VFAST Transactions on Software Engineering
https://vfast.org/journals/index.php/VTSE@ 2024, ISSN(e): 2309-3978, ISSN(p): 2411-6246
Volume 12, Number 4, October-December 2024 *pp:70-86*

Leveraging Machine Learning And Deep Learning Models for Proactive Churn Customer Retention

Hira Farman [1*], **Samar Raza Talpur** [2], **Usman Amjad** [2], **Govari shankar** [3], **Umm e Laila** [4], **Lubaba Naseem** [5]

[1]Department of Computer Science, Iqra University, Karachi, Pakistan; [2]Department of Computer Science, Sukkur IBA University, Sindh, Pakistan; [3]Department of Computer Science & Information Technology TIEST Constituent Institute of NED University Karachi, Pakistan; [4]Department of Computer Science, Institute of Business Management Karachi, Pakistan ; [5]Department of Computer Science, Iqra University, Karachi Pakistan

Keywords: *Churn, Prediction, Binary Classification, Machine Learning, Models, Pre-Processing, Data Driven, Deep Learning, Retention, Machine learning, Deep learning, LSTM, CNN, ANN, Retention, Customer Analysis.*

Journal Info:
Submitted:
October 06, 2024
Accepted:
November 17, 2024
Published:
November 23, 2024

Abstract

Customer attrition is especially an issue in industries such as retail, banking, and telecommunications where customer acquisition costs are significantly higher than the costs of retaining repeat customers. The customer lack of interest is now predictable through machine learning models, and deep learning has become instrumental in early intervention for retention. In order to assess the quality of churn prediction, the study tests six basic machine learning techniques: random forest, logistic regression, and the k-nearest neighbors method, as well as four deep learning techniques: long short term memory (LSTM), bidirectional LSTM, convolutional neural networks (CNN), and artificial neural networks (ANN). The performance of the model is then assessed via the evaluation matrices, including the accuracy, precision, recall, and F1-score from the customer's behavioral data after feature extraction from large datasets. The study reveals that DL models offer improved handling of the churn and non-churn customer classification and Random Forest as well as other ML models comparable accuracy. This research can conclude that LSTM and ANN models outshine in actual-world churn prediction circumstances, especially when long-term consumer behavior evaluation is required. To enhance the current outcomes of a given prediction model, this research focuses on data preprocessing and the utilization of bootstrapping, feature extraction, and the combination of multiple models. The implications of the study provide specific practical recommendations for firms to effectively manage customer churn and increase customer retention by employing data-dealing techniques.

***Correspondence author email address:** hira.farman@iqra.edu.pk
DOI: 10.21015/vtse.v12i4.1928

1 Introduction

Sustaining a customer base is crucial for maintaining steady growth and profitability in the fiercely competitive world of modern business. In industries such as banking, retail, and telecommunications, where keeping existing consumers is more economical than attracting new ones, customer churn and the loss of consumers to competitors is a major problem. Machine learning has developed into an effective method for churn prediction with the rise of data-driven methodologies, allowing companies to promptly deploy retention initiatives.

Particularly in the telecom industry, methods like Random Forest, Logistic Regression, and ensemble models like XG-Boost have demonstrated great accuracy in churn prediction [1] [2] [3]. Bi-LSTM and CNN are two examples of hybrid deep learning models that have further improved prediction skills, demonstrating the promise of cutting-edge methods [4] . By combining deep learning, machine learning, and sophisticated preprocessing approaches to create a reliable churn prediction model, this work seeks to advance previous studies. By analyzing the patterns and behaviors of customers, the model works towards assisting industries in improving customer retention and decreasing attrition. This strategy stretches the limits of churn prediction models for useful business applications while integrating state-of-the art techniques from earlier research [5] [6]. A similar telecom study that used machine learning classifiers like Random Forest, KNN, and Decision Trees achieved 99% accuracy, precision, and recall for customer churn, which demonstrates the reliability of these models for prediction [7, 8].In another examination of churn prediction in the telecom industry through regression analysis and machine learning, it was noted that analytics could help enhance retention strategy [9].

1.1 Problem Statement

As the world becomes business-orientated, it is becoming more important than ever to keep customers in the competitive business marketplace of today. The rate at which the customers leave a service of a particular company, in other words, customer churn, presents serious threats to organizations in a variety of industries. Companies can reduce revenue loss and increase customer satisfaction by implementing retention measures by identifying and anticipating churn patterns. By creating a concrete prediction model that makes use of machine and deep learning and machine forecasting, clients fall out rate with the company. The study aims to deliver actionable insights to help organizations optimize their marketing efforts, improve customer engagement, and eventually cultivate long-term customer loyalty by evaluating historical data, including demographic information, consumer interactions, and behaviors.

1.2 Customer Churn

Customer churn [1, 2] It is a term that describes the behavior of customers that cease their relationship with a business, which can have significant financial implications. Customer leaving the service for a better alternative translates to loss of revenue, and it decreases profitability. But every problem has solutions, so what if there was a way to predict this churn and take immediate measures to retain valuable customers? This is where the concept of churn prediction comes in, which is the process of identifying customers who are likely to stop using a company's products or services, often by cancelling a subscription or not making repeat purchases [10].

1.3 Research Questions

- **RQ1:** Comparison of Machine learning and deep learning for the analysing customer chrun prediction lie at?
- **RQ2:** The impact of data preprocessing on the performance of ML and DL models for customer churn prediction?
- **RQ3:** Which specific model, between Machine Learning and Deep Learning approaches, provides the best balance of accuracy, precision, recall, and F1-score for customer churn prediction the dataset?

2 Literature Review

2.1 Related work

Several studies have shown how crucial machine learning methods are for predicting client departures,

especially in the telecom sector. The capacity of algorithms like Random Forest, Logistic Regression, and k-Nearest Neighbors (kNN) to handle intricate patterns in huge datasets makes them widely used. For example, one study uses Random Forest, Logistic Regression, and kNN that achieves a 96.3% cross-validation score [1]. Another study used nine months of historical customer data to identify the optimal machine learning strategies for customer retention in a competitive environment [2] by focusing on early churn detection. Customer retention is more cost-effective than obtaining new customers in the telecom industry due to the massive amounts of data generated on a daily basis. This has led to further research into algorithms such as Random Forest, KNN, and Decision Tree, with Random Forest achieving 99% accuracy, precision, and recall in one study [10]. Several hybrid deep learning models have been investigated; one of these models, which combines CNN and Bi-LSTM, achieved 81% accuracy [4].Beyond telecom, the retail sector has also adopted machine learning techniques for churn prediction. Research uses the RFM model and K-means clustering, which successfully identify the customers at risk, aiding in retention strategies [11]. The banking sector has also applied various models such as KNN, SVM, Decision Trees, and Random Forest, with Random Forest emerging as the most effective and accurate in handling imbalanced datasets [12] [13]. Another telecom study highlighted the superiority of decision trees for churn prediction [14]. Ensemble methods like Gradient Boosting Machines (GBM) and XGBoost have outperformed others, with Ada boost and XGBoost achieving the highest accuracy in one telecom study at 81.71% and 80.8%, respectively, with an AUC score of 84% [3]. Other research improvised churn prediction through advanced data preprocessing, achieving better accuracy and insights for managers [5]. In terms of model architecture, Churn Net combined several techniques, achieving an accuracy rate of 97.52% [6]. Research focusing on American businesses highlighted Random Forest as the top-performing model with 96.25% accuracy, providing actionable insights for targeted retention strategies [15]. XGBoost and Gradient Boosting were also shown to improve churn prediction accuracy significantly, especially when combined with data balancing techniques [16]. A hybrid neural network model combining SOM and ANN demonstrated higher accuracy than single models in the telecom industry [7]. Customer churn in subscription-based services has also been studied, with churn prediction models emphasizing personalized retention strategies. An accuracy of more than 80% was found in research on churn likelihood, highlighting the significance of churn prediction for revenue stability and corporate decision-making [17]. Customer attrition has become a major concern due to the rise in subscription-based services. Different research employed eleven classifiers, including Random Forest and KNN, to predict churn; after hyper parameter adjustment, Random Forest achieved the greatest Area under the Curve (AUC) score of 85% [18]. Furthermore, decision trees outperformed logistic regression with 99% accuracy in a comparison of random forest and decision tree models for churn prediction, highlighting the necessity of using several models in churn analysis [19]. A unique churn prediction model for the e-commerce industry that combined the LRFM model with the K-means algorithm produced accuracy rates of 99% for total churn and 98% for partial churn [20]. Online firms need to understand the idea of "churn," which refers to customers dropping out, and predictive models are necessary for revenue forecasting and retention tactics [21]. Another telecom research project that included machine learning classifiers such as Random Forest, KNN, and Decision Trees obtained 99% accuracy, precision, and recall, indicating the great accuracy of these models in forecasting customer turnover [8, 9, 22, 23].

