Journal of Engineering Technology and Applied Physics

CRATSM: An Effective Hybridization of Deep Neural Models for Customer Retention Prediction in the Telecom Industry

Johnson Olanrewaju Victor[1], XinYing Chew[1,*], Khai Wah Khaw[2] and Zhi Lin Chong[3]
[1]*School of Computer Sciences, Universiti Sains Malaysia, 11800 Pulau Pinang, Malaysia*.
[2]*School of Management, Universiti Sains Malaysia, 11800 Pulau Pinang, Malaysia*.
[3]*Department of Electronic Engineering, Faculty of Engineering and Green Technology, Universiti Tunku Abdul Rahman (UTAR), 31900, Kampar Perak, Malaysia*.
**Corresponding author*: xinying@usm.my, *ORCiD*: 0000-0001-5539-1959
https://doi.org/10.33093/jetap.2024.6.2.10
Manuscript Received: 8 April 2024, Accepted: 5 June 2024, Published: 15 September 2024

***Abstract* — In the dynamic field of Customer Retention Prediction (CRP), strategic marketing and promotion efforts targeting specific customers are crucial. Understanding customer behavior and identifying churn indicators are vital for devising effective retention strategies. However, identifying customers likely to terminate services presents a challenge, leading to data imbalance issues. Existing CRP studies using Machine Learning (ML) techniques and data imbalance methods face problems such as overfitting and computational complexity. Similarly, recent CRP studies employing Deep Learning (DL) approaches rely on data sampling techniques, which can result in overfitting and a lack of cost sensitivity. Additionally, DL approaches struggle with slow convergence and get stuck in local minima. This paper introduces an effective hybrid of Deep Learning (DL) classifiers focusing on cost-metric integration to address data imbalance issues and period-shift Cosine Annealing Learning Rate (ps-CALR) to accelerate model training, ultimately enhancing performance. Three Telecom datasets, namely IBM, Iranian, and Orange, were used to assess the model performance. Empirical findings show that the hybrid DL classifiers significantly improved CRP over conventional ML. This paper contributes methodological advancements and practical insights for effective customer retention in the telecom industry.**

Keywords*— *Optimization, Residual Network, Attention Mechanism, Tree-structured Network, Cost-sensitive.

I. INTRODUCTION

Customer Retention Prediction (CRP) is a pivotal domain of inquiry within the telecommunications sector, given its direct influence on business profitability and longevity. The exponential surge in digital transformation underscores this importance witnessed globally over the past decade, with a significant portion of the global population relying on telecommunication operators for connectivity. For example, in Malaysia, Global Monitor reported out of 31.83 million individuals, 25.08 million were Internet users in 2018, with the e-commerce sector projected to exceed 20.8% growth by 2020 [1]. The telecom market in Malaysia alone should reach RM38.81 billion by 2023 and a forecast of RM42.27 billion by 2028. Internet users have surpassed 5.19 billion globally, representing 64.6% of the global population [2]. Mobile telecommunications primarily drive this surge in Internet penetration. The telecom industry boasts numerous service providers catering to the diverse needs of consumers. In Malaysia, key players include Maxis, Digi, Celcom, Edotco, and Sacofa Sdn Bhd. These projections highlight the vast market opportunities available to telecom operators and ancillary service providers.

However, customer churn remains a pressing concern for telecom companies, posing significant risks such as revenue loss and diminished customer base. Providers must meticulously analyze factors contributing to customer churn, including dissatisfaction, evolving digital experiences, and

intense market competition. Infrastructure constraints, network failures, poor responsiveness, and software glitches further compound these challenges. Consequently, understanding customer behavior and predicting churn becomes imperative for devising effective marketing strategies and preempting attrition. Machine learning (ML) techniques offer promising avenues for addressing customer retention challenges, as evidenced in many recent studies [3–7]. By leveraging ML-driven predictive analytics, telecom companies can mitigate the risks associated with customer attrition, enhance business intelligence, and foster sustainable growth in an increasingly competitive landscape.

CRP studies encompass a diverse array of focused investigations, which have delved into key areas such as i) user satisfaction [8], ii) comprehensible and interpretability [9], iii) behavioral pattern identification [5, 10], iv) customer segmentation [11], v) profit-driven/maximization [12], vi) customer lifetime value [13, 14], and vii) social network Influence [15, 16]. Also, recent developments and studies in Deep Learning (DL) techniques have proven successful in their application to customer retention prediction [17–20].

Moreover, two things have been significant with CRP research, especially within the telecom industry over the years: the structured nature of customer data and the inherent class imbalance, where the number of customers leaving is typically fewer than those staying. These are perceived as challenges that have long posed obstacles to ML techniques, prompting a research focus on Imbalance Learning (IL). IL solutions are typically categorized into three major classes: data-level, algorithmic-level, and cost-sensitive techniques [21]. While IL has garnered substantial attention in the literature, there remains a growing interest in exploring DL techniques to address IL challenges by augmenting existing approaches or devising entirely new solutions. On the other hand, some scholars have raised arguments regarding the suitability of DL techniques for tabular data problems [22, 23]. They argue that DL architectures, designed to capture inductive biases matching the invariances and spatial dependencies of data, may struggle to identify corresponding invariances in tabular data due to its heterogeneous features, small sample sizes, and extreme values.

However, it is worth noting that no single ML method is universally applicable to all problems, and extensive experimentation is crucial for determining efficacy. In addition to some studies contesting the claims as mentioned earlier [24, 25], the development of tabular-specific DL architectures has emerged as a vibrant area of research, poised to shape the future of CRP. Fayaz *et al.* [25] advocate integrating DL with conventional ML techniques, suggesting potential performance gains from such a combination. Previously, Umayaparvathi and Iyakutti [19] asserted that the efficacy of existing ML methods heavily relies on feature selection techniques, significantly impacting model performance. Feature selection is characterized by computational expense, time consumption, and labor intensiveness and often necessitates domain expertise. DL, on the other hand, promises to automate feature extraction and perform prediction tasks. Building upon these insights, Li *et al.* [18] proposed a hybrid DL model that leverages One-Dimensional Convolutional Neural Networks (1D-CNN) in conjunction with Gradient Boosting Decision Trees (GBDT). The 1D-CNN component is primarily employed for automatic feature extraction and mapping, while GBDT generates the final predictions. Evaluating the model's performance on two telecom datasets, they reported F1-scores of 0.6724 and 0.6122, respectively.

The big question arises: for DL models to compete effectively and maintain relevance alongside conventional ML methods in CRP, how can their design be tailored to achieve enhanced performance, especially amidst the ongoing advancements and innovations within the DL domain? This paper, therefore, introduces a novel hybrid Deep Neural Network (DNN) architecture consisting of a Convolutional-Residual-Attention (CRA) block and a Tree-Structured Multilayer (TSM) architecture called CRATSM. The CRATSM employs period-shift cosine annealing to accelerate the Learning Rate (LR) during training. Notably, this proposed framework represents a unique addition to the existing state-of-the-art CRP models in the telecom industry, offering a fresh and innovative approach. The contributions of this paper are as follows:

i. Development of a Residual-Attention mechanism to improve the base Convolutional Neural Network (CNN) model and implement Tree-Structured techniques to enhance the Multilayer Perceptron (MLP) model.

ii. Optimizing CRATSM model training process using period-shift Cosine Annealing Learning Rate (ps-CALR) to improve its generalization and faster convergence.

iii. Adaptation of the hybrid model as a cost-sensitive IL approach to CRP.

The rest of the paper is organized as follows. Section II provides a review of the literature. Section III outlines the methodology. Section IV presents the empirical results. Finally, Section V concludes the paper.

II. LITERATURE REVIEW

The dynamic nature of the business landscape, particularly for service providers, poses a continuous challenge to find innovative ways to rebrand and redefine their products and services within the competitive market, ensuring they meet customers' expectations and maintain relevance over time [26]. In this context, Customer Relational Management (CRM) systems offer a valuable and cost-effective means to enhance customer experience through data-driven strategies and concepts to achieve business objectives.

One of the critical aspects within the framework of CRM is customer retention, which involves the company's efforts to attract and retain its existing customers (acquiring new customers is 5-6 times more than maintaining them). CRP analyzes customers' behavioral patterns and associated attributes believed to influence these patterns. Therefore, customer data is structured to include both explanatory and target variables, providing insights into the reasons behind customer churn. Customer churn can be attributed to either satisfaction or dissatisfaction with the provided services [18]. Hence, ML offers invaluable prospects in this regard, especially supervised ML.

