[Information Sciences Letters](https://digitalcommons.aaru.edu.jo/isl)

Volume 11
Issue 1 *Jan. 2022*

Article 24

2022

Customer Churn Prediction in Telecommunication Industry Using Deep Learning

Samah Wael Fujo
College of Information Technology, Ahlia University, Bahrain, moaiad.khder@asu.edu.bh

Suresh Subramanian
College of Information Technology, Ahlia University, Bahrain, moaiad.khder@asu.edu.bh

Moaiad Ahmad Khder
Department of Computer Science, Applied Science University, Bahrain, moaiad.khder@asu.edu.bh

Information Sciences Letters
An International Journal

http://dx.doi.org/10.18576/isl/110120

Customer Churn Prediction in Telecommunication Industry Using Deep Learning

Samah Wael Fujo[1], *Suresh Subramanian*[1] *and Moaiad Ahmad Khder* [2,*]

[1]College of Information Technology, Ahlia University, Bahrain
[2]Department of Computer Science, Applied Science University, Bahrain

Received: 7 May 2021, Revised: 21 Jun. 2021, Accepted: 21 Oct. 2021
Published online:1 Jan. 2022

Abstract: Without proper analysis and forecasting, industries will find themselves repeatedly churning customers, which the telecom industry in particular cannot afford. A predictable model for customers will allow companies to retain current customers and to obtain new ones. Deep-BP-ANN implemented in this study using two feature selection methods, Variance Thresholding and Lasso Regression, in addition, our model strengthened by early stopping technique to stop training at right time and prevent overfitting. We compared the efficiency of minimizing overfitting between dropout and activity regularization strategies for two real datasets: IBM Telco and Cell2cell. Different evaluation approaches used: Holdout, and 10-fold cross-validation to evaluate the model's efficiency. To solve unbalanced issue, the Random Oversampling technique was used to balance both datasets. The results show that the model implemented performs well with lasso regression for feature selection, early stopping technique to pick the epochs, and large numbers of neurons (250) into the input and hidden layers, and activity regularization to minimize overfitting for both datasets. In predicting customer churn, our findings outperform ML techniques: XG_Boost, Logistic_Regression, Naïve_Bayes, and KNN. Moreover, our Deep-BP-ANN model's accuracy outperforms the existing deep learning techniques that use holdout or 10-fold CV for the same datasets.

Keywords: Deep learning, customer churn, ANN, backpropagation, lasso regression.

1 Introduction

Telecommunications has long been part of our culture but is attracting more attention now than ever before. Several scientific journals and corporate publications report exciting technological developments in technology and the competition in the mobile telecoms sector. A compelling case is GPRS-based medium broadband networks' rapid growth, which now hit a market penetration level of 70–80 percent in many EU countries. Furthermore, long-distance changes due to the wireless telecoms industry's unbundling and merging with other companies such as IT and media. [1][2]. Telecommunications providers face rising customer support pressure than ever before delivering the triple play package of Audio, Video, and Internet access services. Underlying the difficulty of triple play service is the need to provide consumers with a high-quality experience, either when they use the service or when they ask for help from their service provider [3]. Due to its tremendous growth in recent years, this paper focuses primarily on the telecoms industry. Where almost everybody today has a telecom package, with simple connectivity and a range of service providers [4].

Churn describes the consumer who transfers from one supplier of telephone services to another. [5]. The telecom industry faces a major problem linked to customer turnover, the user who will in the coming years leave their existing relationship with the business/network. This issue can not only impact the company's rapid growth, but it can also affect revenues. Therefore, many Customer Churn Prediction (CCP) models have been implemented but do not deliver the desired CCP performance; this is due to potential variant variables influencing Customer Churn (CC), which would still be unexplored. [6]. Customer may churn for many reasons: the company does not interact and communicate with their clients, complaints were not dealt with promptly, negative comments on social media, press, etc., unsatisfied with the services offered, the present software does not suit needs, the competitor gave the new product a better quality and price, or not supporting advanced network types (4G, 5G, 6G). In most cases, the company does not aware of the reasons behind the

*Corresponding author e-mail: moaiad.khder@asu.edu.bh

customer's leaving. In this case, analysing the data can help understand and predict the customer churn to avoid its occurrence [7].

Model Churn Prediction support evaluating historical business data to identify the clients' list at high risk of churning. It would allow the telecommunications industry to concentrate on a particular category rather than on-individual using retention techniques. Personalized retentions are difficult because companies generally have a large customer base and cannot afford to invest a lot of time and resources. Nevertheless, the possibility of predicting which users are at risk of losing, it may minimize customer retention pains by targeting them specifically to those customers [4]. Today, technology plays a significant role in addressing consumer needs and achieving a degree of satisfaction. It is the business's fundamental need to develop an efficient and productive model for managing consumer churn. There have been many modeling methods that are used to predict CC in various organizations. [8] reviewed the total number of papers that used these methods, as shown in Figure 1.

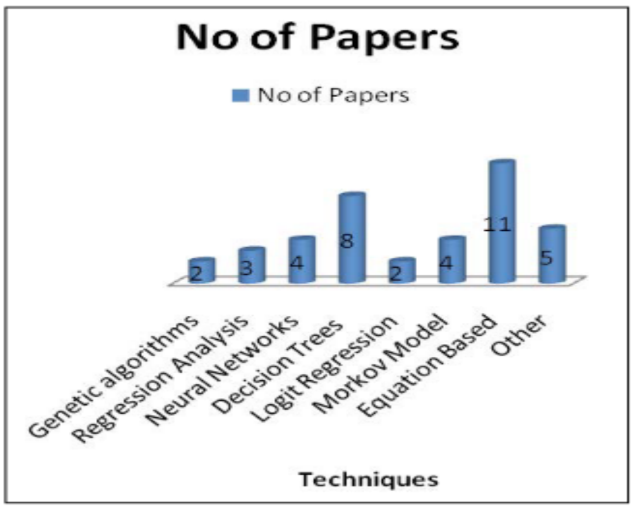

Fig.1: Techniques used to prediction churn.

However, the number of data became huge compared with before. The Machine learning (ML) model works well but not for a massive amount of data. Therefore, in recent years the studies, especially for customer churn prediction, tend to use deep learning (DL) techniques more since they can deal with massive data compared with ML techniques. Deeply structured Learning has stood out as a new field of study in ML since 2006, or more formally named "deep learning or hierarchical learning". DL belongs to a class of "artificial neural networks (ANNs)" that consist of several processing layers [9]. The techniques built through Deep Learning analysis have already affected a wide variety of signals and data processing work across the traditional and new, expanded spectrum over the past couple of years, involving ML and AI [10]. In deep Learning, the "deep" derives from neural networks, typically multiple layers built into deep learning models. The Convolutional Neural Network (CNN) consists of numerous layers of models in which each layer taking inputs from the previous layer, processes it, and outputs it to the next layer in a daisy-chain fashion [11].

2 Related Work

[9] presented a technique for using DL models to remove the manual feature engineering. Three DNN architectures were developed, and the respective CCP model was designed utilizing two CrowdAnalytix and Cell2Cell telecom datasets and compared how they perform with other models. Authors show that models of DL work just as well as conventional classifiers such as SVM and random forest. [12] used the principle of"Transfer Learning (TL)" and"Ensemble-based Meta-Classification," they present a solution to the inherent issues of churn prediction. The proposed "TL-DeepE" approach is implemented in two steps. By fine-tuning many pre-trained Deep Convolution Neural Networks (DCNN), the first stage used TL (CNNs). Predictions from these DCNNs are appended to the original feature vector in the second stage. Therefore, their experiments are carried out using different CNNs as base classifiers and as a meta-classifier, and GP-AdaBoost. The proposed "TL-DeepE" method's efficiency is compared with existing techniques for two common telecommunications datasets, Orange and Cell2cell, by the use of 10-fold cross-validation. The prediction accuracy obtained was 75.4 percent and 68.2 percent by conducting experiments on Orange and Cell2cell datasets, while the AUC was 0.83 and 0.74, respectively. [13] demonstrated that deep Learning enables multi-stage models to represent the data at multiple abstraction levels. It also reduces the time and effort of feature selection considerably as it automatically creates useful features. Authors have compared various models based on accurate customer churn prediction. The study was carried out on the Telco Churn customer dataset of IBM. Artificial neural network (ANN) model achieved a validation data accuracy of 82.83%, which is better than K-Nearest Neighbours (KNN) conventional approach. The results indicate that the multi-layered "ANN" model with "self-learning" capacities and tokenism input data performs better than traditional classification algorithms. [14] suggested a model for the CCP that can be considered to assess the parameters' efficacy, i.e., the lower and the upper distance between the samples using publicly available datasets. Researchers randomly picked the Naïve Bayes classification algorithm as their base-classification. The result of the analysis in this paper shows that "Lower Distance Test Set (LDT)" samples had the best performance comparing it with "Upper Distance Test sets (UDT)." The reason is that the classification performance on the upper-distance samples stays almost the same whenever the test data is increased. Authors have investigated that the CCP model performance in an unclear state in a dataset is inversing proportion to accuracy. The authors mention that a future study could