2.2 Bridging Gaps in Customer Churn Prediction

A Comparison with Prior Studies Existing research in customer churn prediction has extensively explored machine learning and deep learning models, highlighting their effectiveness across various industries such as telecom, retail, and banking. Studies like [1] achieved high accuracy with Random Forest in the

telecom sector, while [11] utilized RFM and K-means clustering for segmentation in retail but lacked predictive modeling. Similarly, deep learning approaches, such as those in [4], demonstrated the potential of hybrid models like CNN with Bi-LSTM, though they often faced computational challenges and imbalanced datasets.

-The findings of the proposed study demonstrate how well different machine learning (ML) and deep learning (DL) algorithms perform in predicting client attrition. Random Forest has the highest accuracy in the ML category (79.02%), but its recall of 0.47 indicates that it has trouble detecting churners (the minority class). The significance of feature selection and data pretreatment was highlighted in the study [1] conducted by India in 2021, when Random Forest obtained 96.3% accuracy with greater recall. With a balanced F1-score of 0.59 and an accuracy of 77.97%, KNN performs admirably but has lesser precision and recall than Logistic Regression (78.75% accuracy and F1-score of 0.56). This is somewhat in line with earlier research. This study emphasizes the need for advanced preprocessing techniques and model tuning to address these limitations and enhance churn prediction accuracy

-ANN has the best accuracy (79.14%) among the deep learning models, but it performs worse in precision and recall (0.71 and 0.67, respectively), which results in a lower F1-score (0.68). This outcome is in line with other research where deeper architectures and hybrid models, such Bi-LSTM (79.03% accuracy), have done better than more straightforward models like ANN. Near-identical performance is shown by Bi-LSTM and LSTM, with Bi-LSTM slightly exceeding in terms of precision, recall, and F1-score. This is consistent with research that combined CNN and Bi-LSTM, which found a similar balance in churn prediction accuracy.

-It is crucial to draw attention to the constraints of feature selection and data balancing, even though the suggested study shows good results overall, particularly with deep learning models like LSTM and Bi-LSTM. Furthermore, the work may offer further light on how models function with unbalanced datasets, a problem frequently mentioned in the literature. The proposed study could improve its contribution to churn prediction research by providing a more thorough comparison with these established results.

3 Comparative Analysis of Past Studies for customer churn retention

The table 1 summarizes the key elements of a comparative study by country, year, algorithms, datasets, target features, accuracy and limitation of the past work.

4 Gap Analysis

Some work had already been done on churn analysis. Some of them are using company dataset, telecom, banking, and utilizing machine learning and deep learning technique.

4.1 Feature Table

Since each feature is responsible for determining specific points in the customer churn rate and improving retention strategies, a variety of research papers that presented the list of features that have been discussed were examined in order to bring excellence to the research work. This was done in order to analyze the various data points that were necessary to make concrete churn prediction. The various elements that were taken from the previous literature paper and its summary are described in Table 2. It displays various useful characteristics and work in an accurate manner.

4.2 Source Table

Table-3 represents the list of 10 selected research papers, which have been read and noted to collect maximum useful information about customer churn prediction employing different machine learning approaches. These sources offer an informed fashion of different approaches and strategies used in churn prediction.

4.3 Mapping Table

TABLE-4 represents the summarized features results in tabular form of relation between the research papers and the features used in other previous works and investigate in our work. It also aids in identifying what feature our research paper has. It combines

Ref Country Year	Algorithms	Dataset	Target Feature	Accuracy	Result	Limitations
[1] India 2021	Logistic Regression, Random Forest, K-nearest Neighbors (KNN)	The dataset used consists of customer records, including features like mobile number, recharge amounts, call usage, and other relevant attributes from June to September 2014.	Telecom sector	Random Forest: 96.3%, KNN: 88.8%, Logistic Regression: 81.72%	Random Forest model performed the best with a cross-validation score of 96.3%. The paper concludes that the proposed churn model produced better results using machine learning techniques.	The paper suggests the need to reduce further features to obtain better accuracy and to introduce more machine learning models for improved performance.
[10] India 2024	Random Forest, K-nearest Neighbors (KNN), Decision Tree	Customer data from the telecom industry.	Telecom sector	Random Forest: 99%, Precision: 99%, Recall: 99%, Overall Accuracy: 99.09%	The Random Forest classifier achieved the highest accuracy (99%) among the tested algorithms. The study emphasizes that analyzing customer churn data helps telecom companies to take timely actions to prevent churn and retain customers, thus reducing profit loss.	The paper mentions that current models face difficulties and hurdles, especially in feature selection, as many information-rich features are neglected during this process. Also, existing models have undesirable results due to the statistical methods used.
[11] USA 2021	RFM (Recency, Frequency, Monetary) model, K-means clustering	The dataset contains transactional data from a UK-based online retail gift store, with records of transactions between December 1, 2010, and December 1, 2011.	Retail sector, specifically online retail businesses.	—-	The study segments customers into 3 clusters based on their RFM scores. Cluster 2 contains the best customers, Cluster 1 contains loyal customers, and Cluster 0 contains at-risk customers. This segmentation allows businesses to identify at-risk customers and apply appropriate retention strategies.	The study is limited to transactional data from a single online retail store. The analysis focuses only on customer segmentation and does not explore other potential predictive models. The paper does not address real-time data processing or dynamic customer behavior changes.
[14] Pakistan 2013	Regression Analysis (Linear and Logistic Regression), Artificial Neural Networks (ANNs), K-Means Clustering, Decision Trees (CHAID, Exhaustive CHAID, CART, QUEST)	Traffic data from a telecom operator with 106,000 customers, including traffic type, destination, rate plan, loyalty, and usage behavior.	Telecom sector	Exhaustive CHAID: Overall accuracy of 70% (recall for churners: 60.5%)	The Exhaustive CHAID algorithm provided the best results with an overall accuracy of 70% and correctly identified 60.5% of churners. The study demonstrated that re-sampling methods effectively addressed the class imbalance problem in the dataset.	The study was limited to a dataset of 106,000 customers and a three-month period. The algorithms' performance might vary with different datasets or longer time periods. Logistic regression and ANNs performed poorly in identifying churners compared to Decision Trees.
[13] Banking sector 2020	k-Nearest Neighbor (KNN), Support Vector Machine (SVM), Decision Tree, Random Forest	The dataset used is from Kaggle, containing information on 10,000 bank customers, with 13 feature vectors and a binary target variable indicating whether the customer has exited the bank.	Banking sector	KNN: 81.65%, SVM: 79.63%, Decision Tree: 78.99%, Random Forest: 85.18% (95.74% after oversampling)	The Random Forest model, particularly after applying oversampling to balance the dataset, achieved the highest accuracy of 95.74%. This model was found to be the most effective in predicting customer churn compared to KNN, SVM, and Decision Tree models.	The study used a relatively small and highly imbalanced dataset. SVM did not perform well with oversampling. Feature selection methods did not improve the performance of tree-based models (Decision Tree and Random Forest).
[12] India 2020	Random Forest, Generalized Linear Model (GLM), Decision Tree, XG-Boost, Artificial Neural Network (ANN)	The dataset consists of savings bank customers' data, including demographic information, transaction data, and other relevant features from the State Bank of India.	Banking sector, specifically savings bank account customers.	Random Forest: 78% accuracy	The Random Forest model outperformed other models with an accuracy of 78%, and key variables were identified for predicting customer churn. The model can help commercial banks deploy targeted campaigns to retain high-value customers.	The model's performance might be affected by the quality and completeness of the data. The study used an under-sampling technique. The findings are specific to the dataset from the State Bank of India and may not generalize to other banks or regions.