There are three ML categories: supervised, unsupervised, and reinforcement learning. In supervised learning, the focus lies on training and estimating models using labeled datasets, represented as, $(x_{11}, y_1), (x_{22}, y_2), \ldots (x_{nm}, y_n)$, where x_m denotes explanatory variables, and y_n represents the target label. A target label comprised of numerical values constitutes a prediction problem. Conversely, a categorical label forms a classification task (binary or multiclass classification) [5]. The literature extensively examines churn prediction using supervised and unsupervised methods [5, 10, 17]. Researchers have also delved into a mixed approach known as semi-supervised learning for churn prediction [27]. Furthermore, reinforcement learning, a branch of ML, has also been applied to churn prediction tasks [28]. These highlights show the adaptability of ML techniques, including DL [17, 18], in addressing challenges related to churn prediction and emphasize the importance of exploring various learning paradigms.

Regarding CRP as primarily an Imbalanced problem, Amin *et al.* [29] conducted a comparative study on oversampling techniques, such as the Synthetic Minority Over-sampling Technique (SMOTE), Mega-Trend-Diffusion Function (MTDF), Adaptive Synthetic Sampling approach (ADASYN), and others. Their experiments revealed that MTDF combined with a genetic algorithm performed best for handling data imbalance in CRP. Xiaojun and Sufang [30] proposed an improved SMOTE technique, SMOTE+AdaBoost, which achieved superior performance compared to other classifiers. Meanwhile, Mqadi *et al.* [31] highlighted the challenge of information loss in undersampling methods like NearMiss. Recent studies have shown promise in synthetic data generation for ML problems, and hybrid data sampling methods have been proposed, albeit with additional costs and data loss. Another approach to address data imbalance is Cost-Sensitive Learning (CSL), which imposes higher penalties for misclassifying the positive class. Various algorithms, such as AdaBoost variants [32] and Bayesian models, have been adapted for CSL. However, CSL methods may face challenges such as context dependency and increased model complexity. Xiao *et al.* [27] combined semi-supervised learning with the Metacost method and random ensemble subspace to improve CRP performance.

Addressing CRP with ML techniques, Kim *et al.* [33] studied CRP using ML techniques by proposing an MLP approach for churn prediction in the telecom industry. Their study incorporated a network variable to analyze its influence on churn prediction among subscribers, revealing improved predictive performance. However, a significant limitation of the research was the absence of hyperparameter tuning for the neural network. Umayaparvathi and Iyakutti [19] introduced three DL methods to explore churn prediction in telecom. Their study aimed to enhance churn model performance by efficiently selecting customer features. The authors demonstrated that DL enables automatic feature extraction, surpassing manual feature engineering in conventional ML approaches. They achieved favorable results without relying on feature engineering, using two telecom datasets with stratified 10-fold cross-validation to address data imbalance. Sniegula *et al.* [34] presented Neural Network (NN) churn prediction in telecom, comparing it with Decision Trees (DT) and K-means. They conducted numerous tests with varying NN configurations, revealing the minimal impact of layer variations on accuracy. Despite achieving comparable results with other classifiers, the study overlooked the effect of different learning rates on optimization. Chouiekh and El Haj [35] investigated consumer behavior in telecom, employing CNN to model rank features derived from pixel-based Call Detail Records (CDR). Their study demonstrated the superior performance of a hybrid CNN-Random Forest (RF), Gradient Boosting (GB), and Support Vector Machine SVM) approach, achieving a notable F1-score of 91% with CONV-RF, suggesting the potential of CNNs to provide understandable and actionable insights.

Furthermore, Kumar and Kumar [20] optimized an Artificial Neural Network (ANN) model by adjusting batch sizes, hidden layers, and epochs, achieving improved performance with specific hyperparameter settings. Domingos *et al.* [36] further emphasized the importance of hyperparameter optimization in churn prediction, considering various parameters such as batch sizes, Activation Functions (AFs), and training optimizers for MLP and DNN models, showcasing their beneficial impact on DL techniques. Similarly, Amatare and Ojo [17] investigated customer churn behavior using MLP and CNN models, aiming to eliminate the need for manual feature selection. Their study, based on data from a major cellular company in Nigeria, compared variants of each model, with MLP achieving 80% and 81% accuracy and CNN achieving 81% and 89% accuracy, respectively. Li *et al.* [18] proposed a hybrid DL model for churn prediction, combining a 1D-CNN for automatic feature extraction with GBDT for final prediction. Their model, trained on two telecom datasets, achieved competitive F1 scores of 67.24% and 61.22%, respectively, compared to traditional ML methods. Tan *et al.* [37] introduced several base classifiers (CNN, LR, DT, and SVM) and combined them using a stacking ensemble technique. This stacking ensemble model harnesses the individual strengths of each base classifier and integrates their collective knowledge through a meta-

learner to enhance prediction accuracy. The study uses the Cell2Cell dataset to assess the effectiveness of the models. The experimental findings indicate that the proposed model outperforms others in churn prediction, achieving a 62.4% f1-score and a 60.62% recall rate.

The literature also underscores the importance of optimizing DL models, particularly regarding batch size and learning rate (LR) [38]. Achieving optimal performance in DL models necessitates careful selection of the LR to ensure both faster convergence and robust generalization during training [39]. The choice of LR is crucial due to the significant challenge posed by local minima traps in DL optimization [40]. Johnson *et al.* [39] proposed an LR method, ps_CALR, which accelerates model training by thoroughly exploring the global minimum loss region. Findings show that the ps_CALR method outperforms the existing Cosine Annealing (CALR) technique, resulting in better model performance.

Comprehending the complexities of customer behavior is crucial for successful CRP using ML techniques. This understanding is the cornerstone for constructing resilient models that anticipate and accommodate different customer segments' varied needs and preferences. Consequently, it leads to improved customer satisfaction and loyalty.

III. METHODOLOGY

In this section, the paper employs the CRoss Industry Standard Process for Data Mining (CRISP-DM) model to explain the design stages of the proposed model.

A. Datasets

This paper utilizes three datasets for the study of CRP), namely IBM Watson, Iranian, and Orange, all sourced from Kaggle. An overview of the datasets is presented in Table I. Ethical considerations regarding the suitability of these datasets for the CRP task include their reflection of the diversity of customer experiences across telecom companies, their varying complexities in terms of numerical and categorical attributes, their accessibility as open-source data without any corporate or individual infringements, and their reflection of close-to-real-life scenarios.

Table I. Overview of the Datasets Used.

Datasets	Number of Instances	Features
IBM	7043	33
Iranian	3150	16
Orange	3333	21

B. Data Preprocessing

Dataset Description: In the case of the IBM dataset, an attribute “total charges” is wrongly defined as a string value (object). The “total charges” attribute was converted to numeric using the pandas’ to_numeric function. In regards to Iranian and Orange datasets, all the attributes are well-defined.

Missing Value: Missing values occur when attribute values for some instances are absent in a dataset, often denoted as "NA" or "_." There are two common approaches to address missing values. Missing values approaches include attributes or instances with missing values removal and imputation techniques [41]. In this paper, mean-value imputation was utilized to address missing values. Specifically, the IBM dataset contained missing values in the "Total charges" and "Churn Reason" attributes, with 11 and 5174 occurrences, respectively. To fix the missing values in the "Total charges" attribute, the computation involved multiplying "monthly charges" by "tenure months" and subtracting the result from "Total charges" to obtain the difference in charges for each customer. At the same time, the "Churn Reason" was dropped.

Categorical Variable Encoding: Categorical variable values were transformed into a numeric sparse form using dummy encoding to facilitate model fitting. The "get dummies" function in Python pandas' library was employed, with the parameter "drop_first" set to 'False' to ensure that all levels appear in the dataset.