deliver empirical results with multiple base-classifier on the balanced dataset. It would be essential to realize the CCP model's impact if the selection method applied to the features by assigning them weights. Moreover, the studies discussed that more research is detailed with other model types in the future, offering an option to compare this study finding and ultimately analyze the effect statistically. [15] used a Deep Learning Method to illustrate churn prediction on a Telco dataset. To construct a nonlinear classifier, a multi-layered NN was developed. The CP model operates on customer attributes, attributes of assistance, attributes of use, and relative attributes. The capacity for turnover as well as the deciding variables are predicted. The qualified model then assigns the final weights to these attributes and calculates the probability of churn for that client. It achieved an accuracy of 80.03 percent. Since the model can include the churn variables, businesses can use it to evaluate and take measures to remove the reasons for such factors.

[16] applied their study to identify the factors affecting customer churn and build an effective model of churn prediction and provide the best interpretation of the results of data visualization. Kaggle's open data website has compiled the dataset. The proposed method for churn prediction analysis covers several phases: pre-processing of data, analysis, implementation of ML algorithms, classifier evaluation, and selecting the best one for CP. Three main actions, cleaning, transformation, and selection of features, involved the data pre-processing process. LR, ANN, and RF were selected as ML classifiers. To select the optimal classifier, classifiers were tested using performance evaluation metrics: precision, accuracy, recall, and error rate. The performance shows that LR outperforms compared to the ANN and RF, based on their analysis. [17] proposed a forecast model to classify the characteristics that significantly affect consumer churn using ML techniques such as "K-nearest-neighbour (KNN)", XG Boost, and RF. To estimate the churn, "IBM Watson" data was analysed. Eventually, ML performed a comparative analysis to determine the best algorithm with the highest accuracy.

Experimental tests reveal that the XGBoost classification provides 0.798, which is better accuracies than the KNN and RF classifications. The "XGBoost" classifier is then used for the collection of features in a proposed framework. The model presented indicates that Fiber Optic users with a higher monthly charge are more influenced for churn

3 Method

The proposed system is implemented based on a supervised classification deep learning technique to predict if the customer is going to churn or not in the telecom industry [18], [19], and [20]. Beside implemented the most common four ML technique LR, NB, KNN, XGBoost in CCP [21], [14], [20], [22], [13], [17], [16], to compare our proposed model with them. The proposed system process is shown in Figure 2.

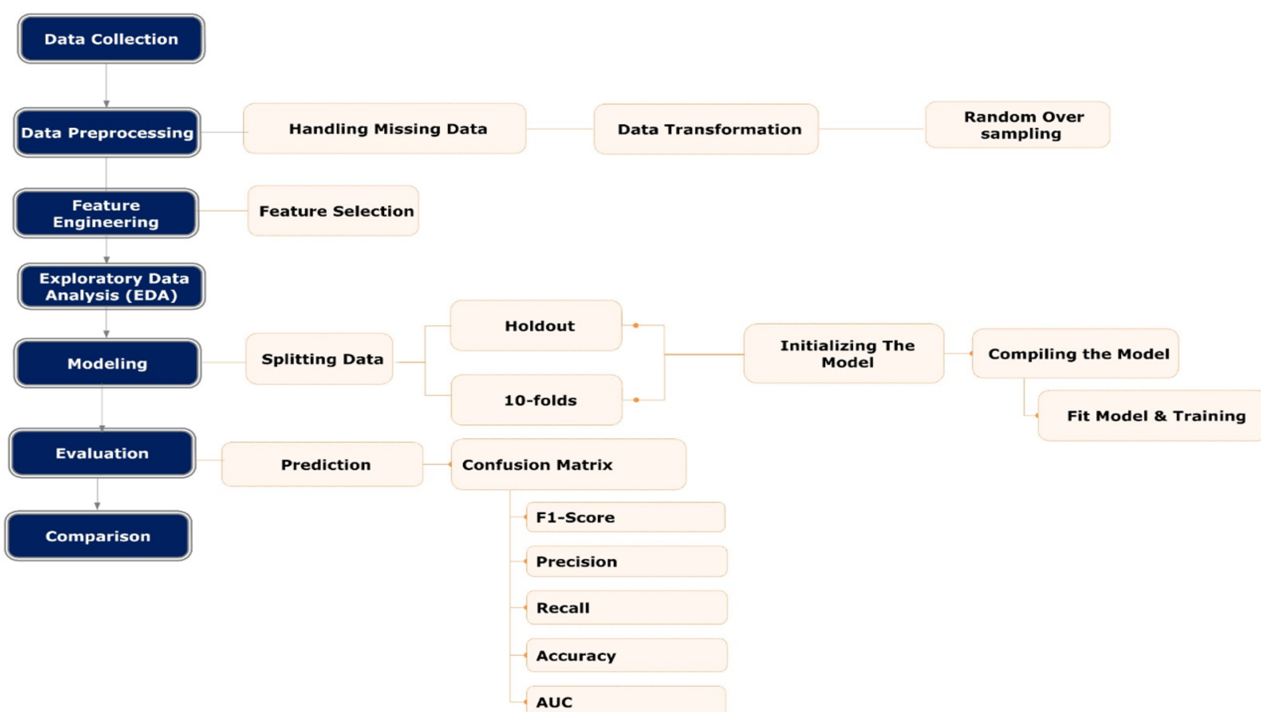

Fig.2: Proposed system processes.

The method includes six main phases: data collection, data pre-processing, feature engineering, random oversampling, modeling, and evaluation. The two most suitable datasets to predict customer churn has been selected. Pre-processing has been implemented on both datasets that include the Exploratory Data Analysis EDA to understand the given datasets and data transformation to pre-process the data. Two different feature selection techniques have been selected to experiment with the performance improvement: Lasso Regularisation and Removing Highly Correlated features (Variance Thresholding) methods. After that, to handle the imbalanced problem in both data sets, random-oversampling technique was used. Then Deep-BP-ANN model was implemented by setting different hyper parameters includes Early Stopping to monitor the generalization error, Model Checkpoint call-back to save the best performance obtained during training. Also, Activity Regularization has been used to avoid overfitting that may cause by random oversampling. The model evaluates the data by using two methods: holdout and K-fold cross validation. Finally, the model performance evaluation was measured using confusion metrics to calculate: F1-Score, Precision, Recall, Accuracy, and AUC.

3.1 Dataset

The data used in this study is the Cell2Cell dataset consists of 71,047 instances and 58 attributes. The Cell2Cell dataset have been used by many researchers [9],[23],[12],[24],[12], and [25]. The data is available at the "Centre for Customer Relationship Management Duke University's" website [26], and it is available in Kaggle [12]. The other data set used in the research is IBM Telco. The dataset consists of different attributes relating to a telecommunications company's customer, which enabled us to deduce a coherent relationship between the customer's actions and the churn. There are 21 attributes and 7044 rows in the dataset, and it is available in Kaggle. The IBM Telco data set has been used by many researchers to predict customer churn [13], [14], [15], [16], [17]. Both datasets have the dependent attribute, which is (churn) that shows whether or not the customer has churned, indicated by (Yes) or (No), respectively. Furthermore, both datasets have divided the attributes into two categorical features within an object type and continuous features within the numeric type. Table 1 shows the characteristics Cell2Cell and IBM Telco datasets.