Table 1. Comparison Analysis of the past studies

Table 2. Feature Table

F#	Feature Name	Description
F1	Customer Demographics	Includes age, gender, location, and other demographic details used to understand customer profiles [1] [11] [24].
F2	Transaction History	Historical data on customer transactions, including purchase frequency, monetary value, and recency [11] [5] [20].
F3	Service Usage Patterns	Data on how customers use the service, including frequency, duration, and type of usage [1] [4] [7].
F4	Customer Interactions	Records of customer interactions with the company, including support tickets, chat logs, and feedback [13] [15] [25].
F5	Subscription Details	Information about customer subscriptions, including start date, end date, subscription type, and renewal history [10] [18] .
F6	Contract Information	Details about customer contracts, such as contract length, terms, and any changes made to the contract [16] .
F7	Payment History	Records of customer payments, including payment frequency, method, and any payment issues [12] [17] [26].
F8	Customer Feedback	Data from customer surveys, reviews, and ratings, providing insights into customer satisfaction and issues [24] [27] .
F9	Loyalty Programs	Participation and engagement in loyalty programs, including points earned, rewards redeemed, and tier status [11] [8] .
F11	Customer Support	Data on customer support interactions, including frequency of contact, resolution time, and customer satisfaction [25] [28].
F12	Usage Frequency	How often the customer uses the product or service, measured over specific time periods [1] [10] [5].

Table 3. Source Table

F	Feature Name	Description
S1	[1]	CUSTOMER CHURN PREDICTION
S2	[2]	Customer churning analysis using machine learning algorithms
S3	[10]	Customer churn prediction in telecom sector using machine learning technique
S4	[4]	Customer churn prediction using composite deep learning technique
S5	[11]	Retail customer churn analysis using RFM model and K-means clustering
S6	[12]	churn prediction for savings bank customers: A machine learning approach
S7	[13]	Machine learning-based customer churn prediction in banking
S8	[3]	Customer churn prediction system : A machine learning approach
S9	[6]	Churn Net : Deep Learning Enhanced Customer Churn Prediction in Telecommunication Industry
S10	[16]	Enhancing customer retention in telecom industry with machine learning driven churn prediction

Table 4. Mapping Table

Feature	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Our Work
F1	✓	X	✓	X	✓	X	✓	✓	✓	X	✓
F2	X	✓	X	X	X	X	X	X	✓	X	✓
F3	✓	✓	✓	X	X	✓	✓	X	X	✓	X
F4	X	X	✓	✓	✓	X	X	✓	✓	X	X
F5	X	X	X	X	X	✓	X	X	X	X	✓
F7	X	✓	X	✓	✓	X	✓	✓	✓	X	✓
F11	✓	X	X	X	X	X	X	X	✓	X	X

and contrasts our work with various structural characteristics of various research papers to demonstrate all correlation between accomplishments of multiple researches. It sets up relationships and differences to previous research works and samples our work to different aspects in order to emphasize patterns, trends and correlation between the achievements of multiple machine learning and deep learning research studies.

5 Methodology

The methodology for predicting customer churn involves several key steps, as shown in figure 1. It is started by collecting and preprocessing the data, followed by the application of machine learning algorithms, and finally, model evaluation. The first steps involve gathering a customer dataset with characteristics including customer demographics shown in figure 1 and transaction histories.

Preprocessing is then carried out on the raw data, which includes cleaning the data to deal with outliers and missing information, then moving on to selecting relevant features helpful for perdition. Then, to estimate customer churn, machine learning models like Random Forest (RF), Logistic Regression (LR), and K nearest neighbors (KNN) are utilized, along with deep learning models like LSTM, Bi-LSTM, CNN, and ANN. Performance indicators like F1-score, accuracy, precision, and recall are used to evaluate the performance of models, and the results are displayed in the form of graphs and charts. These insights are used to help develop customer retention strategies that center on at-risk customers and tailored interventions. In the end, these methods are constantly monitored and improved with continuous feedback loops, which also determine if a customer has been maintained or has left. This proactive plan enables companies to act quickly to lower the customer dropout ratio and increase client retention.

5.1 Dataset Description

This study utilizes the Telco Customer Churn dataset obtained from Kaggle, which has a total of 7,043 customer records with 21 attributes including customer ID, gender, senior citizen status, tenure, service usage (e.g., phone and internet services), contract type and payment methods shown in table 5. The target variable named as churn which indicates whether a customer churned or not.

5.2 Data Pre-Processing

The data preprocessing in this model was specially designed to ensure data quality and readiness for machine learning models. It begins by importing key libraries like pandas for data handling and NumPy for numerical operations. After loading the dataset into a pandas DataFrame, unnecessary columns, such as unique identifiers (e.g., customer_id), are dropped to remove irrelevant information. The model is then designed for data cleaning which corrects any feature errors, such as incorrect data types, ensuring all columns are appropriately formatted, leaving the total data of 20 columns and 7023 rows. Next, quick data visualization steps are included to explore the relationship between the target variable (churn) and other features, providing early insights into patterns and distributions shown in Figure 2 and Figure 3. This exploratory analysis helps to understand the data before diving into more advanced processing.

In terms of data transformation, the script applies both ordinal and one-hot encoding to handle categorical variables. Ordinal encoding is used for features with a clear order, while one-hot encoding is applied to nominal variables, converting them into binary columns. Continuous numerical features are then scaled using Min-Max scaling shown in eq. (1)(, ensuring all values fall within a consistent range, which is crucial for algorithms like K-Nearest Neighbors (KNN) and neural networks.

