, 2021. doi:10.3390/healthcare9081007.

[3] W. Fan, B. Shao, and X. Dong, “Effect of E-service quality on customer engagement behavior in community e-commerce,”

Frontiers in Psychology, vol. 13, no. September, pp. 1–16, 2022. doi:10.3389/fpsyg.2022.965998.

[4] G. CONOLE, "Developing digital literacies through continuing professional development," *Journal Plus Education*, vol. 19, no. 1/2018, pp. 21–30, 2017. doi:10.24250/jpe/1/2018/gc.

[5] V. Roblek, M. Meško, and I. Podbregar, "Mapping of the emergence of Society 5.0: A bibliometric analysis," *Organizacija*, vol. 54, no. 4, pp. 293–305, 2021. doi:10.2478/orga-2021-0020.

[6] A. Burger, T. Oz, W. G. Kennedy, and A. T. Crooks, "Computational social science of disasters: Opportunities and challenges," *Future Internet*, vol. 11, no. 5, pp. 103–116, 2019. doi:10.3390/fi11050103 .

[7] J. E. Raffaghelli, S. Cucchiara, F. Manganello, and D. Persico, "Different views on digital scholarship: Separate Worlds or cohesive research field?," *Research in Learning Technology*, vol. 24, no. 1, pp. 1–16, 2016. doi:10.3402/rlt.v24.32036.

[8] S. Mohamed Asaad El Banna and N. Makram Labib, "Using big data analytics to develop marketing intelligence systems for commercial banks in Egypt," *MATEC Web of Conferences*, vol. 292, no. September, pp. 1–5, 2019. doi:10.1051/matecconf/201929201011.

[9] X. Lin, X. Wang, and N. Hajli, "Building e-commerce satisfaction and boosting sales: The role of Social Commerce Trust and its antecedents," *International Journal of Electronic Commerce*, vol. 23, no. 3, pp. 328–363, 2019. doi:10.1080/10864415.2019.1619907.

[10] N. Yang et al., "Large-scale crop mapping based on machine learning and parallel computation with grids," *Remote Sensing*, vol. 11, no. 12, pp. 1500–1516, 2019. doi:10.3390/rs11121500.

[11] W. Fu, "Research on the construction of early warning model of customer churn on e-commerce platform," *Applied Mathematics and Nonlinear Sciences*, vol. 8, no. 2, pp. 687–698, 2022. doi:10.2478/amns.2022.1.00016.

[12] K. Błoński, "Dysfunctional customer behavior – A review of research findings," *Acta Scientiarum Polonorum. Oeconomia*, vol. 20, no. 2, pp. 3–10, 2022. doi:10.22630/aspe.2021.20.2.10.

[13] L. C. Siebert, A. R. Aoki, G. Lambert-Torres, N. Lambert-de-Andrade, and N. G. Paterakis, "An agent-based approach for the planning of distribution grids as a socio-technical system," *Energies*, vol. 13, no. 18, pp. 4837–4850, 2020. doi:10.3390/en13184837.

[14] Khalikussabir and A. Waris, "The impact of utilitarian value, hedonic value, and brand image of Modern Coffee Shop City of Malang on customer satisfaction," *Jurnal Ekonomi and Bisnis JAGADITHA*, vol. 8, no. 2, pp. 172–178, 2021. doi:10.22225/jj.8.2.2021.172-178.

[15] M. Pondel et al., "Deep learning for customer churn prediction in e-commerce decision support," *Business Information Systems*, vol. 1, no. June, pp. 3–12, 2021. doi:10.52825/bis.v1i.42.

[16] P. V. Kamble, A. Nair, T. Saini, and G. V. Patil, "Churn prediction in banking sector," *INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT*, vol. 07, no. 04, pp. 1–5, 2023. doi:10.55041/ijsrem18977.

[17] B. Mishachandar and K. Anil Kumar, "Predicting customer churn using targeted proactive retention," *International Journal of Engineering and Technology*, vol. 7, no. 2.27, pp. 69–76, 2018. doi:10.14419/ijet.v7i2.27.10180.

[18] M. Meire, K. Hewett, M. Ballings, V. Kumar, and D. Van den Poel, "The role of marketer-generated content in customer engagement marketing," *Journal of Marketing*, vol. 83, no. 6, pp. 21–42, 2019. doi:10.1177/0022242919873903.

[19] V. Sriharsha and S. Giri Babu, "Customer stress prediction in telecom industries using machine learning," *International Journal of Innovative Research in Engineering & Management*, vol. 9, no. 5, pp. 219–222, 2022. doi:10.55524/ijirem.2022.9.5.31.

[20] I. D. Astuti, S. Rajab, and D. Setiyouji, "Cryptocurrency blockchain technology in the digital revolution era," Aptisi Transactions on Technopreneurship (ATT), vol. 4, no. 1, pp. 9–15, 2022.

[21] A. I. L. Wibowo, A. D. Putra, M. S. Dewi, and D. O. Radianto, "Study of Divergence of Go Public Company's Financial Performance Based on Website Before and After Merger Using Window Period Method TIME Frame 2015-2017," Aptisi Transactions On Technopreneurship (ATT), vol. 1, no. 1, pp. 27–51, 2019.

[22] Y. Gu, T. D. Palaoag, and J. S. Dela Cruz, "Comparison of main algorithms in big data analysis of Telecom Customer Retention," *IOP Conference Series: Materials Science and Engineering*, vol. 1077, no. 1, pp. 0–10, 2021. doi:10.1088/1757-

 

899x/1077/1/012045.

[23] A. K. Ahmad, A. Jafar, and K. Aljoumaa, "Customer churn prediction in telecom using machine learning in Big Data Platform," *Journal of Big Data*, vol. 6, no. 1, pp. 1–24, 2019. doi:10.1186/s40537-019-0191-6.

[24] C. Lukita, M. Hardini, S. Pranata, D. Julianingsih, and N. P. L. Santoso, "Transformation of Entrepreneurship and Digital Technology Students in the Era of Revolution 4.0," Aptisi Transactions on Technopreneurship (ATT), vol. 5, no. 3, pp. 291–304, 2023.

[25] N. M. N. Febrianti and G. S. Darma, "Millennials' Intention to Invest through Securities Crowdfunding Platform," Aptisi Transactions on Technopreneurship (ATT), vol. 5, no. 1, pp. 19–30, 2023.

[26] A. Pambudi, N. Lutfiani, M. Hardini, A. R. A. Zahra, and U. Rahardja, "The Digital Revolution of Startup Matchmaking: AI and Computer Science Synergies," in 2023 Eighth International Conference on Informatics and Computing (ICIC), IEEE, 2023, pp. 1–6.

[27] P. Rashi, A. S. Bist, A. Asmawati, M. Budiarto, and W. Y. Prihastiwi, "Influence of post covid change in consumer behaviour of millennials on advertising techniques and practices," Aptisi Transactions on Technopreneurship (ATT), vol. 3, no. 2, pp. 201–208, 2021.

[28] L. Gan, "XGBoost-based e-commerce customer loss prediction," *Computational Intelligence and Neuroscience*, vol. 2022, no. July, pp. 1–10, 2022. doi:10.1155/2022/1858300.

[29] M. Upreti, C. Pandey, A. S. Bist, B. Rawat, and M. Hardini, "Convolutional neural networks in medical image understanding," Aptisi Transactions on Technopreneurship (ATT), vol. 3, no. 2, pp. 120–126, 2021.