Table 1: Characteristics of datasets.

	Cell2Cell	**IBM Telco**
Total of features	58	21
Total of customers	51047	7043
Missing value	Yes	Yes
Churn	28.8%	26.5%
Not churn	71.2%	73.5%
Data distribution	Imbalanced	Imbalanced
Categorical features	23	17
Numerical features	35	4
Dependent feature	1	1
Independent features	57	20

3.2 Pre-processing

One of the main stages of information discovery activities is data pre-processing, which plays a significant role. It requires a variety of steps, such as the transformation of data and the reduction of data. If raw data is converted into low-quality data, learning algorithms' efficiency and accuracy will be compromised. Thus, the collected data can be correctly analyzed by performing proper data pre-processing steps and choosing suitable learning algorithms [27].

Telecom datasets have several issues to address, such as missing values, non-numeric features, inconsistent scales of features, etc. Therefore, before implementing a learning model, it is important to pre-process the data [12]. There are specific and common characteristics of the datasets used, namely (Cell2cell and IBM Telco) as shown in Table 1.

In this research, both data sets (cell2cell and IBM Telco) are pre-processed using almost the same strategies. The pre-processing phase includes handling the missing data, encoding the categorical variables using two methods: "label encoding" and "one-hot encoding". Moreover, scaling the high variance values using the normalization technique and dropping the unnecessary features that do not affect the dependent variable. After that, some of the EDA has applied to get a better understand of the given data. The EDA shows that both datasets are imbalanced, and to deal with this kind of issue, we used the random oversampling technique.

1) Missing Data

Missing data influences statistical analysis through knowledge loss and data pattern irregularities [28].

Since both data sets (Cell2cell and IBM Telco) contain missing values, it is important to know the percentage of missing values and where they are located. Starting with the IBM Telco data set, the total charges contain missing values of about 0.16%, which is less than 30%. The Cell2Cell dataset contains fourteen features with numeric types have missing values with different proportions but still less than 30%.

[29], [30] mentioned that when the missing values are over the threshold of 30%, it is better to remove them; otherwise, take the residual mean for the numeral value and most frequent for categorical values. Therefore, missing data in both data sets are handled by using the simple imputer technique by using Python's Sklearn library. Simple Imputer is used to impute or replace incomplete continuous and categorical variables with various means, median, most frequent, and constant value [31].

[29], [30], filled the missing values of more than 30% with the residual mean, and [9] use the forward filling technique to fill the missing values with more than 95% by the mean.

[28] omitted characteristics with a missing value percentage greater than 0.90.

In our datasets, the missing values occur on the numeric variables, and they are less than 30%, so the best strategy to deal with them is to fill the null with each particular feature's mean.

2. Data transformation

Data transformation techniques can greatly improve the churn prediction's overall performance [16]. Three different data transformation methods have been applied on IBM Telco and Cell2cell data sets: Label encoding, One hot encoding, and Normalization.

Label Encoding (LE): is a basic technique used to map categories as continuous integers in nominal attributes. It is accepted by several data transformation tools, such as Labe-encoder, because of its simplicity "program package from scikit-learn in Python" [27]. LE technique is more useful with categorical variables that contain only two unique values, such as (yes or no) and (male or female).

IBM Telco data set contains six categorical variables with two unique values, which are gender, partner, dependents, phone service, paperless billing, and churn. These features encoded using LE technique where yes replaced by one and no replaced by zero and gender male replaced by one and female replaced by zero. Cell2cell data set contains fifteen features with yes/no these features encoded using LE technique as did in IBM Telco. The homeownership feature is also encoded using the LE technique (known) replaced by one and (unknown) replaced by zero.

One-Hot Encoding (OHE) (or Dummy): is a commonly used approach based on a binary encoding that uses a 1 to denote a group's presence and a 0 to denote its absence. It transforms each category into an N-dimensional vector with an element of (0 or 1) in the nominal attribute containing N categories. This technique does not harm the raw information. However, some issues with the OHE transforms, such as "sparse data" and "high dimensionality," will increase learning algorithms' computational time. Therefore, OneHot should not be used for high cardinality columns [27].

High cardinality means many unique values (a large k). An example of a high cardinality function is a column with customer identification and service area characteristics. Therefore, high cardinality triggers high dimensionality, indicating a multi-dimensional matrix [32]. Hence, these two columns would be dropped before the OHE method is implemented.

Normalization: the two datasets were normalized to improve the results on customer prediction with ML methods. One of the most common normalization methods is MinMaxScaler. "MinMaxScaler" scales and translates each feature individually by rescaling variables into the range (0,1) [33] [29]. The MinMax Scaler is applied only on the high variance features, not the entire data set; otherwise, they cause over rescaling when categorical features are already encoded.

There are many other techniques to rescale the values, such as standardization and binarization, but normalization shows the best performance results after some experiments. Furthermore, several researchers applied this technique in their data to rescale it [29], [12], [34].

3. Resampling the datasets

Imbalanced classification is one of the predictive modelling classification problems which is used in this research. An imbalance happens when, relative to the other classes, one or more classes have meager proportions in the training results. In our case, the churn class has very low proportions compared with the not churn class. For example, customers' features have 80 instances of customers not churned and 20 instances of churn customers, and our training dataset contains only these instances. This reflects an instance of an imbalanced problem of classification. Class imbalance adds critical challenges to the CCP [35], [36], [37].

[38] mentioned that "Any dataset with an unequal class distribution is technically imbalanced. However, a dataset is imbalanced when there is a significant, or in some cases extreme, disproportion among the number of examples of each class of the problem".

Therefore, it has been found that Cell2cell and IBM Telco datasets were unbalanced since the percentage of the second (minority) class representing churn customers is about 28.8% in Cell2cell and 26.5% in IBM Telco's whole dataset.

In order to tackle this problem, several solutions have been provided. By resampling the existing, unbalanced data collection, data-level solutions aim to re-balance the class distribution. Data-level solutions, consisting of several different resampling techniques, are among the most common solutions for the imbalanced issue. "Random oversampling (ROS) and Random under-sampling (RUS)" are given by random sampling techniques [36], [35], [39].

By randomly reproducing the minority instances from the current data set, ROS attempts to rebalance the class distribution in the original data set. On the other hand, Via the random deletion of majority class instances, RUS seeks to balance the class distribution. ROS and RUS are flexible to use and comprehend their strategies. However, RUS will discard potentially valuable data instances, while by making exact copies of current instances, ROS can increase overfitting [35], [45].

In this study, the ROS has been chosen because its drawback can be overcome using the Activity Regularization technique in our Deep-BP-ANN model. ROS improves performance in terms of accuracy, recall, precision, and F1-score in both datasets compared with other studies.

4. Feature selection

A selection of features typically used in the pre-stage of ML, which refers to the selection process of a subset of specific features of a set of features [16]. In this study, the correlation Matrix has been used to visualize how each variable correlates with the other variables. The feature that is above 0.5 or equal were be excluded from the data set using variance Thresholding method. Another feature selection method used in this study is lasso regression which is one of the most prevalent embedded feature selection approaches [40].

"Lasso Regression" has the potential to nullify the influence of an irrelevant variable in the data, which means that it will fully remove the coefficient of a variable to zero and thus is better at decreasing the variance whenever the data consists of several insignificant features. Mathematically, Lasso is = Square Residual Sum + λ * ("Sum of the absolute value of the magnitude of coefficients") [41] [42].

$$\text{Lasso:} \quad \sum_{i=1}^{n}(y_i - \sum_j x_{ij}\,\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \qquad (1)$$

Where λ a sum of shrinkage. λ = 0 It means that all attributes are taken into account and is equal to a linear regression where only squares residual sum is assumed to construct a predictive model. λ = ∞ no attribute is taken into account, i.e., (as λ closes to infinity). The bias increases with an increase in λ, and variance increases with a decrease in λ.