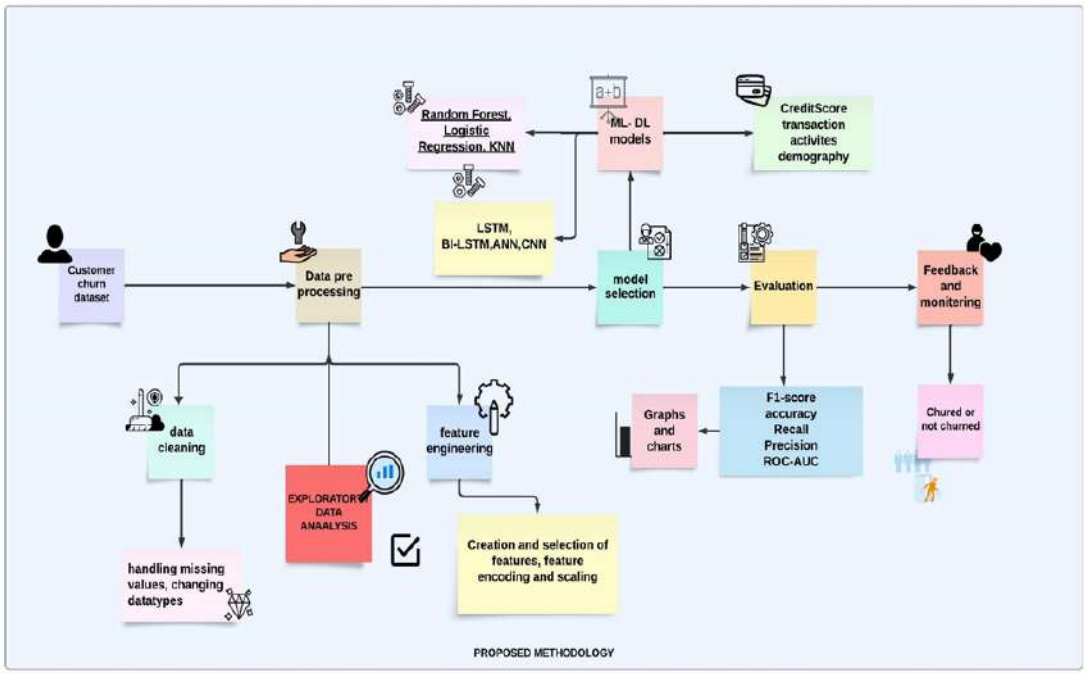

Figure 1. Overview of Methodology

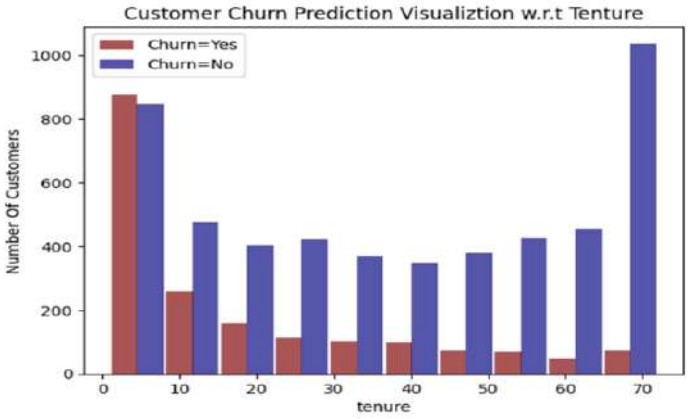

Figure 2. Visualization of number of customers with Tenure

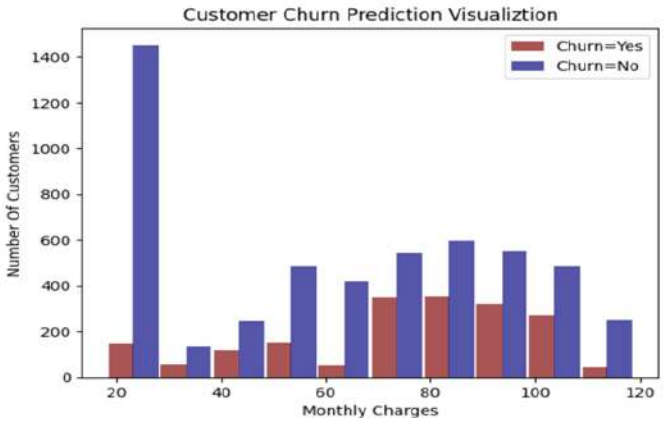

Figure 3. Visualization of target variable (churn)

The data is also visualized to detect which distribution the data follows in order to select which scaling techniques should be used shown in figure 4.

$$minmax = \frac{\min(x)}{\max(x) - \min(x)} (1)$$

As churn data is imbalanced, with a higher number of non-churn cases, this experimental study employs the SMOTE technique to balance the data by over-sampling the minority class. This ensures that the model does not become biased toward the majority class, allowing it to effectively learn from both churn and non-churn cases. Figure 4 illustrates the data distribution with and without balancing. The graph in Figure 4 shows the distribution of the variable "Total Charges," overlaid with a normal distribution curve. The horizontal axis represents the total charges, ranging from 0 to approximately 8000, while the vertical axis shows the density, reflecting the frequency of occurrence for various total charge values. The light purple histogram bars depict the frequency of

Table 5. Dataset attributes description

s.n	attributes Name	Description
1	Customer ID	It is a unique string combination assigned to each customer.
2	gender	It describes the gender of customers either they be male or female.
3	Senior Citizen	This feature identifies customer as senior citizen with 1 as Yes and 0 as No.
4	Partner	This feature represents customer partner (spouse)
5	Dependents	It is a feature that Indicates if the customer has dependents (Yes/No)
6	tenure	It is the period of time the client has been with the business.
7	Phone Service	This shows whether or not the client has phone service.
8	Multiple Lines	This shows whether the client has more than one line. (No phone service, yes, or no)
9	Internet Service	It is the kind of internet service that the client utilizes (fibre optic, DSL, etc.).
10	Online Security	This feature shows whether or not the consumer has internet service or an online security add-on.
11	Device Protection	It shows whether or not the consumer has internet service and device protection add-on.
12	Tech Support	This feature indicates if the customer has tech support add-on (Yes/No/No internet service)
13	Streaming TV	It indicates if the customer has streaming TV service (Yes/No/No internet service)
14	Streaming Movies	This function shows whether or not the user has internet service and can stream movies.
15	Contract	It depends on the customer's contract type (month-to-month, one-year, or two-year).
16	Paperless Billing	This feature shows whether or not the customer is signed up for paperless billing.
17	Payment Method	It is the payment mechanism that the consumer uses (credit card (automated), bank transfer (automated), electronic cheque, or mailed cheque).
18	Monthly Charges	This feature tells the monthly charges incurred by the customer
19	Total Charges	This feature gives out the total charges accumulated by the customer
20	Churn	The target variable that outputs the churned and non-churned values (Yes/No) is this feature.