Normalization: ML algorithms perform more effectively when training or testing set data values follow a normal distribution. However, some variables may have higher magnitude values in real-life datasets than others, leading to bias in predictive models [42]. A standardization technique is applied in the paper to scale the data. The standardization technique, using a Z-score, is defined in Eq. (1) as:

$$x'_{i,m} = \frac{x_{i,m} - \mu_i}{\sigma_i}, \quad (1)$$

where $x'_{i,m}$ is the transformed instances from $x_{i,m}$ with mean, μ_i and variance σ_i

Outlier Detection: Outliers, which are instances diverging significantly from the overall pattern in datasets, were detected using the interquartile range (IQR) rule as defined in Eq. (2) and Eq. (3). Mean imputation was then employed to address the outliers.

$$IQR = Q_3 - Q_1 \quad (2)$$

$$Q_1 - 1.5 * IQR \leq x'_{i,m} \leq Q_3 + 1.5 * IQR, \quad (3)$$

where Q_3 and Q_1 are third and first quartile, respectively.

Dropping Irrelevant Features: The following features were dropped in the IBM dataset, including "Country," "State," "City," "Count," "Zip Code," "Churn Reason," "Churn Score," "Churn Value," "CLTV," "CustomerID," "Lat Long," "Latitude," and "Longitude." For the Iranian dataset, the dropped attributes are "age", "FN", and "FP." No attributes were dropped for the Orange dataset.

C. Proposed Model Approach

Method I: The proposed hybrid CRATSM model combines enhanced 1D-CNN and MLP architectures with ps_CALR to accelerate the training process. The enhanced 1D-CNN model, RCA, builds upon the ID-CNN architecture defined in Eq. (4), incorporating additional Residual layers and Attention techniques. RCA starts with input X and processes it through convolutional layers with increasing filter sizes and kernel dimensions. Each layer is paired with ReLU activations, max pooling, and dropout for regularization. These layers then extract and downsample local features to progressively capture more complex patterns in the data.

A Residual layer with skip connections follows these layers to facilitate the training of the deep CNN and mitigate the vanishing gradient problem [40, 43, 44]. While solving the vanishing gradient that may often arise, the model may suffer convergence problems and computational complexity, leading to overfitting. Hence, an Attention mechanism inspired by Bahdanau *et al.* [45] is introduced in the design to help maintain the model's stability and ease gradient flow, sharpening its focus on relevant parts of the input sequence. Furthermore, the Attention [45] layers dynamically adjust the attention weights based on the input data, allowing the model to adapt to the specific characteristics of the dataset. This adaptability helps the model capture relevant patterns and dependencies, even in datasets with varying structures or distributions. Ultimately, this combination enhances the model's performance [46].

The conventional CNN is expressed as:

$$z_i = f\left(\sum_{j=0}^{m-1} \omega_j x_{i+j} + b\right), \quad (4)$$

where z_i denotes the output at position i in the feature map, f represents the activation function (such as Rectifier Linear Unit (ReLU) or Sigmoid), ω_j are the weights of the filter applied to the input x, x_{i+j} are the input values at position $i + j$, b is the bias, and m is the filter size.

The second architecture in the hybrid model is an enhanced MLP, known as TSM. The concept of TSM is derived from Neural Decision Trees (NDT), which Balestriero [47] inspired. NDT employs traditional Decision Tree (DT) techniques but integrates them within a neural network framework. Traditional DTs are non-parametric supervised learners that split data into subsets based on feature values at each node, forming a tree-like structure where each path from root to leaf represents a classification rule. In a similar approach, the MLP in Eq. (5) consists of layers of neurons, and activation functions are created and interlaced as a hierarchical structure analogous to decision nodes and leaf nodes in a DT, as illustrated in Fig. 1, thus forming the TSM.

The TSM takes input X and splits it over n dense layers acting as decision nodes. These decision nodes' outputs are further processed by additional MLP sub-networks, with the endpoints functioning as the decision paths similar to leaf nodes in a DT. Each leaf node processes information from its respective decision node pathway, refining the non-linear transformation process described in Eq. (5). Additional layers, such as dropout layers, are added to regularize the training and prevent overfitting. The outputs from the leaf nodes are combined to make the final prediction. Unlike the explicitly tree-like decision nodes with soft splits used in NDT, the TSM relies on standard dense layers to effectively capture complex patterns and relationships in the data. For further details on NDT, readers can refer to [47].

$$ReLU(wx + b) \quad (5)$$

The existing MLP model is expressed as:

$$y_j^k = \sigma\left(\sum_i w_i^{kj} + b_j^k\right), \quad (6)$$

where $\sigma(x) = \frac{1}{1+\exp(x)}$, y_j^k is the output of node j in an ith layer, w_i^{kj} is the weight of the connection from node i on layer k-1 to node j on layer k, x_i^{k-1} is the output of node i on layer k-1 and b_j^k is the "bias" of cell j on layer k. The activation function is a sigmoid as defined by $\sigma(x)$.

Then, the proposed model is mathematically expressed based on Eq. (4) and Eq. (5), defining the CRA part as:

$$f\left(\sum_{j=0}^{m-1} \omega_j x_{i+j} + b\right) + Res(x) + Att(x), \quad (7)$$

where $Res(x)$ and $Att(x)$ are the residual block and Attention mechanism, respectively, and the TSM part is expressed as:

$$y_j^k = \sigma\left(\sum_{\substack{i \\ \ell \in \mathcal{L}}} w_i^{kj} + b_j^k\right), \quad (8)$$

where ℓ represents individual prediction nodes, and $\mathcal{L}$ denotes the set of all prediction nodes.

The concatenation process of the hybrid proposed model is expressed as follows:

$$Y_i^k = y_j^k \oplus Z_i^{k,l} \quad (9)$$

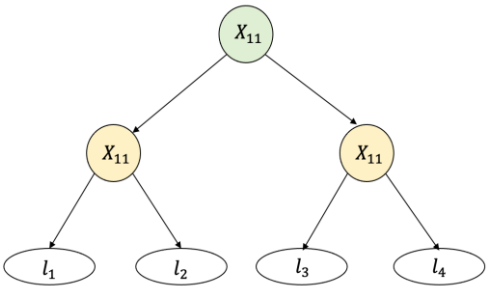

Fig. 1. A simple DT.

Method 2: To accelerate the training process of the proposed hybrid model, ensuring its better generalization and faster convergence, the ps_CALR was introduced. The ps_CALR was proposed in [39] as an LR optimization method from the *cosine annealing* technique to solve hyperparameter tuning of LR and model convergence to the global minimum. The ps_CALR is expressed in Eqs. (9) – (11):

$$P_s = \frac{5\pi}{4} \tag{10}$$

$$C_a = cos\left(\frac{P_s(mod(t-1,\lfloor T/M \rfloor)-E_w)}{\lfloor T/M-E_w \rfloor} - offset\right) \tag{11}$$

$$\alpha(t)_{CALR_{PS,EW}} = \begin{cases} \alpha_0 \cdot \frac{t+1}{E_w}, & E_w < t, set\ initial\ E_w \\ \frac{\alpha_0}{2}(C_a+1), & E_w \geq t \end{cases}, \tag{12}$$

where t is the iteration number, T is the total number of training iterations, M is the number of cycles as in the original function, f is a monotonically decreasing function, E_w represents the number of warm-up epochs, an $offset$ represents the offset value, and α_0 defines the initial LR. Notably, the periodic shift P_s and $offset$ are constants that do not impede the model's hyperparameter tuning, whereas the warm-up epoch initialization is carefully chosen for the underlying task. For details of ps_CALR, see [39].

Method 3: The class distribution, shown in Fig. 1, indicated a significant imbalance in the datasets. To ensure the proposed hybrid model performed optimally, a cost-sensitive approach was adopted using a Class Imbalance Weight (CIRW) technique [48]. Consequently, the hybrid model becomes inherently cost-sensitive to minority classes during training. The CIRW technique is expressed as:

$$Class\ Weight_k = \frac{M_t}{k \times M_k}, \tag{13}$$

where k is the number of classes. M_t represents the total observations, and M_k is the samples in each class k in the dataset.

Since the focus is on the minority class, the weight penalty is obtained by penalizing the majority class by 1 and the rest of the minority classes with the magnitude, CW_k of the majority class. The penalty is expressed as:

$$CIRW_k = \frac{CW_k}{CW_{majority_k}}, \tag{14}$$

where $CW_{majority_k}$ represents the CW of the majority class

Consequently, the penalty for a cost-sensitive metric is of the form:

$$\left[\frac{CW_j}{CW_{majority_k}}, \frac{CW_{j+1}}{CW_{majority_k}}, \ldots, \frac{CW_k}{CW_{majority_k}}\right], \tag{15}$$

where $j = 1, \ldots, k$.