5. Exploratory data analysis (EDA)

One of the most important challenges in developing the mobile telecoms service industry is retaining clients. The EDA enables a service provider to track the product service that influences the client to churn and suggest or facilitate the best service or strategy to attract and retain the customer [43], [16].

The technique used to find out the top ten important features to predict customer churn in both data sets is the XGBoost. An advantage of using XG Boost is that it is relatively easy to retrieve important scores for each attribute after the boosted trees are created. In general, significance offers a score that shows how useful or important each feature was within the model in constructing the boosted decision trees. The more an attribute is used with decision trees to make key decisions, the greater its relative value. Attributes in the dataset, this value is determined directly, allowing attributes to be ranked and compared to each other. The feature's significance is then averaged over all decision trees within the model [44], [45].

4 Deep-Bp-Ann Model

After several experiments, the optimal number of neurons is 250 neurons for input and hidden layers. Besides, the optimal number of hidden layers were two hidden layers. These optimal options work well for both data sets: IBM Telco and Cell2cell.

When dealing with binary classification problems as customer prediction, the best activation function to use in the input layer and hidden layers is Rectified Linear (ReLU) activation function and the sigmoid activation function in the output layer [29], [47], [48].

The selecting number of training epochs one of the biggest challenges with training neural networks. Many epochs will contribute to the training dataset overfitting, whereas too few will result in an under fit model. Early stopping is a technique that allows to assign a large number of epochs randomly and stop training once the model performance on a validation dataset stops from improving [49].

The early stopping call-back technique is used in recent research for different problems when building the deep learning model [50],[51],[52],[12]. Nevertheless, no research applied this technique when predicting customer churn in the telecom industry using deep learning.

To compile the model and optimize the weights, Adam, the backpropagation DL optimization algorithm, was selected, which can measure adaptive learning rates for each parameter as the "gradient descent optimization algorithms" of the ANN model. Simultaneously, binary "cross-entropy" was used as an objective function to minimize the loss function during model training.

The activity regularization method has been used to reduce Deep-BP-ANN overfitting on predicting a binary classification problem (customer churn) [45].

5 Model Validation

The most significant aspect of developing a supervised model is model validation. One must have sensible data separating strategy to construct a model with excellent generalization efficiency, which is essential for model validation [53]. Two common methods for validation are used Hold out (90% for training and 10% testing) and 10-Fold CV.

6 Evaluation

Most studies using AUC and Accuracy to evaluate their model in CCP. Thus, this study used the confusion matrix to evaluate the implemented model and AUC metric. The accuracy, Recall, F1-Score, and precision can be obtained from the confusion matrix as shown in Table 2.

Table 1: Confusion Matrix

		ACTUAL	
		Churn	not-churn
PREDICTED	Churn	TP	FN
	Not-Churn	FP	TN

These metrics are suitable for analysing any model built using balanced and unbalanced datasets [9].

The following explains how each term described in our model:

1. A customer who is churning (positive) and classified as churn (positive) is called “true positive (TP)”.
2. A customer who is not churning (negative) and classified as not churning (negative) is called true “negative (TN)”.
3. A customer who is not churning (negative) and classified as churn (positive) is called “false positive (FP)”.
4. A customer who is churning (positive) and classified as not churning (negative) called “false negative (FN)”.

Accuracy computes the percentage of customers who are correctly churned or not churned.

Accuracy: $$\frac{TP+TN}{TP+TN+FP+FN} \tag{2}$$

Precision the percentage of TP to the sum of TP and FP is knowing as precision.

Precision: $$\frac{TP}{TP+FP} \tag{3}$$

Recall focuses on the number of FN thrown into a prediction mixture. The recall is also known as sensitivity or true positive rate, and it is calculated as the following:

Recall: $$\frac{TP}{TP+FN} \tag{4}$$

The F1-score becomes one only if both “precision and recall” are 1. Furthermore, since both “precision and recall” are high, then the F1 score is raised. The harmonic mean of (precision and recall) is the F1-score (Krishnan, 2020).

F1-Score: $$\frac{Precision \times Recall \times 2}{Precision + Recall} \tag{5}$$

The AUC evaluation metric is also used to measure the efficiency and performance of a binary Classifier. The AUC provides a more powerful evaluation metric than other evaluation metrics, which measures a supervised classification’s overall performance by considering all potential cut-off points on the receiver’s operating features curve [47].

7 Result

A preliminary look at the overall churn rate on both data sets shows that in IBM Telco, 73.5% of the customers are not churning still belong to the company, and 26.5% are churned, as shown in Figure 3. Whereas in Cell2cell, the number of not churned customers is 71.2%, while 28.8% are churned, as shown in Figure 4. The results mean that this is an imbalanced classification problem. ML and DL algorithms work well only if the number of records of each class is approximately equal, as mentioned earlier. Therefore, both datasets are skewed; this issue can be handled in several ways; in this research, the ROS technique has been used to handle the imbalanced issue.

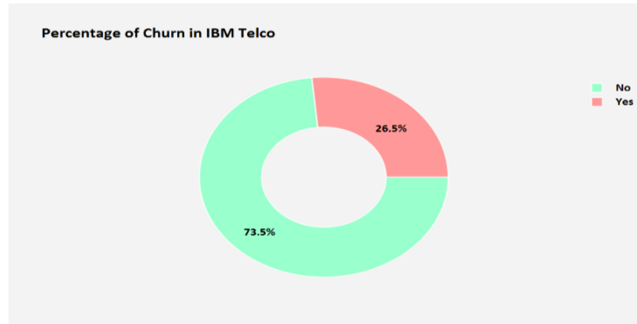

Fig. 3: Churn rate in IBM Telco

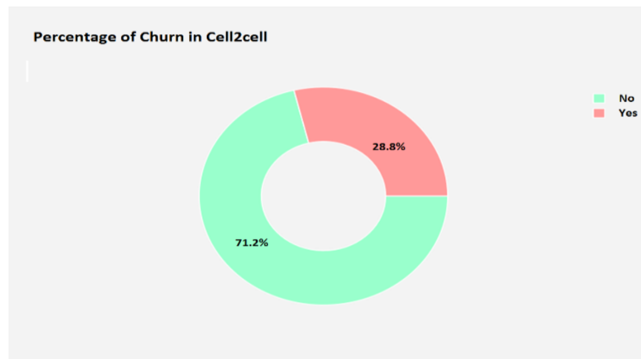

Fig. 4: Churn rate in cell2cell

The top ten influencing features to the churn using lasso regularization for feature selection and XG Boost for feature importance techniques in IBM Telco and Cell2cell are shown in Figure 5 and Figure 6. The level of importance is different based on the f score. The highest influencing features with the highest f score, the total charge is the highest importance following by tenure, in IBM Telco. Whereas, in Cell2cell, PercChangeMinutes, and PerChangeRevenues is the highest significant features to the churn.

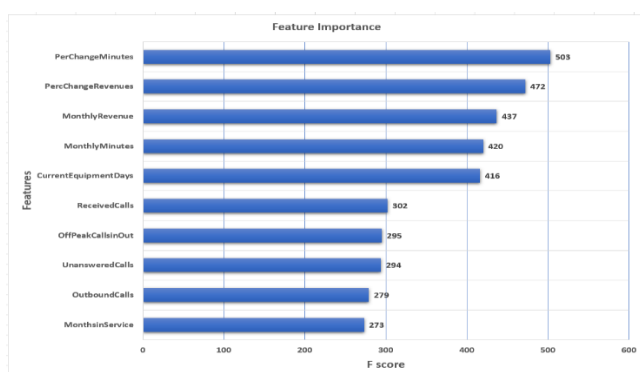

Fig.5: Top 10 influencing features in Cell2cell.

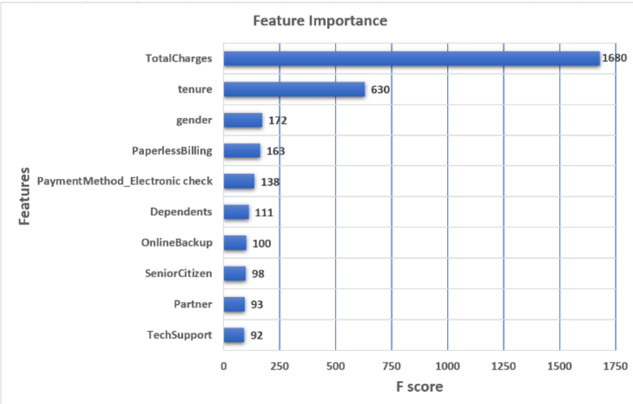

Fig. 6: Top 10 influencing features in IBM Telco.