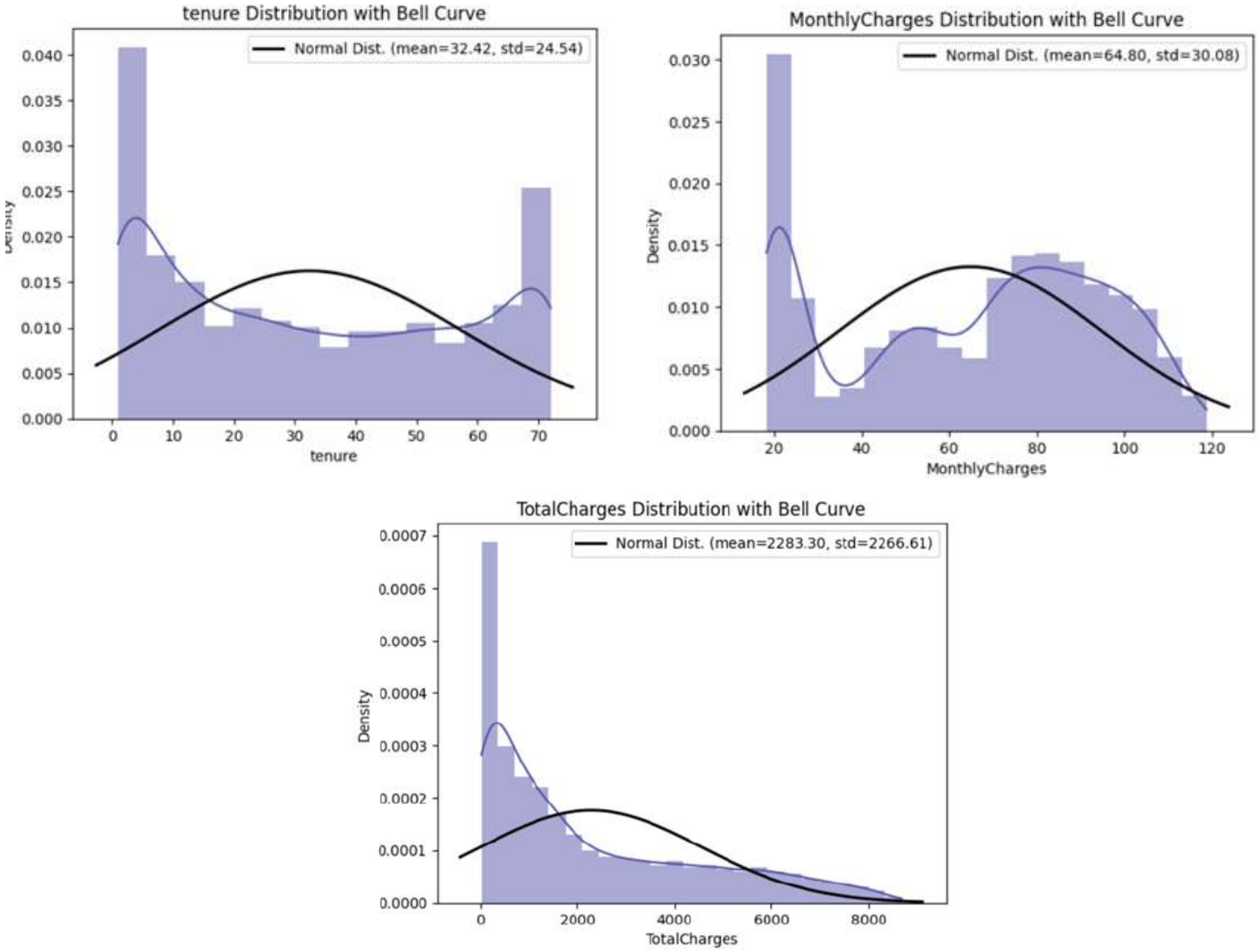

Figure 4. Distribution of Data

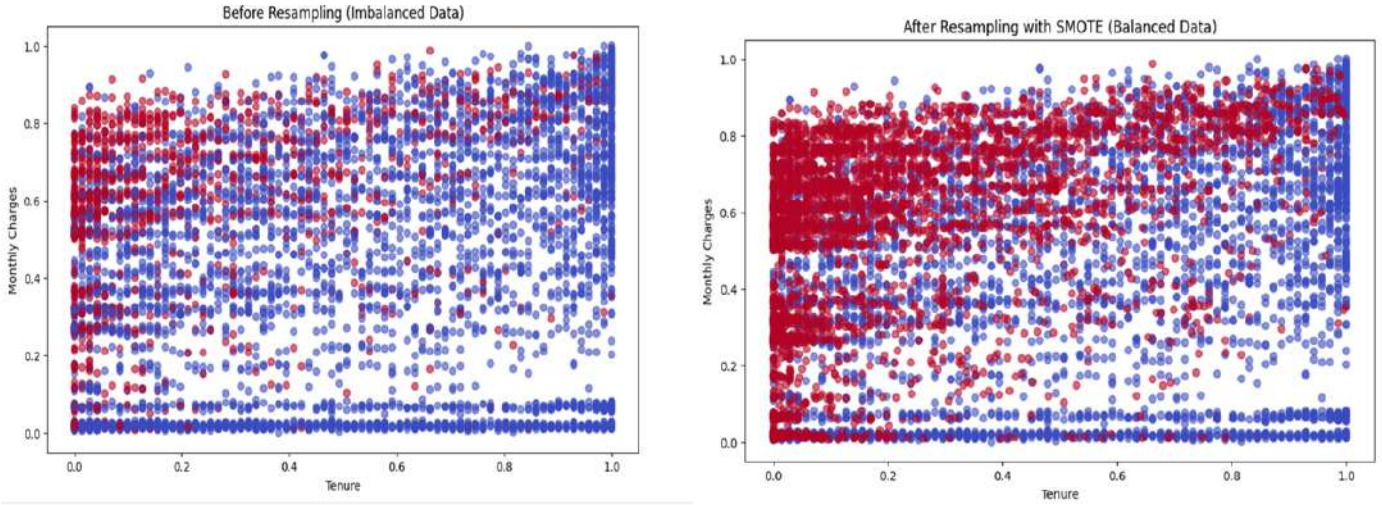

Figure 5. Scatter Plot of data before and after resampling

different charge amounts. As seen in the graph, there is a distinct peak at the lower end, indicating a large concentration of data points within this range. The actual normal distribution, represented by the black curve, spans a range from approximately 1957.66 to 2609.12, with an average value of 2283.30 and a standard deviation of 2266.61.The data reflects the extent to which the Total Charges variable approximates a normal distribution.

However, the elongated tail on the right side indicates a positive skew, suggesting significant variability in the charges, with the majority of values falling in the lower range. This plot effectively highlights the outliers and skewness in the Total Charges distribution, providing valuable context for subsequent analysis and decision-making

Figure 5 shows the data distribution before and after applying SMOTE for resampling. The top scatter plot (Before Resampling) highlights the class imbalance, with a larger number of non-churn cases (in blue) compared to churn cases (in red). After SMOTE is applied (bottom scatter plot), the data is balanced, showing an equal distribution of both churn and non-churn cases. This balanced dataset allows the model to learn effectively from both classes, minimizing the risk of bias toward the majority class.

5.2.1 Test and Train Split

The dataset is split into training and testing sets using the 'train_test_split()' method with a ratio of 80-20 from scikit-learn, allowing for an unbiased evaluation of model performance.

5.3 Models utilized in study

This study discusses the use of numerous algorithms used for churn analysis including both machine learning and deep learning algorithm, testing them out with various feature to obtain maximum accuracy. The description for each model is given below:

5.3.1 Machine Learning Models

Machine learning (ML) is a branch of artificial intelligence (AI) and computer science that focuses on the using data and algorithms to enable AI to imitate the way that humans learn, gradually improving its accuracy [51]. There are various models that are used for prediction purposes but in this study three models are utilized K Nearest Neighbors (KNN), Random Forest (RF), and Logistic Regression (LR).

A. k-Nearest Neighbors (k-NN)

In churn prediction, k-NN can be utilized to classify customers as "churn" or "not churn" based on the characteristics of their closest neighbors figure 6. The model relies on historical customer data, such as usage patterns, payment history, and demographic information. The Distance Metric: Commonly uses Euclidean distance to find the nearest neighbors shown in eq. (2):

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{d}(x_{ik} - x_{jk})^2} \qquad (2)$$

The Voting Mechanism: For a new customer instance (x), where yi is the churn label shown in eq. (3).

$$class(x) = \arg\max_{c \in C} \sum_{i=1}^{k} I(y_i = c) \quad (3)$$

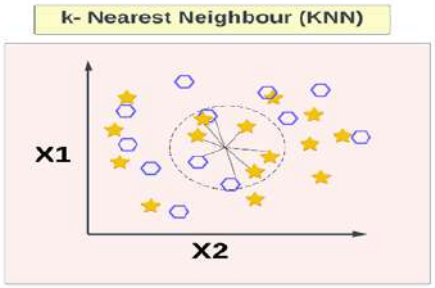

Figure 6. Nearest Neighbor Representation

B. Logistic Regression:

Logistic regression is effective for binary classification problems like churn prediction. It estimates the probability that a customer will churn based on their features (e.g., service usage, customer support interactions) Figure 7. The Logistic Function is shown in eq. (4):

$$P(Y = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n)}} \qquad (4)$$

The Cost Function (Log-Likelihood) is shown in eq. (5):

$$L(\beta) = \sum_{i=1}^{m} \left[y_i \log(P(y_i \mid x_i)) + (1 - y_i) \log(1 - P(y_i \mid x_i)) \right] \quad (5)$$

Where y_i is the actual churn label.