The architectural framework of the proposed model is depicted in Fig. 2, showcasing the proposed hybrid CRATSM model sandwiched with ps_CALR for optimizing LR, as well as CIRW for the data balancing method. By leveraging the strengths of both models, this integrated approach provides a robust framework for accurately capturing complex feature interactions and focusing on the most relevant parts of the input data.

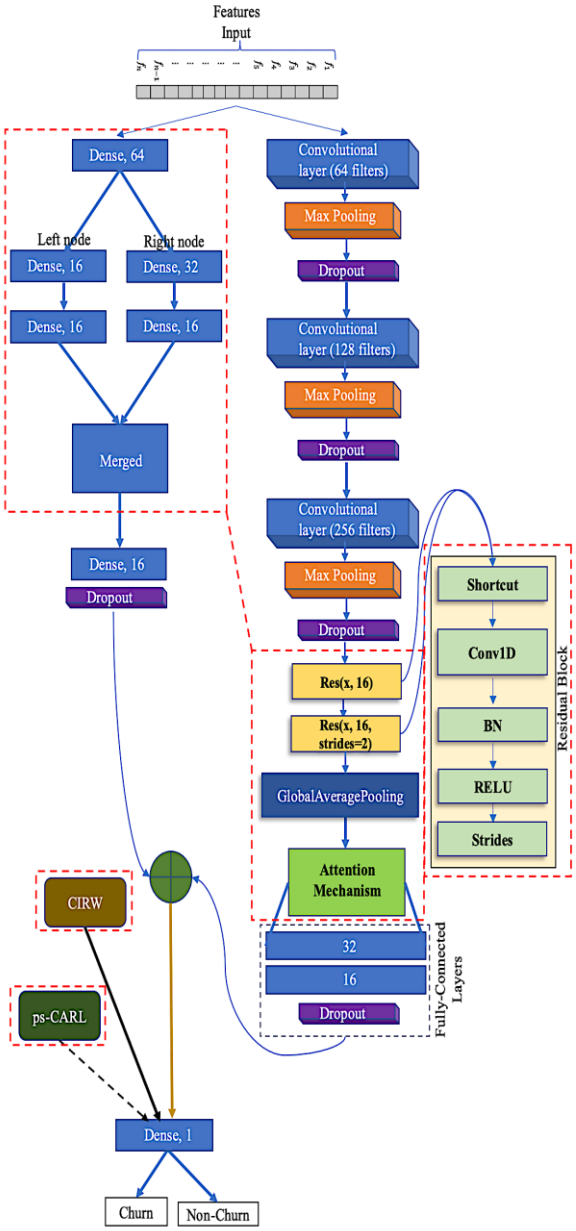

Fig. 2. Proposed CRATSM model architecture.

D. Performance Measures

The CRP is primarily a data-imbalanced problem. In this regard, performance metrics such as F1-score,

geometric mean (Gmean), and Area Under the Curve (AUC) are considered more suitable than accuracy for the proposed model evaluation. The summary of the metrics formulae is presented in Table II.

Table II. Performance Metrics.

Metric	Formula
Accuracy (Acc)	$\frac{(TP+TN)}{N}$
Recall (Sensitivity)	$\frac{TP}{TP+FN}$
Specificity	$\frac{TN}{TN+FP}$
Precision or PPV	$\frac{TP}{(TP+FP)}$
F1-score	$2 \times \frac{Precision.Recall}{(Precision+Recall)}$
Gmean	$\sqrt{Recall * Specificity}$

Where N represents the total observations, True Positive (TP) is the number of cases accurately identified as positive, False Positive (FP) cases are those categorized as negative but are positive, and True Negative (TN) is the number of instances accurately identified as negative. False Negative (FN) is the number of situations mistakenly labeled as positive when they are negative. Positive Predictive Value (PPV).

Precision-Recall (PR) Curve: The PR-Curve illustrates the trade-off between Precision and Recall for various thresholds in a binary classification task, which is particularly useful for handling imbalanced problems. Each point in the curve represents a different threshold for classifying instances as positive, demonstrating how precision and recall vary with these thresholds, as depicted in Fig. 3. Achieving high recall is straightforward by predicting most instances as the majority class. It indicates that the model prioritizes identifying positive instances (the minority) over negatives.

E. Experimental Setup

The proposed hybrid model incorporates two primary DNN models: MLP and CNN. Initial experiments were conducted to evaluate the effectiveness of CIRW and ps_CALR on these base models. The performance of CIRW was compared against six other oversampling techniques: Adaptive Synthetic-ADASYN (AD), BorderlineSMOTE (BS), GeometricSMOTE (GS), KMeansSMOTE (KMS), SMOTE (SM), and SVMSMOTE (SVMS). Following this, the performance of ps_CALR was assessed in comparison to the existing CALR. Finally, the hybrid model was integrated with CIRW and ps_CALR, and its performance was evaluated.

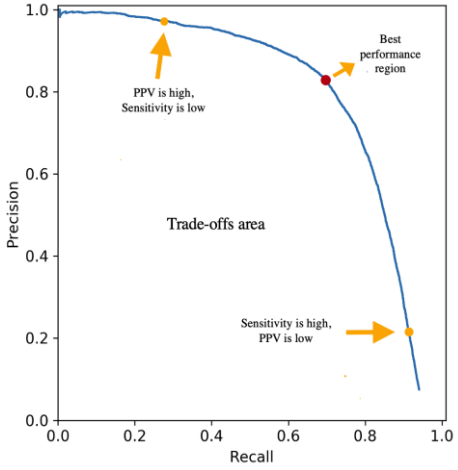

Fig. 3. PR curve for model performance.

F. Model training and Hyperparameters

The hyperparameters of both the base and proposed models are detailed in Tables III-V, respectively. The models were trained with batch sizes of 32 and 30 epochs. Adam optimizer with LR set at 0.01 was employed in the initial base models training, and the same value was used as the initial large LR for the ps_CALR experiment. Furthermore, the respective datasets were split into three parts, namely train, validation, and test sets. A split train_test_split function was used in the ratio 80:20 first to split the respective datasets into train and test sets.

Table III. Hyperparameter setting for MLP Base Model.

MLP Model	Shape/Units	AF
Input shape	(39,) – IBM (27,) – Iranian (66,) – Orange	
Dense Layer 1	64	Relu
Dense Layer 2	32	Relu
Dense Layer 2	16	Relu
Output Layer	1	Sigmoid

Table IV. Hyperparameter setting for CNN Base Model.

CNN Model	CNN Block			Fully Connected Layers (FCL)		Output Layer
Parameters	Conv Layer (CL) 1	CL 2	CL 3	1st	2nd	
Input Shape	(39, 1) – IBM (27, 1) – Iranian (66, 1) – Orange					
Filters	64	128	256			
Kernel Size	3	6	6			
Padding	Same	Same	Same			
Max Pooling	✓	✓	✓			
BatchNormalization	✓	✓	✓			
Dropout	0.25	0.25	0.25		0.4	
Flatten	✓					
AF	Relu	Relu	Relu	Relu	Relu	Sigmoid
Units				32	16	1

Table V. Hyperparameter setting for Proposed CRATSM model.

Proposed Model	**CNN+Residual+Attention (CRA)**			**TSM**		
Parameters	**CL 1**	**CL 2**	**CL 3**	**Dense_1**	**Dense_2**	**Dense_3**
Input Shape	(39, 1) – IBM (27, 1) – Iranian (66, 1) – Orange			(39,) – IBM (27,) – Iranian (66,) – Orange	Dense_1	Dense_1
Filters	64	128	256	**Leaf Nodes**		
Kernel Size	3	6	6	Leaf_1	16	Dense_3
Padding	Same	Same	Same	Leaf_2	16	
Max Pooling	✓	✓	✓	AF	Relu	
Dropout	0.25	0.25	0.25	Concatenate	Leaf_1	Leaf_2
AF	Relu			Merged	16	Merged
Residual Block with ConvID				Output	1	Merged
Res. Parameters	1st	2nd				
Filters	16	16				
Kernel Size	3	3				
Padding	Same	Same				
BatchNormalization	✓	✓				
Stride		2				
AF	Relu					
GlobalAveragingPooling	✓					
Attention Block	✓					
FCL						
	1st	2nd				
Units	32	16				
AF	Relu					
Dropout		0.4				
Output						
Concatenate	CNN+Res+Att_Block			TSM_Block		
FCL	64, Relu					
Model Output	1, Sigmoid					

IV. RESULTS AND DISCUSSION

The results of the various experiments are presented and discussed in this section.