The results show that the classification from the learned Deep-BP-ANN has been improved. Moreover, the Deep-BP-ANN is the best in classifying the churn and not churn customers, followed by XG Boost. The Deep-BP-ANN classified (430) customers with a rate of 41.55% as True-negative and (482) 46.57% as True-positive. While XG Boost classified (411) with a rate of 39.71% as True-negative, and (477) with a rate of 46.09% as True-positive.

The results show that the best classification results were obtained by Deep-BP-ANN, followed by XG Boost obtained from IBM Telco. The Deep-BP-ANN classified (2523) customers with 34.71% as True-negative and (3246) customers with a rate of 44.66% as True-positive. At the same time, XG Boost classified (2506) customers with a rate of 34.48% as True-negative and (2727) customers with a rate of 37.52% as True-positive.

The evaluation performance of our model for IBM Telco is shown in Figure 7. Our model outperforms XG Boost on its results in ACC, Recall, Precision, F1-Score, and AUC. The Deep-BP-ANN model achieved the best accuracy by a hold-out evaluation of 88.12% for IBM Telco and Precision 84.71%, Recall 93.04%, F1-Score 88.68%, and AUC 88.12%. The results of the evaluation performance in Cell2cell using holdout methods are shown in Figure 8.

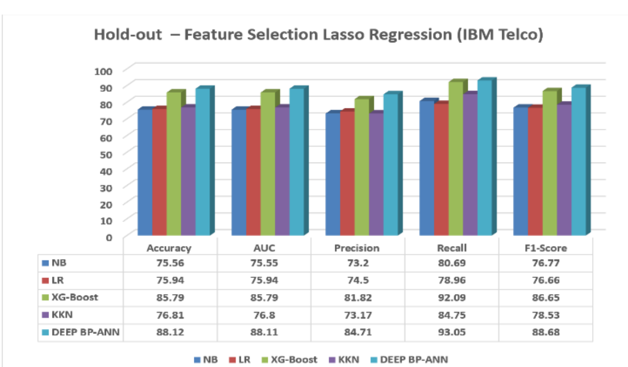

Fig. 7: The Evaluation Performance using Holdout in IBM Telco.

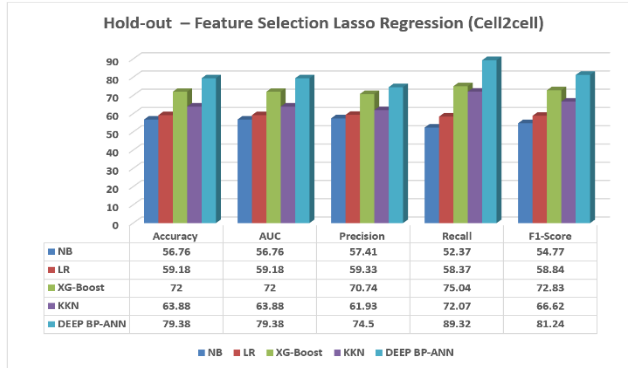

Fig.8: The Evaluation Performance using Holdout in Cell2cell Telco.

The result shows that Lasso Regression greatly affects the Cell2cell data set since some features are irrelevant to the response variable. The improvement has been noticed for Deep-BP-ANN and XG-Boost models. However, still Deep-BP-ANN model outperforms XG Boost and other classifiers. The Deep-BP-ANN model achieved the best accuracy with a hold-out evaluation, 79.38% for Cell2cell and Precision 74.5%, Recall 89.32%, F1-Score 81.24%, and AUC 79.38%.

The last experiment applied 10-fold cross-validation to evaluate the model obtained from the previous hold-out method by using Lasso Regression, 250 neurons, 32 batch size, 214 epochs (based on the nearest point obtained early stopping technique), and Activity Regularization.

The best classification results obtained by Deep-BP-ANN followed by XG Boost. The model able to classified (4082) customers with a rate of 39.45% as True-negative. Moreover, the model able to classified (4786) customers 46.25% as True-positive.

The 10-fold cross-validation is also applied to the Cell2cell data set to evaluate the last experiment's model results that used the hold-out method. The results show that Deep-BP-ANN can classify 29.03% as True-negative and can classify 44.87% as true positive.

The obtained classification results from Deep-BP-ANN is outperformed XG Boost, LR, NB, and KNN machine learning models. Therefore, our model can classify customer churn for both data sets better than other machine learning classifiers. However, to ensure that our model works better than other deep learning techniques for both datasets, the comparison between our model results and other studies' results have been shown in the dissection.

The evaluation performance of the Deep-BP-ANN using 10-fold CV is shown in Figure 9. The results show that the best performance achieved by Deep-BP-ANN followed by XG Boost. The best ACC with a 10-fold CV evaluation achieved by Deep-BP-ANN, 86.57% for IBM Telco and Precision 81.59%, Recall 94.45%, F1-Score 87.55%, and AUC 86.57%.

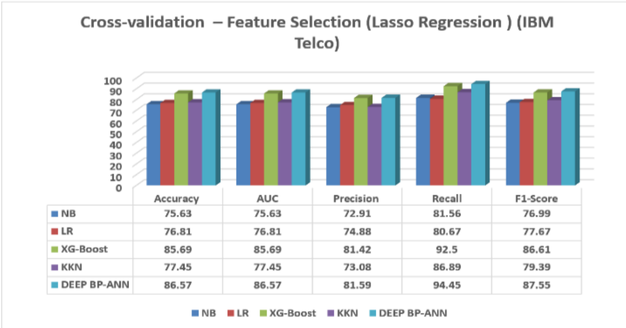

Fig. 9: The Evaluation Performance with Lasso Regression in IMB Telco (10-fold)

On the other hand, the evaluation performance of the 10-fold CV for Cell2cell in this experiment is shown in Figure 10.

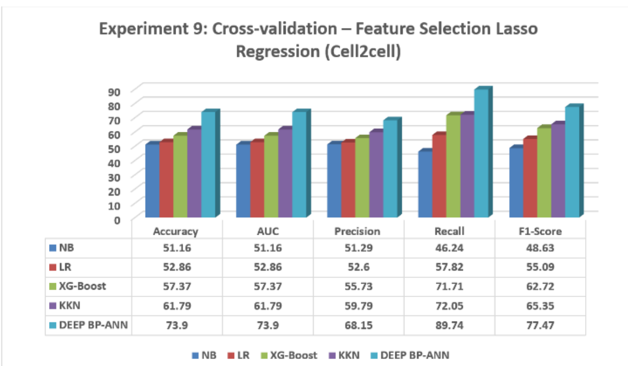

Fig. 10: The Evaluation Performance with Lasso Regression in Cell2cell (10-fold)

The results show that a great improvement is achieved by 10-fold CV for the Cell2cell dataset compared with current studies that applied deep learning on the Cell2cell dataset. The results show that the ACC achieved by Deep-BP-ANN is 73.09%. These improvements are achieved after applied the Lasso Regression technique for feature selection, increased neurons by 218 new neurons since it was 32, and select the nearest numbers of epochs achieved by the Early Stopping technique, which is 214.

8 Discussion

The summary of the two last experiments results obtained by Deep-BP-ANN by two different evaluation methods (Simple holdout and 10-fold CV) for both datasets is shown in Table 3.

