International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Identifying Customer Churn in Insurance Company Information Using Novel Ensemble Technique

Thivakaran TK[1], **Neeraj Sharma**[2], **Pradeep Kumar Shah**[3], **Mohammad Shahid**[4]

Submitted: 24/04/2023 **Revised**: 19/06/2023 **Accepted**: 29/06/2023

Abstract: For insurance firms, client churn, or the absence of customers, is a major problem because it affects revenue and sales. This study suggests a unique method for recognizing client churn in an insurance firm called modified monarch butterfly optimized random forest (M2BO-RF) to tackle this issue. This approach aims to improve the precision and efficacy of customer churn prediction using both RF and modified monarch butterfly optimization (M2BO). The foraging behavior of monarch butterflies, renowned for their impressive navigational abilities and an effective quest for supplies, served as the model for the MBO algorithm. The fundamental components of the MBO technique are incorporated into the proposed MMBORF algorithm, which changes the conventional RF technique. We performed trials on a dataset from an insurance company to assess the efficacy of the MMBORF method. The experimental findings showed that, in terms of prediction precision, Recall, accuracy, and F1-score, MMBORF surpasses other algorithms. Insurance businesses may create focused retention strategies and deploy resources effectively due to the algorithm's ability to detect probable churners successfully. Ultimately, this can improve customer satisfaction, lower customer churn, and contribute to insurance businesses' success.

Keywords*:Customer churn, insurance, business, prediction, and modified monarch butterfly optimized random forest (M2BO-RF)*

1. Introduction

An analysis of the potential for a client to stop using an item or resource is known as a customer churn study [1]. It implies that clients are forced to select another business due to pressure. Before abandoning the customer's goods or work, it is to detect this issue and take preventative steps. It is a statistical technique applied to fields including figuring out current clients' accounts, looking at client exits, and calculating client evacuation. It must move quickly to interact with its clientele before a policy expires because the expense of acquiring new clients is significantly higher than the expense of maintaining current ones [2].

This research aims to use multiple ML techniques and produce the best model to predict customer churn percentage. The data set comprises macroeconomic variables and information related to client activity and ethnicity. Churn forecasting uses the information in the instance of a Finnish insurance company [3]. The approaches recommended in the research contrast with how the insurer already does churn statistics. They begin by outlining the pertinent ML ideas before continuing to study the research in customer churn forecasting. Evaluating customer attrition in the aviation sector is a challenging topic that has baffled several carriers [4]. A disproportionate amount of client turnover information and unstable data cause problems.

The presence of loud demonstrations in the info set significantly impacts the sampling quality and predictive accuracy of classification systems. However, some contemporary sampling methods and ensemble models effectively address this issue of class inequalities. The growth of internet-based technologies offers businesses an outstanding chance to interact with current or prospective clients [5].

The study aims to create and implement a novel collaborative method for identifying client turnover in the insurance business. The project attempts to increase the efficacy and accuracy of churn prediction models by merging several algorithms and using the knowledge found in data from insurance companies. The objective is to give insurance businesses insightful data that will help them proactively manage customer retention and churn rates.

The remaining sections of this paper are as follows: Part 2 describes related works; Part 3 explains methodology; Part

[1]Professor, Department of Computer Science and Engineering, Presidency University, Bangalore, India, Email Id: thivakaran@presidencyuniversity.in

[2]Assistant Professor, Department of Electrical Engineeing, Vivekananda Global University, Jaipur, India, Email Id: neeraj.sharma@vgu.ac.in

[3]Assistant Professor, College of Computing Science and Information Technology, Teerthanker Mahaveer University, Moradabad, Uttar Pradesh, India, Email id: pradeep.rdndj@gmail.com

[4]Associate Professor, Department of Artificial Intelligence & Machine Learning (AIML), Noida Institute of Engineering and Technology, Greater Noida, Uttar Pradesh, India, Email id: mohammadshahid@niet.co.in

4 summarizes results; and Part 5 accomplishes with a conclusion.