A. Exploratory Data Analysis (EDA)

The churn rate distribution for each dataset is illustrated in Fig. 4. In the case of the IBM dataset, subfigure (a) reveals that the ratio of non-churners to churners is approximately 3:1. Similarly, the churn rate distributions in the Iranian and Orange datasets, depicted in subfigures (b) and (c), show an imbalance ratio of roughly 6:1. Additionally, it was observed that the Orange dataset exhibits the highest class imbalance among the datasets. This insight indicates that the proportion of customers leaving is significantly skewed towards those staying, resulting in an imbalanced data problem.

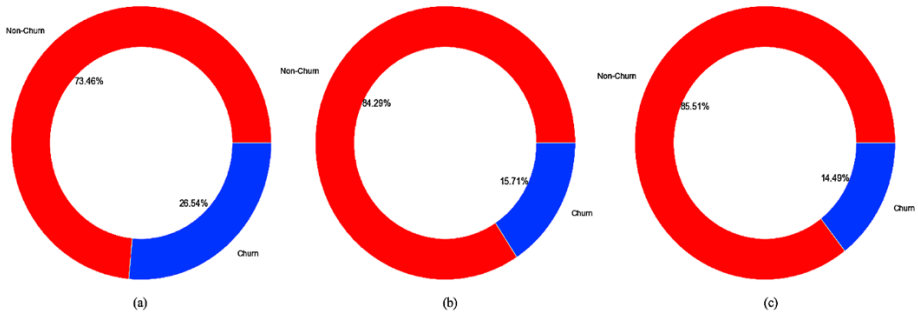

Fig. 4. Class Distribution (a) IBM, (b) Iranian, and (c) Orange Datasets.

B. CIRW with Base Models

The results of CIRW are, as shown in Fig. 5. In the context of IBM, the result indicates that SVMS achieves the highest F1-score of 58.86% with MLP, outperforming other oversampling methods, closely followed by AD, as shown in Fig. 5(a). Similarly, KMS leads among oversampling methods with a 59.39% F1-score, followed by GS with 58.59% using the CNN base model in Fig. 5(b). Comparing oversampling methods with the proposed CIRW, CIRW exhibits significant enhancements with both MLP and CNN base models, indicating its competitive edge over existing methods in Figs. 5(a) and (b).

Regarding the Iranian dataset, SM achieves an impressive 88.26% F1-score with the MLP base model, depicted in Fig. 5(c), while SM emerges as the top performer with an 82.61% F1-score with the CNN base model in Fig. 5(d). Incorporating CIRW further improves performance, achieving F1-scores of 90.49% and 83.30% for MLP and CNN, respectively, surpassing oversampling methods, as shown in Figs. 5 (c) and (d).

Training the MLP model with the Orange dataset, SM demonstrates a noteworthy F1-score of 54.24%, followed by SVMS, as shown in Fig. 5(e). AD leads with a 59.76% F1-score for the CNN model, with SM at 53.33% in Fig. 5(f). However, CIRW significantly outperforms oversampling methods across both MLP and CNN models, as depicted in Figs. 5(e) and (f).

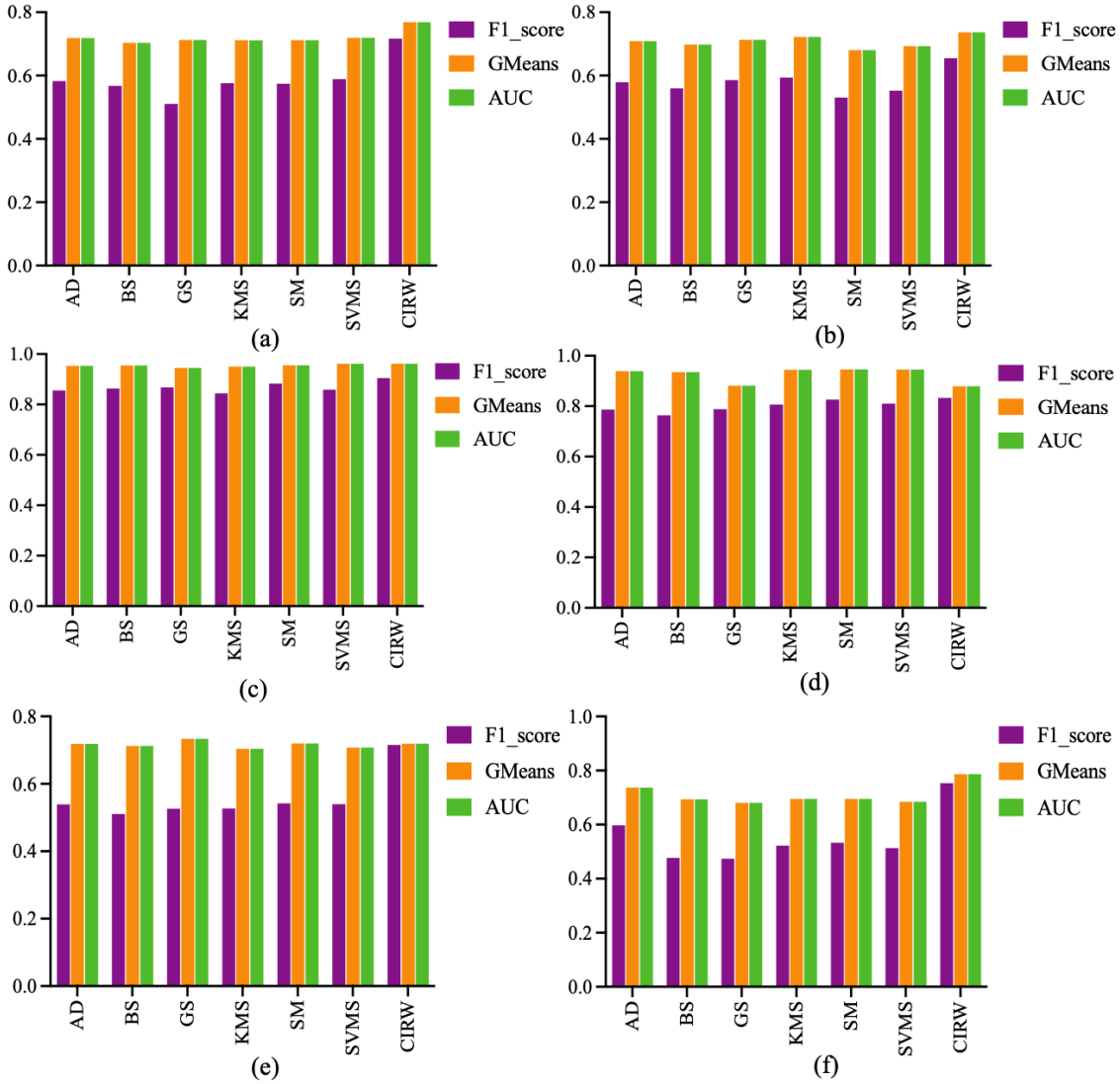

Fig. 5. Base models empirical results for data balancing methods: (a) MLP with CIRW on IBM (b) CNN with CIRW on IBM (c) MLP with CIRW on Iranian (d) CNN with CIRW on Iranian (e) MLP with CIRW on Orange (f) CNN with CIRW on Orange.

C. ps_CALR with Base Models

Combining CALR perturbation with MLP and data sampling methods in the IBM dataset context reveals that SVMS achieves the highest F1-score of 59.83%. However, employing CIRW enhances performance, achieving an F1-score of 71.70%. Similarly, using ps-CALR shows KMS with the best F1-score of 62.53%, but MLP+CIRW+ps_CALR leads to an improved F1-score of 72.14% to outperform MLP+CIRW+CALR and MLP+CIRW, as shown in Fig. 6(a).

Experimenting with a different model, CNN further validates ps-CALR+CIRW, achieving better performance improvement than CNN+CALR+CIRW and CNN+CIRW in Fig. 6(b). Also, in the Iranian dataset, employing CALR with GS yields the highest F1-score of 88.46%, while CIRW boosts performance to 90.70%. With MLP+ps-CALR, SVMS achieves the best F1-score of 89.72%, but combining MLP+ps-CALR with CIRW further enhances performance to 90.70%, as shown in Fig. 6(c).