The results show that the holdout method outperforms the 10-fold CV method because the holdout method is applied with the early stopping technique. Each time the model is training, it will stop training at the perfect time with the best test validation results. In contrast, in the 10-fold CV, although the number of epochs selected based on the nearest point chosen by early stopping technique when applied on holdout, these epochs are constant for the entire 10-fold, which means the model may not stop at the right time for each running and may lead to giving results that not the best. However, the early stopping technique can be applied with the 10-fold CV which will be applied in the future work. Nevertheless, the Deep-BP-ANN model results, either by Holdout or 10-fold CV, outperform the other research studies, and the evidence is shown below in Table 3.The results show that the holdout method outperforms the 10-fold CV method because the holdout method is applied with the early stopping technique. Each time the model is training, it will stop training at the perfect time with the best test validation results. In contrast, in the 10-fold CV, although the number of epochs selected based on the nearest point chosen by early stopping technique when applied on holdout, these epochs are constant for the entire 10-fold, which means the model may not stop at the right time for each running and may lead to giving results that not the best. However, the early stopping technique can be applied with the 10-fold CV which will be applied in the future work. Nevertheless, the Deep-BP-ANN model results, either by Holdout or 10-fold CV, outperform the other research studies, and the evidence is shown below in Table 3.

Table 2. The hold-out and cross-validation results obtain from Deep-BP-ANN in (IBM Telco and Cell2cell)

	Deep-BP-ANN results obtained from Holdout and 10-fold CV (IBM Telco)		*Deep-BP-ANN results obtained from Holdout and 10-fold CV* (Cell2cell)	
	Simple Holdout	**10-fold CV**	**Simple Holdout**	**10-fold CV**
Accuracy	88.12%	86.57%	79.38%	73.90%
Precision	84.71%	81.59%	74.50%	68.15%
Recall	93.05%	94.45%	89.32%	89.74%
F1-Score	88.68%	87.55%	81.24%	77.47%
AUC	88.11%	86.57%	79.38%	73.90%

The implemented Deep-BP-ANN model is compared with the other studies that used the same dataset and deep learning techniques. Table 4 summarize the recent studies with their achieved accuracy results that used the same dataset, either IBM Telco or Cell2cell, with deep learning techniques. The summary table shows that out of thirty studies about the CCP, five of them are compatible with our study, which used the same data set and deep learning techniques. Two of these studies worked with the Cell2cell dataset, and three of them worked with the IBM Telco dataset. Therefore, these studies are used to compare our model results with them since our contribution is to overcome deep learning techniques used to PCC in terms of accuracy.

Table 3. Summary of Compatible studies with our study

NO	AUTHORS & YEAR	DATASET	CLASSIFIERS	EVALUATION METHOD	ACCURACY
1	(Umayaparvathi & Iyakutti, 2017)	Cell2cell	CNN LFNN	10-fold CV	CNN and FNN – 71.66%
2	(U. Ahmed et al., 2019)	Cell2cell	Transfer Learning + CNN (TL+CNN)	10-fold CV	68.2%
3	(Momin et al., 2020)	IBM Telco	ANN	Simple holdout 10% for testing and 90% for training	82.83%
4	(Agrawal et al., 2018)	IBM Telco	multi-layered ANN	Simple Holdout 80% for training and 20% for testing.	80.03%
5	(Mohammad NI et al., 2019)	IBM Telco	ANN	Simple Holdout 70% for training and 30% for testing.	85.55%

Table 5 shows a comparison between existing deep learning models and implemented Deep-BP-ANN model for the Cell2cell dataset. The results show that our model outperforms CNN and FFN that implemented by [9]. The implemented Deep-BP-ANN predicted customer churn with ACC of 73.90%, which over their model by (2.24%) using the same evaluation method (10-fold CV). In another study, [12] used transfer learning and CNN on the Cell2cel and

orange datasets. Their obtained ACC result for Cell2cell is 68.2%, while our model ACC is 73.90% using the same evaluation method (10-fold CV). The implemented model outperforms the accuracy of this study with a rate of (5.52%). Moreover, [9] and [12] applied only the cross-validation method to evaluate the model while our model applied two evaluation methods (Holdout and 10-fold CV).

Table 4. The Comparison between Our Model and Other Studies for Cell2cell Dataset

	Existing Model ACC		Implemented model ACC	
Study Authors	***Holdout***	***10-fold CV***	***Holdout***	***10-fold CV***
(Umayaparvathi & Iyakutti, 2017)	Not applied	CNN: 71.66% LFNN: 71.66%	**79.38%**	**73.90%**
(Ahmed et al., 2019)	Not applied	TL+CNN: 68.2%	**79.38%**	**73.90%**

Table 6 compares existing deep learning models' accuracy and our model results with the IBM Telco dataset. The results show that three different recent studies used ANN deep learning to predict customer churn. [13] develop an ANN model, which achieved an accuracy of 82.83%, while our model achieved an accuracy of 88.12% using the same evaluation method (Holdout).

Table 5. The recent studies used deep learning to PCC

	Existing Model ACC		Implemented model ACC	
Study Authors	***Holdout***	***10-fold CV***	***Holdout***	***10-fold CV***
(Momin et al., 2020)	ANN: 82.83%	Not applied	**88.12%**	**86.57%**
(Agrawal et al., 2018)	ANN: 80.03%	Not applied	**88.12%**	**86.57%**
(Mohammad et al., 2019)	ANN: 85.55%	Not applied	**88.12%**	**86.57%**

The implemented Deep-BP-ANN outperforms the accuracy of ANN that was implemented by [13] with a rate of (5.29%) using holdout and with a rate of (3.74%) using 10-fold CV. Another study used ANN by [15] achieved an accuracy of 80.03%, while our model over their accuracy by (8.09%) using the same evaluation method (holdout). Also, the implemented model outperforms their accuracy results using 10-fold CV by (6.54%). [16] applied ANN in their study and achieved a good accuracy result, which is 85.55%, but still, our model accuracy result outperforms their result by a rate of (2.57%) using the same evaluation method (holdout). Moreover, the implemented Deep-BP-ANN model outperforms their result with a rate of (1.02%) using 10-fold CV, and their study not applied this evaluation method.

9 Conclusion

Customer churn is a critical issue that needs to be analysed and predict churn before it occurs to help companies, especially in the telecom industry to avoid the reasons that cause the churn. Thus, the company will retain its customers and attract new ones since it has a good reputation obtained from loyal customers. Therefore, a model that can predict customer churn and deal with a huge amount of data is significant.

Today, companies deal with billions of data that cannot be handled by machine learning, so deep learning techniques are essential. Most of the current studies recognized how important customer churns predicting models are, especially in the telecom industry. Many researches have been starting applying deep learning algorithms to build a model that can predict customer churn accurately. However, there is still a gap in the existing models that used deep learning techniques since their accuracy needs to be improved. Herein lies the importance of our research in implementing a deep learning model that can predict customer churn with high accuracy compared with existing models and deal with a huge amount of data.

In this study, the Deep-BP-ANN model has been implemented to predict customer churn and compared its performance with four states of art machine learning techniques: XG Boost, LR, NB, and KNN. Moreover, the obtained accuracy compared with other studies that used deep learning techniques. Experiments were conducted using two real-world datasets IBM Telco (7043 customers) and Cell2Cell (51047 customers). Experiments results showed that the implemented Deep-BP-ANN outperforms the ACC, Precision, Recall, and F1-Score machine learning classifiers. Furthermore, our model outperforms the other studies that used the deep learning techniques in terms of the accuracy for both datasets.

The Lasso regularization has been used as a feature selection technique to delete non-important attributes from the datasets. This technique provides great benefits when the datasets contain a large number of features. In our model, the Lasso regression technique greatly impacts the performance, especially for the Cell2cell dataset, since it contains more features than IBM Telco.

One advantage of using sets of decision tree techniques such as “gradient boosting” is that the trained predictive model can accurately estimate the features’ importance. In this study, the XG Boost algorithm has been used to identify which features have impacted the churn prediction. The results show that the most significant features that impact the CCP are total charge, tenure (the number of months the customers have stayed with the company), the percentage change in minutes of use, and the percentage change in revenues.

In this study, the ROS technique has been applied to balance both datasets, and the Activity regularization technique is used to reduce the overfitting that may happen because of the ROS. The results show effective performance by the Deep-BP-ANN model in ACC, Recall, Precision, and F1-Score obtained with these techniques.