2. Related Works

Authors [6] examined the effectiveness of uniform classification groups for predicting client churn in the finance, assurance, and electronic communications industries. Various ML methods are modeled for this intent to show how well the suggested group architecture works. K-NN, logistical regression, naive Bayes, assistance vector machines, random forests, decision trees, and preceptor multiplication procedures are among the models used in the study.

Authors [7] examined the main benefit of this research is that it demonstrated the suggested solution's higher efficiency by comparing its efficiency to that of conventional models. In particular, this study demonstrated the greater efficacy of Accelerating procedures in combination with hybrid reproducing techniques. This data can aid in researching and applying customer churn prediction because it shows the effectiveness of cutting-edge techniques.

Authors [8] examined the accuracy of their predictions of customer turnover in the wireless communications sector or other sectors; scientists can create and run trials using more cutting-edge ensemble and hybrid ML algorithms. To help individuals better recognize clientele's needs and uphold a stronger relationship with them, experiments can be conducted with intelligent optimized systems and models that are more readily available and interpretable. It might also be a future study area that integrates ML and qualitative social science techniques better to assist decision-makers and practitioners in the telecom industry.

Authors [9] examined the creation of profit-driven modeling strategies that maximize profits. To support the results of this work, further experimental proof of the extra value of uplift over predictive modeling—obtained by looking at an increased number of usage types—is required. These findings indicate that ascension modeling is a superior paradigm to predictive modeling for assisting in maintaining clientele operations and probably afterward.

Authors [10] examination was conducted using a variety of metrics, including memory, precision, reliability, and f1-score, and the results were presented by deploying the software package MATLAB. The study also found that, compared to results produced using standard methods, the suggested scheme is a reasonable and marginally better strategy for the insurance industry to assess client churn. To execute an extensive assessment of patron departure variables and identify the most important factors driving client abandonment, applying particular algorithms like different decision trees with interpretable conclusions is also advised. This problem makes the results less accurate and helps us understand the causes of client churn. Future tests might bag or pack combined a maximum of two approaches, according to the study's findings, to integrate them.

Authors [11] examined the data solution that can be used to create a decision support system for telecom operators and users that is unbreakable regarding privacy across the globe. To complete this task, more unnecessary research will be needed. One practical approach is considering pertinent churn prediction characteristics for scattered mode variable optimization that removes privacy ambiguity.

Authors [12] examined RF and AB, who performed exceptionally well in the data collection used in this analysis and earlier investigations. Additionally, CCP can be committed with a certain amount of confidence using the available data supply, and machine learning (ML) models might perform at least marginally higher than statistical algorithms in terms of productivity. However, there was little difference in the outcome parameter distribution across each algorithm's assessed variables. A more significant performance gain was observed when comparing the results on characteristics provided by the financial institution to data chosen by an algorithm.

Authors [13] examined that the conventional methods of forecasting churn heavily rely on demographic, product usage, and revenue features. However, more recent studies have successfully incorporated aspects of churn theories connected with evaluating social networks. Additionally, there needs to be more empirical research that uses the information-rich content of customer-company interactions that occur over e-mail, live chat, and other channels. Most research has been done here. Thirdly, there is room to examine how composite sampling techniques affect the accuracy of models. Much research has yet to be done on this in the literature. Finally, there is no formalized standard for the appropriate evaluation parameters when applying systems to datasets with uneven churn. This is a murky issue that needs further focus.

Authors [14] examined the research done to forecast their churn. It will assist banks in identifying the appropriate services, service quality, and process gaps so they can close them and keep clients. The ExtraTreeClassifier's achievement was deemed adequate for predicting the turnover of new clients. According to the data collection characteristics and the fine tweaking of the hyperparameters, tree-based classifiers and ensembles performed best. The effectiveness of the distribution-based classifications could have been better. The investigation is descriptive and has a small sample size and geographic scope. They advise expanding it to equivalent or comparable locations with an additional sample size to increase its reliability and generalizability. A longitudinal study can also shed light on the rising or falling trend in adolescent client retention for retail financial services.