Similarly, the performance improvement was observed with CNN+ps_CALR+CIRW, which obtained an F1-score of 89.93% in the CNN model category in Fig. 6(d). Furthermore, in the context of the Orange dataset, SVMS achieves a 58.54% F1-score with CALR, whereas CIRW improves it to 75.36% in Fig. 6(e). Similarly, BS achieves a 61.99% F1 score, but ps_CALR+CIRW achieves a superior F1 score of 77.53%. A similar performance trend was observed with a CNN model with CNN+ps-CALR+CIRW, reaching 80.41% F1-score to surpass CNN+CALR+CIRW and CNN+CIRW, as shown in Fig. 6(f).

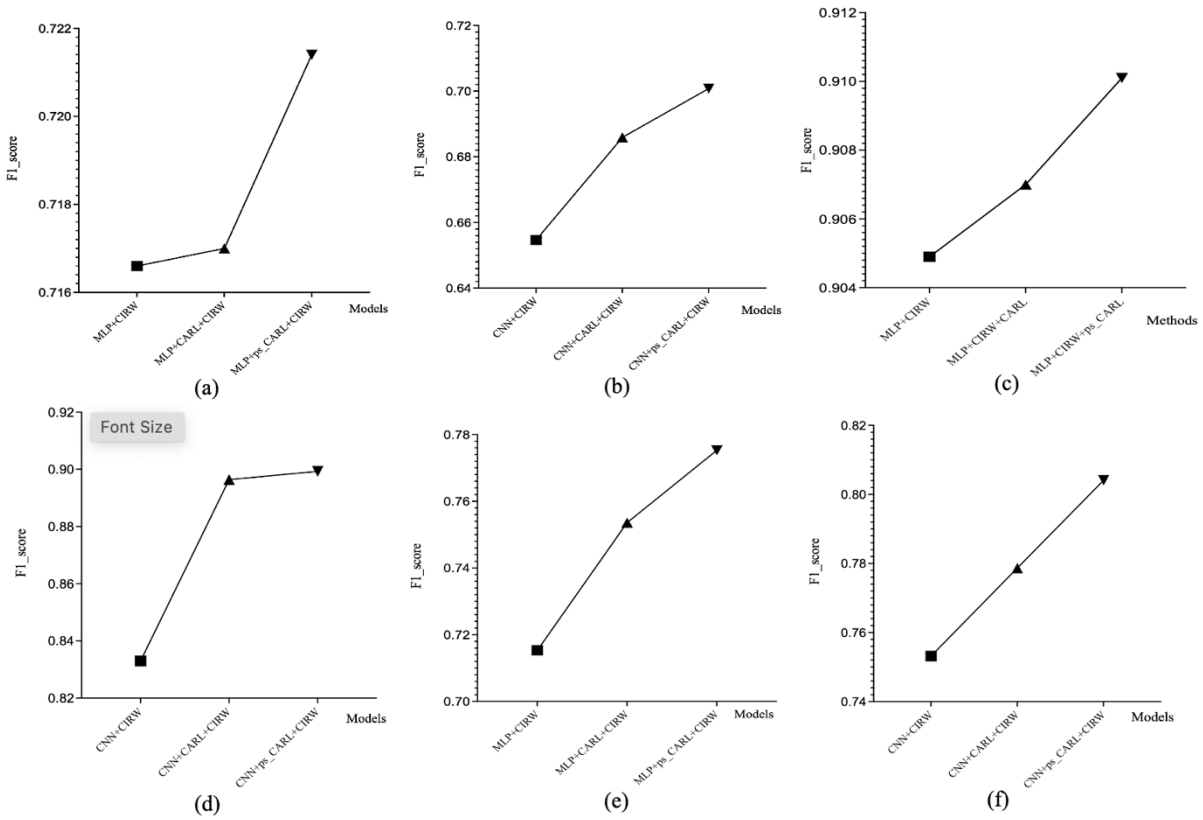

Fig. 6. Experimenting ps_CALR optimization with base models: (a) MLP with CIRW and ps_CALR on IBM (b) CNN with CIRW and ps_CALR on IBM (c) MLP with CIRW and ps_CALR on Iranian (d) CNN with CIRW and ps_CALR on Iranian (e) MLP with CIRW and ps_CALR on Orange (f) CNN with CIRW and ps_CALR on Orange.

D. Proposed Hybrid Model

The findings from the proposed hybrid model across the selected datasets are summarized in Table VI. Analysis of the confusion matrix reveals instances of misclassification, with 132, 5, and 52 cases falsely predicted as churn in the IBM, Iranian, and Orange datasets, respectively. Conversely, false positives of 152, 24, and 16 instances are noted in the IBM, Iranian, and Orange datasets, respectively, as shown in Fig. 7. This indicates the commendable performance of the proposed CRATSM model, aligning well with cost-sensitivity considerations within the business context. Effective customer retention hinges on the model's ability to identify potential churners, minimizing false negatives to prevent long-term financial losses.

Table VI. Performance of the proposed model.

Datasets	**Accuracy**	**F1-score**
IBM	0.7984	0.7458
Iranian	0.9540	0.9193
Orange	0.8980	0.8880

In addition, the PR curve in Fig. 8 highlights the optimal performance of the model with the Iranian data and a notable trade-off between precision and recall for the IBM and Orange datasets. It underscores the model's effectiveness in identifying positive instances, particularly customers likely to switch service providers.

However, performance discrepancy is evident, with IBM exhibiting the lowest performance. This may be attributed to its complex feature patterns, class imbalance distribution, sparse characteristics, and class separability issues that impact model generalization. Conversely, the model effectively captures patterns in the Iranian and Orange datasets, showcasing adaptability to varying complexities.

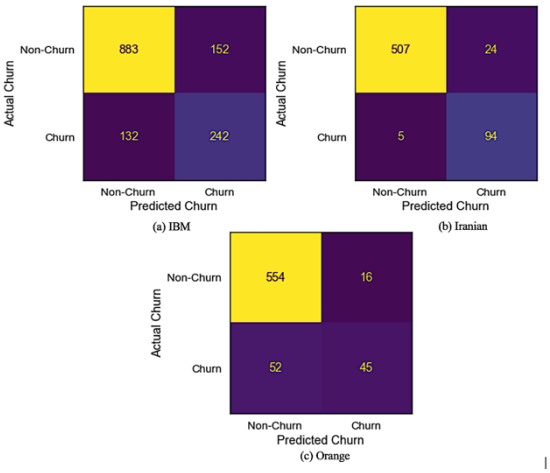

Fig. 7. Confusion matrix of the CRATSM model for the three datasets.

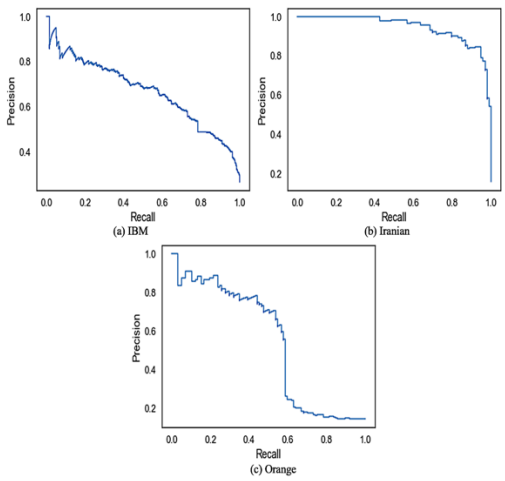

Fig. 8. PR-Curve of the CRATSM model.

E. Comparative Analysis of the Proposed Model

The performance of the proposed hybrid CRATSM model was assessed and compared with nine conventional ML algorithms, including RF, DT, NB, and (K-Nearest Neighbor) KNN. Other include LGB, ADA Boosting (ADAB), Gradient Boosting (GB), SVM, and XGB, which are designed to handle unbalanced data. The ML techniques not agnostic toward imbalanced data treatment were paired with data sampling techniques from previous experiments.

Analysis of the IBM indicated that RF achieved the highest F1-score of 72.30% among non-cost-sensitive ML techniques, followed by NB at 70.55%, KNN at 66.99%, and DT at 66.73%. Among the cost-sensitive MLs, GB performed the best, scoring 73.96%. However, the proposed model surpassed all ML methods with an F1-score of 74.58%, as shown in Fig. 9(a).