Moreover, the tuning parameters of the implemented model have a strong impact on improving the performance, such as increase the number of neurons, minimize the number of hidden layers, select an appropriate batch size, and the number of epochs. These parameters can be tuned with multiple trials to achieve the best selections except the number of epochs. To select the best number of epochs, the early stopping technique has been used, which can stop training after a particular number of epochs under the condition of obtaining the best test validation results.

The proposed model achieved the best results within 250 neurons, one hidden layer including input and output layers, 32 batch size, and 4000 epochs. The early stopping technique almost stops with approximately 211 to 214 epochs for both datasets.

Deep-BP-ANN has obtained the best results for IBM Telco using the holdout method with an accuracy of 88.12%, Precision 84.71%, Recall 93.05%, F1-Score 88.68%, and AUC 88.11%. On the other hand, using 10-fold CV, the best performance achieved by Deep-BP-ANN too, where ACC is 86.57%, Precision 81.59%, Recall 94.45%, F1-Score 87.55%, and AUC 86.57%. Moreover, the Deep-BP-ANN model also achieved the best performance for the Cell2cell dataset. The highest ACC using hold out method is 79.38%, Precision 74.50%, Recall 89.32%, F1-score 81.24%, and AUC 79.38%. The 10-fold CV method also gives the best performance by Deep-BP-ANN with an ACC of 73.90%, Precision 68.15%, Recall 89.74%, 77.47%, and AUC 73.90%.

The obtained results either for IBM Telco or Cell2Cell dataset and either using holdout or cross-validation are the best compared with the other existing studies that used deep learning techniques such as CNN, LFNN, and ANN.

10 Recommendation

In general, IBM Telco achieved an accuracy higher than the Cell2cell dataset, although the IBM Telco has fewer records than Cell2cell, and it commonly known that deep learning works well with a large amount of data. IBM Telco contains more significant features related to the customer churn than the features in the Cell2cell dataset.

Most researches select the Cell2cell dataset when they work with deep learning because it contains many records compared with other available datasets. Before depending on a large amount of data, the researchers must know that deep learning works well with data sets that contain significant features to the target. Due to that, this study recommend that researches to apply feature importance techniques such as XG Boost as we did earlier and visualize the data based on the dependent and independent variables to decide which data set contains more significant and relevant features.

Moreover, deep learning, such any machine learning techniques, works well with a balanced dataset. It is tough to obtain a balanced dataset when dealing with the customer churn problem because most customers are not churning. This study recommends that the telecom industry to take the required churned customer data from the history or archive to balance the dataset used to train the model. Another way, by using balance techniques such as the ROS technique.

Furthermore, the telecom industry should focus on the Bill Payment Analysis details besides how long the customer belongs to the company to build a strong model that can predict customer churn. Besides that, the telecom industry should give the customer a good offer to convince them to subscribe with one or two years because our results show that the customer with a short contract is more likely to churn.

The telecom industry should give a good discount for the customers who are subscribed with fiber optic service because our results show that this group of customers is more likely to churn than others. The telecom industry should avoid electronic check methods while paying the financial receivables and guide the customers to use mailed check, bank transfer, or credit card. Since the electronic check is in an electronic format, it can be processed in some steps, and the customer may get confused. The results prove that the customers who follow the payment methods as electronic check are most likely to churn. Also, the telecom industry should give more attention to technical support since the customers who are not serving with technical support are more likely to churn.

11 Future Work

In future work, the early stopping technique will be applied to the 10-fold cross-validation to achieve the higher accuracy as achieved with the hold-out evaluation method. Moreover, the implemented model will perform on more telecom datasets that contain a huge amount of data, such as millions of records, to obtain higher accuracy since deep learning works well with more data.

Compare the model performance with more machine learning techniques such as random forest and decision tree.

The implemented model will develop into a web application using the flask technique and deploy it using Heroku Service to benefit the telecommunication industries.

Acknowledgement

The authors are grateful to the Ahlia University, Bahrain and Applied Science University, Bahrain, for the full financial support granted to this research paper.

Competing interests

The author declares that she has no competing interests.

References

[1] B. Gates, Why word is used, Microsoft Lett., **69(5)**, 53-60, 1998.

[1] Berkhout, A. J., Van Der Duin, P. A. (2007). New ways of innovation: An application of the cyclic innovation model to the mobile telecom industry. International Journal of Technology Management., **40(4)**, 294–309, 2007.

[2] Khder, M. A., Fujo, S. W., Sayfi, M. A. (2021). A roadmap to data science: background, future, and trends. International Journal of Intelligent Information and Database Systems., **14(3)**, 277-293, 2021.

[3] Perez, M. J., Flannery, W. T. (2009). A study of the relationships between service failures and customer churn in a telecommunications environment. PICMET: Portland International Center for Management of Engineering and Technology, Proceedings, October 2008, 3334–3342, 2009.

[4] PK, D., SK, K., A, D., A, B., VA., K. (2016). Analysis of Customer Churn Prediction in Telecom Industry using Decision Trees and Logistic Regression. In 2016 Symposium on Colossal Data Analysis and Networking (CDAN) (Pp. 1-4). IEEE., **88(5)**, 4, 2016.

[5] Mishra, A., Reddy, U. S. (2018). A comparative study of customer churn prediction in telecom industry using ensemble based classifiers. Proceedings of the International Conference on Inventive Computing and Informatics, ICICI 2017, Icici, 721–725, 2018.

[6] Amin, A., Al-Obeidat, F., Shah, B., Adnan, A., Loo, J., Anwar, S. (2019). Customer churn prediction in telecommunication industry under uncertain situation. Journal of Business Research., **94**, 290–301, 2019.

[7] Kloutix. (2018). Are your best telecom customers leaving you? https://www.citationmachine.net/apa/citea-website/new

[8] Ahmad Naz, N., Shoaib, U., Shahzad Sarfraz, M. (2018). a Review on Customer Churn Prediction Data Mining Modeling Techniques. Indian Journal of Science and Technology., **11(27)**, 1–7, 2018.

[9] Umayaparvathi, V., Iyakutti, K. (2017). Automated Feature Selection and Churn Prediction using Deep Learning Models. International Research Journal of Engineering and Technology(IRJET)., **4(3)**, 1846–1854, 2017.

[10] Deng, L., Yu, D. (2013). Deep learning: Methods and applications. Foundations and Trends in Signal Processing., **7(3–4)**, 197–387, 2013.

[11] Wehle, H. (2017). Machine Learning, Deep Learning, and AI: What's the Difference? In International Conference on Data Scientist Innovation Day, Bruxelles, Belgium., July.

[12] Ahmed, U., Khan, A., Khan, S. H., Basit, A., Haq, I. U., Lee, Y. S. (2019). Transfer Learning and Meta Classification Based Deep Churn Prediction System for Telecom Industry., 1–10, 2019.

[13] Momin, S., Bohra, T., Raut, P. (2020). Prediction of Customer Churn Using Machine Learning. EAI International Conference on Big Data Innovation for Sustainable Cognitive Computing., 203–212, 2020.

[14] Amin, A., Al-obeidat, F., Shah, B., Adnan, A., Loo, J., Anwar, S. (2018). Customer churn prediction in telecommunication industry using data certainty Customer churn prediction in telecommunication industry using data certainty. Journal of Business Research., 2019, 1–12, 2018..

[15] Agrawal, S., Das, A., Gaikwad, A., Dhage, S. (2018). Customer Churn Prediction Modelling Based on Behavioural Patterns Analysis using Deep Learning. 2018 International Conference on Smart Computing and Electronic Enterprise (ICSCEE)., 1–6.

[16] Mohammad NI, SA, I., MN, K., OM, Y., A, A. (2019). Customer Churn Prediction In Telecommunication Industry Using Machine Learning Classifiers. Proceedings of the 3rd International Conference on Vision, Image and Signal Processing., 1–7, 2019.

[17] Pamina, J., Beschi Raja, J., Sathya Bama, S., Soundarya, S., Sruthi, M. S., Kiruthika, S., Aiswaryadevi, V. J., Priyanka, G. (2019). An effective classifier for predicting churn in telecommunication. Journal of Advanced Research in Dynamical and Control Systems., **11(1 Special Issue)**, 221–229.