Authors [15] examined the forecast number of clients who will leave a wireless provider; this study combined CNN and Random Forest classifiers with the Relief-F curve optimization technique. Estimating the number of departing clients is both crucial and challenging. To develop more effective retention strategies, mobile phone providers are investing more in developing precise churn forecasting systems. Relief-F, Random Forest, and CNN models were evaluated and educated in this investigation to forecast client turnover in telecommunications.

3. Methodology

3.1 IMBO Algorithm

It is possible to determine the total amount of monarchs present at field 1 and field 2 using the formulas ceil ($o * NP$) (NP_1, subgroup 1, SP1) and $NP - NP_1$ (NP_2, subgroup 2, SP2), correspondingly. Subgroups 1 and 2 are

3.1.1 The operator for Butterfly Adjustment

If q and is not greater than p for butterfly q, the l th element is given as.

$$w_{i,l}^{s+1} = w_{best,l}^{s} \tag{4}$$

where $w_{i,l}^{s+1}$ is the lth component of w_i. According to this $w_{best,l}^{s}$ the lth component of the best person w_{best}. However, if $rand$ is more than o, it can be represented as

$$w_{i,l}^{s+1} = w_{q_3,l}^{s} \tag{5}$$

where $w_{q_3,l}^{s}$ is the l th component of w_{q3}. Here,

$q_3 \in \{1,2,\dots,NP_2\}$.Since rand exceeds BAR in this instance, it may be determined with a different formula.

$$w_{i,l}^{s+1} = w_{i,l}^{s+1} + \alpha \times (cw_l - 0.5) \tag{6}$$

where dx is the butterfly i move stride.

The MBO technique has gained more and more attention from academics and technologists despite being developed a decade ago. They have proposed various methods to enhance the fundamentals of the MBO technique's search functionality. The MBO has also been used to solve a wide range of issues in real-life situations effectively. As previously reported, the MBO uses a set amount of butterflies in fields 1 and 2, but all newly created butterfly species produced by the migration operator are allowed. In this research, adapts itself and demanding techniques will be used to suggest an innovative version of the MBO algorithm. There is a thorough explanation of the SPMBO process.

referred to as SP1 and SP2, accordingly. In this case, ceil(w) rounds w to the nearest integer that is not w. Consequently, the following Equation produces w_l^{s+1}, l when$q \le o$formula is shown in Equation 1.

$$w_{j,l}^{s+1} = w_{q_1,l}^{s} \tag{1}$$

Whereas $w_{j,l}^{s+1}$ represents the lth element w_i and $w_{q_1,l}^{s}$ represents the l th element of w_{q1}. Randomly, Butterfly q_1 is selected from SP1. R is stated in a particular format in Equation (1).

$$q = rand * peri \tag{2}$$

Where micro denotes the duration of the transition window. When$q > o$, on the other hand, $w_{j,l}^{s+1}$ is provided by.

$$w_{j,l}^{s+1} = w_{q_2,l}^{s} \tag{3}$$

Where butterfly q_2 is randomly selected from SP2 and $w_{q_2,l}^{s}$ is the l th element of w_{q2} .

As stated in Section 3.1, there are alternately ceil ($o * NP$) (NP_1, subgroup 1) and $NP - NP_1$ (NP_2, subgroup 2) butterflies on field 1 and land 2. Throughout every stage of improvement manipulation, they are fixed and remain the same. Here, an adaptable approach adjusts the setting p in an evolving manner and updates as follows:

$$o = b + at \tag{7}$$

where b and a are constants provided by and s is the present iteration

$$b = \frac{O_{min}s_n - O_{max}}{s_n - 1} \tag{8}$$

$$a = \frac{O_{max} - O_{min}}{s_n - 1} \tag{9}$$

Where $O_{min}$and $O_{max}$are the lower and upper bounds of the variable o, correspondingly, and s_n is the maximum output. $O_{min}$and $O_{max}$are is in the range [0, 1].