C. Random Forest:

Random Forest helps improve the robustness of churn predictions by combining multiple decision trees Figure 8. Each tree is trained on a random subset of data, allowing for diverse decision-making based on various customer characteristics. The Prediction from Trees

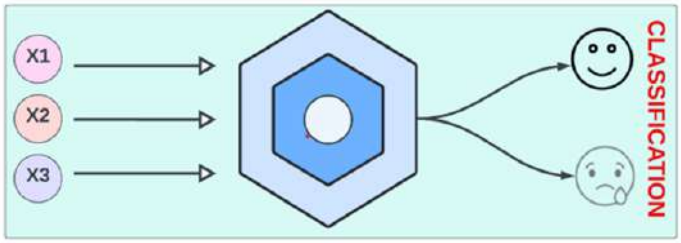

Figure 7. Logistic Regression classification

for a new customer is shown in eq. (6):

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} f_t(x) \tag{6}$$

The Feature Importance can be calculated using the Gini impurity or mean decrease in accuracy.

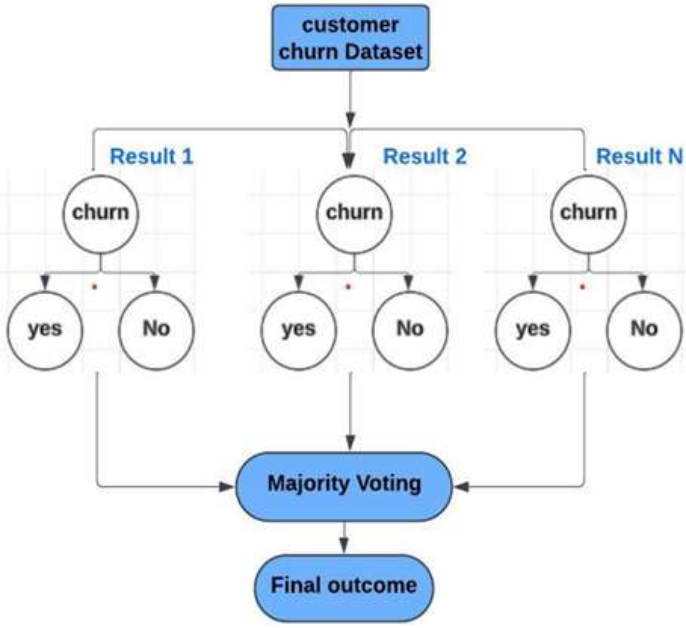

Figure 8. Random Forest Tree Structures

5.3.2 Deep Learning Models:

A subset of machine learning techniques based on neural networks and representation learning is called deep learning. The focus of the field, which draws inspiration from biological neuroscience, is "training" artificial neurons to process data by stacking them in layers. In this work, the following deep learning methods are applied.

Long Short-Term Memory (LSTM):

LSTMs are suitable for churn prediction, especially when analyzing time-series data related to customer behavior over time (e.g., monthly usage metrics). They can capture long-term dependencies in customer activity leading up to churn. The LSTM Cell Structure consists of forget, input, and output gates Figure 9.

The Cell State Update is shown in eq. (7):

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \tag{7}$$

The Hidden State Update is shown in eq. (8):

$$h_t = o_t \cdot \sigma(C_t) \tag{8}$$

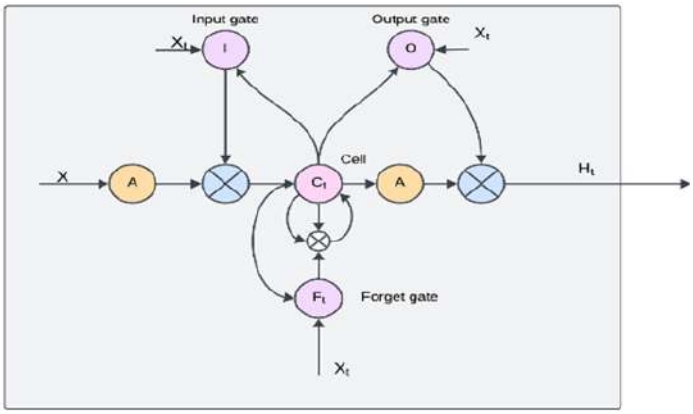

Figure 9. Structure of LSTM

Bidirectional LSTM (BI-LSTM):

BI-LSTM enhances LSTM capabilities by processing input sequences in both directions shown in Figure 10, making it powerful for churn prediction by considering not only past behavior but also future trends in customer activity.

The Forward LSTM eq. is shown in (9):

$$h_t^{\text{Forward}} = \text{LSTM}(x_t, h_{t-1}^{\text{Forward}}) \tag{9}$$

The Backward LSTM is shown in eq. (10):

$$h_t^{\text{Backward}} = \text{LSTM}(x_t, h_{t+1}^{\text{Backward}}) \tag{10}$$

The Final Output is shown in eq. (11):

$$h_t^{\text{Final}} = [h_t^{\text{Forward}}, h_t^{\text{Backward}}] \tag{11}$$

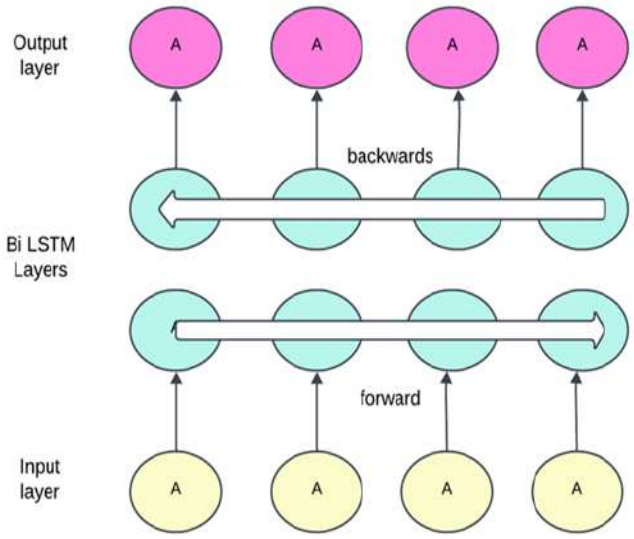

Figure 10. Structure of Bi-Directional LSTM

Convolutional Neural Network (CNN):

While CNNs are typically used for image processing, they can also be adapted for churn prediction by analyzing structured customer data and the neural structure shown in Figure 11. For example, they can extract patterns from customer interaction sequences.

The Convolution Operation: For an input feature matrix X and filter K is shown in eq. (12):

$$Y[i,j] = (X * K)[i,j] = \sum\sum X[m,n]K[i-m, j-n] \quad (12)$$

The Activation Function: Commonly uses ReLU shown in eq. (13):

$$f(x) = \max(0, x) \quad (13)$$

The Pooling Operation: Reduces dimensions shown in eq. (14):

$$Y[i,j] = \max_{m,n \in \text{pool}} X[m,n] \quad (14)$$

D. Artificial Neural Network (ANN)

Artificial Neural Networks (ANNs) can model complex relationships in customer data to predict churn. They consist of multiple layers of neurons, allowing for non-linear relationships to be learned effectively, as shown in Figure 12.