[18] Lecun, Y., Bengio, Y., Hinton, G. (2015). Deep learning. Nature., **521(7553)**, 436–444, 2015.

[19] Keramati, A., Jafari-Marandi, R., Aliannejadi, M., Ahmadian, I., Mozaffari, M., Abbasi, U. (2014). Improved churn prediction in telecommunication industry using data mining techniques. Applied Soft Computing Journal., **24**, 994–1012

[20] Sundararajan A, Gursoy K. (2020). Telecom customer churn prediction. March. https://doi.org/10.7282/t3-76xmde75.

[21] Ahmad, A. K., Jafar, A., Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in big data platform. Journal of Big Data., **6(1)**, 2019.

[22] Xiong, A., You, Y., Long, L. (2019). L-RBF: A customer churn prediction model based on lasso + RBF. Proceedings - 2019 IEEE International Congress on Cybermatics: 12th IEEE International Conference on Internet of Things, 15th IEEE International Conference on Green Computing and Communications, 12th IEEE International Conference on Cyber, Physical and So., 621–626, 2019.

[23] De Caigny, A., Coussement, K., De Bock, K. W. (2018). A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. European Journal of Operational Research., **269(2)**, 760-772, 2018.

[24] Idris, A., Khan, A., Soo, Y. (2013). Intelligent churn prediction in telecom: employing mRMR feature selection and RotBoost based ensemble classification. Applied Intelligence., **93(3)**, 659–672, 2013.

[25] Idris, A., Khan, A. (2012). Genetic Programming and Adaboosting based churn prediction for Telecom. IEEE International Conference on Systems, Man, and Cybernetics (SMC), October, 1328–1332.

[26] Neslin, S. (2002). Cell2Cell: The churn game. In Hanover, NH: Tuck School of Business, Dartmouth College.

[27] Li, Q., Xiong, Q., Ji, S., Wen, J., Gao, M., Yu, Y., Xu, R. (2019). Using fine-tuned conditional probabilities for data transformation of nominal attributes. Pattern Recognition Letters., **128**, 107–114, 2019..

[28] Xie, J., Li, Y., Wang, N., Xin, L., Fang, Y., Liu, J. (2020). Feature Selection and Syndrome Classification for Rheumatoid Arthritis Patients with Traditional Chinese Medicine Treatment. European Journal of Integrative Medicine, 34(October 2019), 101059.

[29] Karanovic, M., Popovac, M., Sladojevic, S., Arsenovic, M., Stefanovic, D. (2018). Telecommunication Services Churn Prediction - Deep Learning Approach. 2018 26th Telecommunications Forum, TELFOR 2018 - Proceedings - IEEE, January 2019.

[30] Wadikar, D. (2020). Customer churn prediction [Technological University Dublin]. In Masters Dissertation. Technological University Dublin.

[31] Kumar, A. (2020). Imputing Missing Data Using Sklearn SimpleImputer - DZone AI. Zone.Com. [32] Hale, J. (2020). Smarter Ways to Encode Categorical Data for Machine Learning. Medium. https://towardsdatascience.com/smarter-ways-to-encodecategorical-data-for-machine-learning-part-1-of-3-6dca2f71b159

[33] Pires, I. M., Hussain, F., Garcia, N. M., Lameski, P., Zdravevski, E. (2020). Homogeneous data normalization and deep learning: A case study in human activity classification. Future Internet., **12(11)**, 1–14, 2020.

[34] ZHOU, J., YAN, J., YANG, L., WANG, M., XIA, P. (2019). Customer Churn Prediction Model Based on LSTM and CNN in Music Streaming. DEStech Transactions on Engineering and Technology Research, aemce, 254–261.

[35] Zhu, B., Baesens, B., Seppe, K. L. M. (2017). An empirical comparison of techniques for the class imbalance problem in churn prediction. Information Sciences., **408**, 84–99, 2017.

[36] Kuhn, M., Johnson, K. (2013). Applied Predictive Modeling. Springer. https://link.springer.com/book/10.1007

[37] Zhu, B., Guicai, X., Yuan, Y., Yiqin, D. (2018). Investigating Decision Tree in Churn Prediction with Class Imbalance. In Proceedings of the International Conference on Data Processing and Applications., 11–15, 2018.

[38] Fern´andez, A., Garc´ıa, S., Galar, M., Prati, R. C., Krawczyk, B., Herrera, F. (2018). Learning from Imbalanced Data Sets. Springer Berlin. https://link.springer.com/book/10.1007

[39] Amin, A., Anwar, S., Adnan, A., Nawaz, M., Howard, N., Qadir, J., Member, S., Hawalah, A., Hussain, A., Member, S. (2016). Comparing Oversampling Techniques to Handle the Class Imbalance Problem: A Customer Churn Prediction Case Study. IEEE Access., **4(Ml)**, 7940–7957, 2016.

[40] Madeline, M. (2019). Intro to Feature Selection Methods for Data Science. Medium. https://towardsdatascience.com/intro-to-feature-selectionmethods-for-data-science-4cae2178a00a

[41] Bhandari. (2015). Towards an Efficient Method of Modeling "Next Best Action" for Digital Buyer's Journey in B2B—In:

Machine Learning and Data Mining in pattern Recognition.

[42] Kumarappan, S. (2020). Feature Selection by Lasso and Ridge Regression-Python Code Examples. Medium. https://medium.com/@sabarirajan.kumarappan/featureselection-by-lasso-and-ridge-regression-python-codeexamples-1e8ab451b94b

[43] Ahn, J. H., Han, S. P., Lee, Y. S. (2006). Customer churn analysis: Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry. Telecommunications Policy., **30(10–11)**, 552–568, 2006.

[44] Hastie, T., Tibshirani, R. Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Science Business Media. https://silo.pub/the-elements-of-statistical-learning-datamining-inference-and-prediction-2nd-edition-springerseries-in-statistics.html

[45] Brownlee, J. (2020). A Gentle Introduction to Imbalanced Classification. Machine Learning Mastery. https://machinelearningmastery.com/what-is-imbalancedclassification/::text=Another way to describe the,percent in a third class.

[46] Krishnan, H. (2020). Confusion Matrix, Accuracy, Precision, Recall, F1 Score. Medium. https://medium.com/analytics-vidhya/confusion-matrixaccuracy-precision- recall-f1-score-ade299cf63cd

[47] Wang, C., Han, D., Fan, W., Liu, Q. (2019). Customer Churn Prediction with Feature Embedded Convolutional Neural Network: An Empirical Study in the Internet Funds Industry. International Journal of Computational Intelligence and Applications., **18(1)**, 1–19, 2019.

[48] Zhong, J., Li, W. (2019). Predicting customer churn in the telecommunication industry by analyzing phone call transcripts with convolutional neural networks. ACM International Conference Proceeding Series, Part F1481(March)., 55–59, 2019.

[49] Bisong, E. (2019). Regularization for Deep Learning. In: Building Machine Learning and Deep Learning Models on Google Cloud Platform. (C. Apress, Berkeley (ed.)). Springer.

[50] Kazi, M., Eljack, F., Mahdi, E. (2020). Predictive ANN Models for Varying Filler Content for Cotton Fiber / PVC Composites based on Experimental Load Displacement Curves. Composite Structures, August., 112885, 2020.

[51] Jajodia, T., Garg, P. (2019). Image Classification – Cat and Dog Images. International Research Journal of Engineering and Technology (IRJET)., **06(12)**, 570–572, 2019.

[52] Liang, X., Angelopoulou, A., Kapetanios, E. (2020). A Multi-modal Machine Learning Approach and Toolkit to Automate Recognition of Early Stages of Dementia Among British Sign Language Users. Bartoli A., Fusiello A. (Eds) Computer Vision – ECCV 2020 Workshops. ECCV 2020. Lecture Notes in Computer Science., **12536**, 278–293, 2020.

[53] Xu, Y., Goodacre, R. (2018). On Splitting Training and Validation Set: A Comparative Study of Cross - Validation , Bootstrap and Systematic Sampling forEstimating the Generalization Performance of Supervised Learning. Journal of Analysis and Testing., **2(3)**, 249–262, 2018.