For the Iranian dataset, RF attained the highest F1-score of 91.39% among non-cost-sensitive ML methods, followed by KNN at 87.46%, DT at 85.14%, and NB at 75%. LGB achieved the highest performance among cost-sensitive MLs, scoring 90.28%. The proposed model outperformed all ML methods, achieving an F1-score of 91.93% in Fig. 9(b).

On the Orange dataset, RF obtained the highest score among non-cost-sensitive ML methods at 84.09%, followed by DT at 76.95%, KNN at 54.55%, and NB at 52.12%. GB achieved the highest performance among cost-sensitive MLs at 87.06%. The proposed model demonstrated superior performance with an F1-score of 88.80%, surpassing traditional ML methods, as depicted in Fig. 9(c).

Overall, the proposed CRATSM model exhibited superior performance across all datasets, outperforming conventional ML methods in the F1-score.

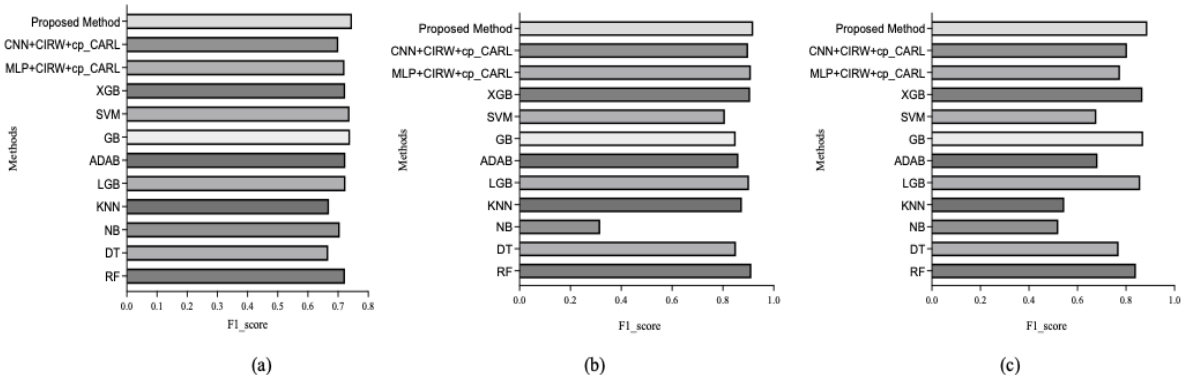

Fig. 9. Comparative analysis of the proposed model on (a) IBM, (b) Iranian, and (c) Orange.

In addition, the proposed model is compared with existing works in the literature, as presented in Table VII. The finding shows that the proposed model competes favorably with existing models in the literature, with notable improvement in CRP model performance.

Table VII. Comparative analysis with past studies.

ML Models	Models	Dataset	F1-score	Accuracy
[16]	ANN	IBM	-	0.7648
[18]	IDCNN-GBDT	Orange	0.6122	0.8003
[49]	ANN	Iranian	0.8810	-
[50]	CNN+Embedding	Orange	0.8937	0.8116
This Paper	CRASTM	IBM Iranian Orange	**0.7445** **0.9193** **0.8880**	**0.7984** **0.9540** **0.8980**

V. CONCLUSION

This paper has discussed the practical approach to DL model design for CRP. The approach demonstrated the hybrid of two DNN models: MLP and CNN. A TSM was formed from the base MLP model, while CRA improved the ID-CNN model. In addition, the hybrid CRATSM model was sandwiched with p_CALR to optimize its LR, ensuring better generalization and faster convergence while exploring CIRW to solve the class imbalance in the datasets. Three Telecom datasets were used to investigate the performance of the hybrid model. Empirical findings show the model competes favorably with conventional ML techniques and other CRP models in existing studies.

Even as customer churn remains a pressing concern for telecom companies, efforts will continue to be geared towards an effective CRP to provide a reliable decision-making process.

For future works, successful innovation based on Transformer [46, 51] is suggested to explore churn prediction. Also, Polyak [52] as promising LR optimization techniques could also be investigated for improving model convergence and generalization

ACKNOWLEDGEMENT

This work is funded by Ministry of Higher Education Malaysia, Fundamental Research Grant Scheme [Grant Number: FRGS/1/2022/STG06/USM/02/4], for the Project entitled “Efficient Joint Process Monitoring using a New Robust Variable Sample Size and Sampling Interval Run Sum Scheme”.

REFERENCES

[1] Global Monitor, *Malaysia Telecommunication Market Report*, pp.1–4. 2021.

[2] Statista, *Number of internet and social media users worldwide as of July 2023*, https://www.statista.com/statistics/ [Accessed 2023/10/20]

[3] N. Y. Nhu, T. Van Ly and D. V. Truong Son, “Churn Prediction in Telecommunication Industry using Kernel Support Vector Machines,” *PLoS One*, vol. 17, pp. 0267935, 2022.

[4] M. Zhu and J. Liu, “Telecom Customer Churn Prediction Based on Classification Algorithm,” in *2021 Int. Conf. Aviation Safety and Inform. Technol.*, pp. 268–273, 2021.

[5] N. Alboukaey, A. Joukhadar and N. Ghneim, “Dynamic Behavior Based Churn Prediction in Mobile Telecom,” *Expert Sys. Appl.*, vol. 162, pp. 113779, 2020.

[6] N. I. Mohammad, S. A. Ismail, M. N. Kama, O. M. Yusop and A. Azmi, “Customer Churn Prediction in Telecommunication Industry Using Machine Learning Classifiers,” in *Proc. 3rd Int. Conf. Vision, Image and Sign. Process.*, no. 34, pp. 1-7, 2019.

[7] S. Saleh and S. Saha, “Customer Retention and Churn Prediction in The Telecommunication Industry: A Case Study on A Danish University,” *SN Appl. Sci.*, vol. 5, pp. 173. 2023.

[8] B. Do Chung, J. H. Park, Y. J. Koh and S. Lee, “User Satisfaction and Retention of Mobile Telecommunications Services in Korea,” *Int. J. Hum. Comput. Interact.*, vol. 32, pp. 532–543, 2016.

[9] L. Y. Chen, Y. Chen, Y. D. Kwon, Y. Kang and P. Hui, “IAN: Interpretable Attention Network for Churn Prediction in LBSNs,” in *Proc. of the 2021 IEEE/ACM on Adv. in Soc. Netw. Analy. and Mining*, pp. 23–30, 2021.

[10] M. Al-Mashraie, S. H. Chung and H. W. Jeon, “Customer Switching Behavior Analysis in The Telecommunication Industry via Push-Pull-Mooring Framework: A Machine Learning Approach,” *Comput. Ind. Eng.*, vol. 144, pp. 106476, 2020.

[11] A. Amin, B. Shah, A. Khattak, F. Moreira, G. Ali, Á. Rocha and S. Anwar, “Cross-company Customer Churn Prediction in Telecommunication: A Comparison of Data Transformation Methods,” *Int. J. Inf. Manage.*, vol. 46, pp. 304–319, 2019.

[12] S. Maldonado, J. López and C. Vairetti, “Profit-based Churn Prediction Based on Minimax Probability Machines,” *Eur. J. Oper. Res.*, vol. 284, pp. 273–284, 2020.

[13] E. King and J. Rice, “Analysis of Churn in Mobile Telecommunications: Predicting the Timing of Customer Churn,” *AIMS Int. J. of Mgt.*, vol. 13, pp. 127-141, 2019.

[14] A. Amin, F. Al-Obeidat, B. Shah, A. A. J. Loo and S. Anwar, “Customer Churn Prediction in Telecommunication Industry using Data Certainty,” *J. Bus. Res.*, vol. 94, pp. 290-301, 2019.

[15] S. M. Kostić, M. I. Simić and M. V. Kostić, “Social Network Analysis and Churn Prediction in Telecommunications Using Graph Theory,” *Entropy*, vol. 22, pp. 753, 2020.

[16] N. Gamulin, M. Štular and S. Tomažič, “Impact of Social Network to Churn in Mobile Network,” *Automatika*, vol. 56, pp. 252–261, 2015.

[17] S. A. Amatare and A. K. Ojo, “Predicting Customer Churn in Telecommunication Industry Using Convolutional Neural Network Model,” *IOSR J. Comput. Eng.*, vol. 22, pp. 54–59, 2020.