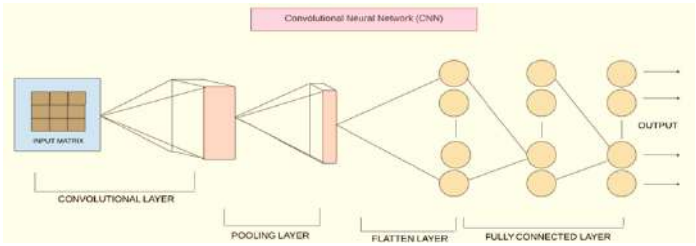

Figure 11. Structure of Convolutional Neural Network

The Forward Propagation: For a network with n input:

$$z^{(l)} = W^{(l)}a^{(l-1)} + b^{(l)} \quad (1)$$

$$a^{(l)} = \sigma(z^{(l)}) \quad (2)$$

The Loss Function: For churn prediction, Mean Squared Error (MSE) can be used:

$$L(y, \hat{y}) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \quad (3)$$

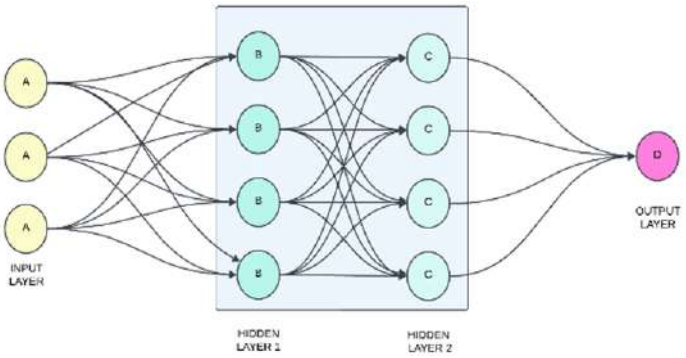

Figure 12. Structure of Artificial Neural Network

6 Model Evaluation Criteria

In this study, the following techniques were utilized to assess the performance of each model. The evaluation metrics include:

6.1 Accuracy

Accuracy is a metric that simply shows the number of times the model has correctly predicted whether an instance belongs to the target class or not:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FN + FP} \quad (4)$$

6.2 Precision

Out of all the occurrences projected as positive, precision is the percentage of accurately predicted positive instances (true positives):

$$\text{Precision} = \frac{TP}{TP + FN} \tag{5}$$

6.3 Recall (Sensitivity)

Out of the total number of actual positive occurrences, recall is the percentage of accurately anticipated positive instances (true positives):

$$\text{Recall} = \frac{TP}{TP + FP} \tag{6}$$

6.4 F1 Score

The F1 score, which balances the two evaluation metrics, is the balanced average of precision and recall:

$$\text{F1 Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

7 Results Discussion

While ML models like Random Forest perform well in terms of accuracy, DL models such as LSTM and Bi-LSTM offer more balanced performance across precision, recall, and F1-score. ANN shows the highest accuracy, but LSTM and Bi-LSTM deliver more consistent results across all metrics, making them more suitable for tasks requiring a balance between identifying both classes effectively.

7.1 Model Performance with Machine Learning Algorithms

Table 6 shows that KNN achieves a respectable accuracy of 77.97%, with a balanced F1-score of 0.589947, indicating decent performance but slightly lower precision and recall. Random Forest leads in accuracy among the ML models with 79.02%. However, its recall is relatively low at 0.473262, suggesting that it may struggle to identify the minority class. Logistic Regression shows solid overall performance with 78.75% accuracy and a balanced precision (0.621359) and recall (0.513369), resulting in a reasonable F1-score of 0.562225.

7.2 Model Performance with Deep Learning Algorithms

ANN shows the highest accuracy at 79.14%, but its precision and recall (0.71 and 0.67) are slightly lower than the other DL models, giving it the lowest F1-score (0.68) among deep learning methods. LSTM and Bi-LSTM perform similarly, with 79.03% and 78.68% accuracy, respectively. Bi-LSTM performed with a slightly better balance between precision, recall, and F1-score compared to the other implemented models. CNN has also achieved 78.68% accuracy, but its precision and recall remain on par with the other models, leading to a competitive F1-score of 0.71.

Table 6. Comparative Results of Machine Learning Models Across Performance Metrics

Model	Accuracy (%)	Precision	Recall	F1-Score
KNN	77.97	0.583770	0.596257	0.589947
Random Forest	79.02	0.627660	0.473262	0.539634
Logistic Regression	78.75	0.621359	0.513369	0.562225

Table 7. Comparative Results of Deep Learning Models Across Performance Metrics

Metric	LSTM	Bi-LSTM	CNN	ANN
Accuracy (%)	79.03	78.68	78.68	79.14
Precision	0.73	0.73	0.73	0.71
Recall	0.70	0.71	0.70	0.67
F1-Score	0.71	0.72	0.71	0.68

Table 8. Model Parameters and Configuration Details for Implementation

Model Parameter	Validation Split	Dropout Rate	Loss	Activation Function	Optimizer	Epochs Run	Batch Size
LSTM	0.2	0.1	binary_crossentropy	Sigmoid	Adam	10	32
BI-LSTM	0.2	0.3	binary_crossentropy	Tanh	RMSprop	20	32
CNN	0.2	0.3	binary_crossentropy	ReLU	Nadam	20	32
ANN	0.2	0.3	binary_crossentropy	Sigmoid	Adam	20	32

7.3 Architecture Utilized in Implementation

This study compares four neural networks: LSTM, BI-LSTM, CNN, and ANN, as shown in Table 8. Each model uses a 20% validation split and the binary cross-entropy loss function. LSTM has a 10% dropout rate, while others use 30% to prevent overfitting. Activation functions vary: CNN uses ReLU, BI-LSTM uses tanh, and both LSTM and ANN use sigmoid. Different optimizers are employed (CNN: Nadam, BI-LSTM: RMSprop, LSTM & ANN: Adam). LSTM trains for 10 epochs, while CNN, ANN, and BI-LSTM train for 20 epochs with a batch size of 32. This setup allows for a thorough performance comparison and model optimization.

8 Conclusion

This study demonstrates the significance of machine learning and deep learning techniques for predicting customer attrition and creating proactive retention strategies. In comparing several models, it was found that standard machine learning algorithms, such as Random Forest and Logistic Regression, performed well in terms of accuracy, while deep learning models, such as LSTM and Bi-LSTM, offered a more balanced approach and produced strong results across important metrics like precision, recall, and F1-score. By using these models to anticipate churn, better understand customer behavior, and launch timely retention campaigns, organizations can reduce attrition rates and enhance long-term client loyalty. By continuously refining these models and applying advanced preprocessing techniques, businesses in various industries, including telecommunications, finance, and retail, can harness the potential of predictive analytics to not only mitigate churn but also drive sustained growth and profitability.

Author Contributions

Hira Farman: Idea, Methodology, Writing- Original draft preparation **Samar Raza Talpur**: Supervision **Govari shankar**: Analysis **Umm e Laila**: Simulation work **Lubaba Naseem**: Software, Validation.

Compliance with Ethical Standards

It is declare that all authors don't have any conflict of interest. It is also declare that this article does not contain any studies with human participants or animals performed by any of the authors. Furthermore, informed consent was obtained from all individual participants included in the study.