[18] S. Li, G. Xia and X. Zhang, “Customer Churn Combination Prediction Model Based on Convolutional Neural Network and Gradient Boosting Decision Tree,” in *Proc. 2022 5th Int. Conf. Algorit., Comput. and Artif. Intellig.*, no. 12, pp. 1-6, 2022.

[19] V. Umayaparvathi and K. Iyakutti, “Automated Feature Selection and Churn Prediction using Deep Learning Models,” *Int. Res. J. Eng. and Technol.*, vol. 4, pp. 1846–1854, 2017.

[20] S. Kumar and M. Kumar, “Predicting Customer Churn using Artificial Neural Network,” *Commun. in Comput. and Inform. Sci.*, vol. 1000, pp. 299-306, Springer, Cham, 2019.

[21] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk and F. Herrera, *Learning from Imbalanced Data Sets*, Springer, 2018.

[22] R. Shwartz-Ziv and A. Armon, “Tabular data: Deep learning is not all you need,” *Information Fusion*, vol. 81, pp. 84–90, 2022.

[23] L. Grinsztajn, E. Oyallon and G. Varoquaux, “Why Do Tree-Based Models Still Outperform Deep Learning on Typical Tabular Data?” *Adv. Neural Inf. Process Sys.*, vol. 35, pp. 1-14, 2022.

[24] Y. Gorishniy, I. Rubachev, V. Khrulkov and A. Babenko, “Revisiting Deep Learning Models for Tabular Data,” *Adv. Neural Inf. Process Sys.*, vol. 23, pp. 18932–18943, 2021.

[25] S. A. Fayaz, M. Zaman, S. Kaul and M. A. Butt, “Is Deep Learning on Tabular Data Enough? An Assessment,” *Int. J. Adv. Comput. Sci. and Appl.*, vol. 13, pp. 466–473, 2022.

[26] B. N. R. Chagas, J. Viana, O. Reinhold, F. M. F. Lobato, A. F. L. Jacob, R. Alt, “A Literature Review of The Current Applications of Machine Learning ond Their Practical Implications,” *Web Intellig.*, vol. 18, pp. 69–83, 2020.

[27] J. Xiao, L. Huang and L. Xie, “Cost-Sensitive Semi-Supervised Ensemble Model for Customer Churn Prediction,” in *2018 15th Int. Conf. Serv. Sys. and Serv. Manage.*, Hangzhou, China, pp. 1-6, 2018.

[28] M. Panjasuchat and Y. Limpiyakorn, “Applying Reinforcement Learning for Customer Churn Prediction,” *J. Phys. Conf. Ser.*, vol. 1619, no. 012016, 2020.

[29] A. Amin, S. Anwar, A. Adnan, M. Nawaz, N. Howard, J. Qadir, A. Hawalah and A. Hussain, “Comparing Oversampling Techniques to Handle the Class Imbalance Problem: A Customer Churn Prediction Case Study,” *IEEE Access*, vol. 4, pp. 7940–7957, 2016.

[30] X. Wang and S. Mu, “E-commerce Customer Churn Prediction Based on Improved SMOTE and AdaBoost,” in *13th Int. Conf. on Serv. Sys. and Serv. Manage.*, Kunming, pp. 1–5, 2016.

[31] N. M. Mqadi, N. Naicker and T. Adeliyi, “Solving Misclassification of The Credit Card Imbalance Problem using Near Miss,” *Math. Probl. Eng.*, vol. 2021, pp. 1–16, 2021.

[32] M. Fathian, Y. Hoseinpoor and B. Minaei-Bidgoli, “Offering A Hybrid Approach of Data Mining to Predict The Customer Churn Based on Bagging and Boosting Methods,” *Kybernetes*, vol. 45, pp. 732–743, 2016.

[33] K. Kim, C. H. Jun and J. Lee, “Improved Churn Prediction in Telecommunication Industry by Analyzing A Large Network,” *Expert Sys. Appl.*, vol. 41, pp. 6575–6584, 2014.

[34] A. Sniegula, A. Poniszewska-Maranda and M. Popovic, "Study of Machine Learning Methods for Customer Churn Prediction in Telecommunication Company," in *Proc. 21st Int. Conf. Inform. Integr. and Web-based Appl. & Serv.*, pp. 640-644, 2019.

[35] A. Chouiekh and E. H. I. El Haj, "Deep Convolutional Neural Networks for Customer Churn Prediction Analysis," *Int. J. Cognit. Inform. and Nat. Intellig.*, vol. 14, pp. 1–16, 2020.

[36] E. Domingos, B. Ojeme and O. Daramola, "Experimental Analysis of Hyperparameters for Deep Learning-Based Churn Prediction in The Banking Sector," *Computation*, vol. 9, no. 34, pp. 1-19, 2021.

[37] Y. L. Tan, Y. H. Pang, S. Y. Ooi, W. H. Khoh and F. S. Hiew, "Stacking Ensemble Approach for Churn Prediction: Integrating CNN and Machine Learning Models with CatBoost Meta-Learner," *J. Eng. Technol. and Appl. Phys.*, vol. 5, pp. 99–107, 2023.

[38] J. Jepkoech, D. M. Mugo, B. K. Kenduiywo and E. C. Too, "The Effect of Adaptive Learning Rate on the Accuracy of Neural Networks," *Int. J. Adv. Comput. Sci. and Appl.*, vol. 12, pp. 736–751, 2021.

[39] O. V. Johnson, C. Xinying, K. W. Khaw and M. H. Lee, "ps-CALR: Periodic-Shift Cosine Annealing Learning Rate for Deep Neural Networks," *IEEE Access*, vol. 11, pp. 139171–139186, 2023.

[40] D. Chen, F. Hu, G. Nian and T. Yang, "Deep Residual Learning for Nonlinear Regression," *Entropy*, vol. 22(2), no. 193, pp. 1-14, 2020.

[41] M. L. Yadav and B. Roychoudhury, "Handling Missing Values: A Study of Popular Imputation Packages in R," *Knowl. Based Sys.*, vol. 160, pp. 104–118, 2018.

[42] D. Singh and B. Singh, "Investigating the Impact of Data Normalization on Classification Performance," *Appl. Soft Comput.*, vol. 97, pp. 105524, 2020.

[43] Z. Allen-Zhu and Y. Li, “What Can Resnet Learn Efficiently, Going Beyond Kernels?” *Adv. in Neur. Inform. Process. Sys.* 32, 2019.

[44] F. He, T. Liu and D. Tao, “Why Resnet Works? Residuals Generalize,” *IEEE Trans. Neur. Netw. and Learn. Sys.*, vol. 31, no. 12, pp. 5349-5362, 2020.

[45] D. Bahdanau, K. Cho and Y. Bengio, “Neural Machine Translation by Jointly Learning to Align and Translate,” in *Int. Conf. Learn. Repres., 2015*.

[46] Z. Niu, G. Zhong and H. Yu, "A Review on The Attention Mechanism of Deep Learning," *Neurocomputing*, vol. 452, pp. 48–62, 2021.

[47] R. Balestriero, “Neural Decision Trees,” *arXiv preprint*, arXiv:1702.07360, 2017.

[48] O. V. Johnson, C. XinYing, K. W. Khaw and M. H. Lee, "A Cost-Based Dual ConvNet-Attention Transfer Learning Model for ECG Heartbeat Classification," *J. Inform. and Web Eng.*, vol. 2, pp. 90–110, 2023.

[49] A. Keramati, R. Jafari-Marandi, M. Aliannejadi, I. Ahmadian, M. Mozaffari and U. Abbasi, "Improved Churn Prediction in Telecommunication Industry using Data Mining Techniques," *Appl. Soft Comput. J.*, vol. 24, pp. 994–1012, 2014.

[50] T. W. Cenggoro, R. A. Wirastari, E. Rudianto, M. I. Mohadi, D. Ratj and B. Pardamean, "Deep Learning As A Vector Embedding Model for Customer Churn," *Procedia Comput. Sci.*, vol. 179, pp. 624–631, 2021.

[51] M. Moradi and M. Dass, "Applications of Artificial Intelligence in B2B Marketing: Challenges and Future Directions," *Indus. Market. Manage.*, vol. 107, pp. 300–314, 2022.

[52] X. Wang, M. Johansson and T. Zhang, "Generalized Polyak Step Size for First Order Optimization with Momentum," in *Proc. 40th Int. Conf. Mach. Learn.*, no. 1489, pp. 35836–35863, 2023.