References

[1] B. Senthilnayaki, M. Swetha, and D. Nivedha, "Customer churn prediction," *IARJSET*, vol. 8, pp. 527–531, 2021.

[2] B. Prabadevi, R. Shalini, and B. R. Kavitha, "Customer churning analysis using machine learning algorithms," *International Journal of Intelligent Networks*, vol. 4, pp. 145–154, 2023.

[3] P. Lalwani, M. K. Mishra, J. S. Chadha, and P. Sethi, "Customer churn prediction system: A machine learning approach," *Computing*, vol. 104, no. 2, pp. 271–294, 2022.

[4] A. Khattak, Z. Mehak, H. Ahmad, M. U. Asghar, M. Z. Asghar, and A. Khan, "Customer churn prediction using composite deep learning technique," *Scientific Reports*, vol. 13, no. 1, p. 17294, 2023.

[5] J. B. Brito, G. B. Bucco, R. Heldt, J. L. Becker, C. S. Silveira, F. B. Luce, and M. J. Anzanello, "A framework to improve churn prediction performance in retail banking," *Financial Innovation*, vol. 10, no. 1, p. 17, 2024.

[6] S. Saha, C. Saha, M. M. Haque, M. G. R. Alam, and A. Talukder, "Churnnet: Deep learning enhanced customer churn prediction in telecommunication industry," *IEEE Access*, 2024.

[7] H. Farman, A. W. Khan, S. Ahmed, D. Khan, M. Imran, and P. Bajaj, "An analysis of supervised machine learning techniques for churn forecasting and component identification in the telecom sector," *Journal of Computing & Biomedical Informatics*, vol. 7, no. 1, pp. 264–280, 2024.

[8] A. Bugajev, R. Kriaužienė, O. Vasilecas, and V. Chadysas, "The impact of churn labelling rules on churn prediction in telecommunications," *Informatica*, vol. 0, no. 0, pp. 1–31, 2022.

[9] H. Farman, N. Islam, S. A. Ali, D. Khan, H. A. Khan, M. Ahmed, and A. Farman, "Advancing rainfall prediction in pakistan: A fusion of machine learning and time series forecasting models," *International Journal of Emerging Engineering and Technology*, vol. 3, no. 1, pp. 17–24, 2024.

[10] S. K. Wagh, A. A. Andhale, K. S. Wagh, J. R. Pansare, S. P. Ambadekar, and S. H. Gawande, "Customer churn prediction in telecom sector using machine learning techniques," *Results in Control and Optimization*, vol. 14, p. 100342, 2024.

[11] N. Bagul, P. Berad, P. Surana, and C. Khachane, "Retail customer churn analysis using rfm model and k-means clustering," *Int. J. Eng. Res. Technol.*, vol. 10, no. 03, 2021.

[12] P. Verma, "Churn prediction for savings bank customers: A machine learning approach," *Journal of Statistics Applications & Probability*, vol. 9, no. 3, pp. 535–547, 2020.

[13] M. Rahman and V. Kumar, "Machine learning based customer churn prediction in banking," in *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, pp. 1196–1201, IEEE, 2020.

[14] S. A. Qureshi, A. S. Rehman, A. M. Qamar, A. Kamal, and A. Rehman, "Telecommunication subscribers' churn prediction model using machine learning," in *Eighth International Conference on Digital Information Management (ICDIM 2013)*, pp. 131–136, IEEE, 2013.

[15] N. Gurung, M. R. Hasan, M. S. Gazi, and F. R. Chowdhury, "Ai-based customer churn prediction model for business markets in the usa: Exploring the use of ai and machine learning technologies in preventing customer churn," *Journal of Computer Science and Technology Studies*, vol. 6, no. 2, pp. 19–29, 2024.

[16] A. Sikri, R. Jameel, S. M. Idrees, and H. Kaur, "Enhancing customer retention in telecom industry with machine learning driven churn prediction," *Scientific Reports*, vol. 14, no. 1, p. 13097, 2024.

[17] P. Baruah and B. Sarma, "Customer churn prediction using ensemble techniques and algorithms," *Educational Administration Theory and Practice Journal*, vol. 92, 2024.

[18] Y. Jiang, "Customer churn analysis prediction based on cluster analysis and machine learning algorithms," in *Proceedings of the 3rd International Conference on Business and Policy Studies*, 2024.

[19] E. D. Ramírez Martínez, L. R. García Oyervides, and J. A. García Quijano, "Churn detection on bank customers," tech. rep., Tecnológico de Monterrey, 2024.

[20] D. Das and S. Mahendher, "Comparative analysis of machine learning approaches in predicting telecom customer churn," *Educational Administration Theory and Practice Journal*, 2024.

[21] P. Baruah, P. Deshmukh, P. Karanjawane, D. Chaudhari, and N. M. Ranjan, "Churn prediction in telecommunication industry," in *Proceedings of the 2023 International Conference for Advancement in Technology (ICONAT)*, (Goa, India), 2023.

[22] N. N. Koranchirath, "Predictive modelling and customer retention: A machine learning approach to analyze churn," *International Journal of Computer Techniques*, vol. 11, no. 2, 2024.

[23] H. Farman, D. Khan, S. Hassan, M. Hussain, and S. A. A. Usmani, "Analyzing machine learning models for forecasting precipitation in australia," *Journal of Computing & Biomedical Informatics*, vol. 7, no. 1, pp. 439–458, 2024.

[24] P. Khare and S. Arora, "Predicting customer churn in saas products using machine learning," *International Research Journal of Engineering and Technology (IRJET)*, vol. 11, no. 5, pp. 754–765, 2024.

[25] K. C. Moulia, C. V. Raghavendran, V. Y. Bharadwaj, G. Y. Vybhavi, C. Sravani, K. M. Vafaeva, R. Deorari, and L. Hussein, "An analysis on classification models for customer churn prediction," *Cogent Engineering*, vol. 11, no. 1, p. 2378877, 2024.

[26] K. H. Akhila, N. Swathy, and M. Manjula, "Churn predictions in marketing," tech. rep., SVTB College, Mannampatta, 2024.

[27] G. Sam, P. Asuquo, and B. Stephen, "Customer churn prediction using machine learning models," *Journal of Engineering Research and Reports*, vol. 26, no. 2, pp. 181–193, 2024.

[28] H. H. Ahmed, M. H. Khafagy, and M. R. Kaseb, "A novel model for partial and total churn prediction in e-commerce," tech. rep., Fayoum University, 2024.

[29] A. Manzoor, M. A. Qureshi, E. Kidney, and L. Longo, "A review on machine learning methods for customer churn prediction and recommendations for business practitioners," *IEEE Access*, 2024.

[30] X. Jiang, "Customer churn data analysis using data mining," in *Proceedings of the 2nd International Conference on Software Engineering and Machine Learning*, 2024.

[31] E. Alihosseinzadeh, *Siamese Networks for Telecommunication Customer Churn Data in a Few-Shot Learning Context*. PhD thesis, Norwegian University of Life Sciences (NMBU), 2024.

[32] Y. Wei, "Telco customer churn prediction," in *Highlights in Science, Engineering and Technology SDPIT 2024*, vol. 92, Johns Hopkins University, 2024.

[33] N. Karunanithi and N. Nithi, "A neural network approach for software reliability growth modeling in the presence of code churn," in *Proceedings of the International Symposium on Software Reliability Engineering (ISSRE)*, June 2024.

[34] S. Wu, "Customer churn prediction in telecom based on machine learning," in *Highlights in Science, Engineering and Technology CMLAI 2024*, vol. 94, (Shanghai University of Finance and Economics), pp. 113–118, 2024.