In the fundamental MBO approach, the butterfly adjustment operation updates all butterflies if o =0, and the movement operation updates all butterflies if o =1. In our subsequent trials, $O_{min}$and $O_{max}$are gave the values 0.1 and 0.9, respectively, to expand the value of p's range beyond these two unique circumstances. Equation (7) shows that the variable o changes linearly between the smallest limit $O_{min}$and the highest limit O_{max}.

This section will conduct a more thorough analysis of the movement function. All newly developed butterfly individuals are accepted as the new butterfly individuals for the following iteration in the fundamental MBO process. The number of butterflies will deteriorate due to this iteration, and the velocity of confluence will be slowed down if the newly formed butterfly individual is worse than the previous one. More critically, the number of people will fluctuate if this occurs in the later stages of the inquiry.

In this study, the fundamental MBO method is given an aggressive approach. Only freshly created butterflies with higher fitness levels will be accepted and passed on to the following generation. The resulting species will be better, and the algorithm will evolve properly thanks to this choice strategy. Following the introduction of the demanding plan of action, the new butterfly individual is provided as

$$w_{j,new}^{s+1} = \begin{cases} w_j^{s+1}, & e\left(w_j^{s+1}\right) < e(w_j^s) \\ & w_j^s, \ else \end{cases}, \quad (10)$$

Where $e\left(w_j^{s+1}\right)$and $e(w_j^s)$ is the efficiency of butterflies w_j^{s+1}andw_j^s, accordingly, and $w_{i,l}^{s+1}$is a freshly created butterfly that will be passed on to the following generation.

Thirteen standard issues are addressed using the suggested method to assess the effectiveness of the proposed SPMBO methodology. The MBO and SPMBO algorithms are attempting to find the smallest possible value of a function because the 13 standard issues are minimum functional.

3.2 Random forest algorithm

The RF technique's fundamental concept is as listed below: First, selecting features is done on the decision tree to improve the purity of the classified information set. In this case, the integrity assessment criterion is the GINI index.

$$\Delta H = 1 - \sum_{j=1}^{r} O_j^2 - \sum_{i=1}^{l} \frac{C_i}{|C|} H_{C_i} \quad (11)$$

Where G corresponds to the GINI operation, q is the number of categories in sample C, o_j is the ratio of class j sample to all specimens, and l is the number of sections sample C is split over, i.e., the size of $C_i$data collections. Cluster split occurs when the gain value of the GINI index in the Equation reaches its highest level.

The RF comprises the several decision trees that were constructed, and a forecast is finished using a simple plurality casting system. The formula displays the last categorization decision.

$$K(W) = max \sum_{j=1}^{r} J(k_j(w) = z) \quad (12)$$

Where k_j stands for the categorizing method of the jth decision tree, $K(W)$ stands for the combined classification algorithm, and z is the objective parameter. The indicating variable is $(\cdot)$.

The cartographic decision tree is a cornerstone of the consolidated RF method, which successfully increases prediction accuracy and does not need to consider multicollinearity problems between parameters. However, the algorithm is highly complicated and computationally intensive for high-dimensional data and needs more generalizability.

4 Result

In Python and statistics, the database of a Customer Churn in an Insurance Companyprovided confidential information for this investigation. The selected information spans one year and is related to the characteristic. It includes information gathered by 72,445 different users [16]. The information gathered for this study was gathered from numerous outlets. The incorporation of the data was done in a way that minimizes redundancy and contradictions to collect accurate information.

Accuracy is important in assessing how well a categorization algorithm is performing. It displays the percentage of properly categorized occurrences out of all examined examples. To determine accuracy, we must divide the number of forecasts by the number of right predictions (true positives and true negatives). Figure 1 shows the results of accuracy for various methods. The accuracy of 88.23% obtained via logistic regression showed a strong performance. The accuracy of naive Bayes was 86.36%, which was less than that of logistic regression (LR). With an accuracy of 85.6 percent, the decision tree was marginally less accurate than both logistic regression and naive Bayes. The M2BO-RF model outperformed the others, achieving the highest accuracy (98.47%).

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions} \quad (13)$$

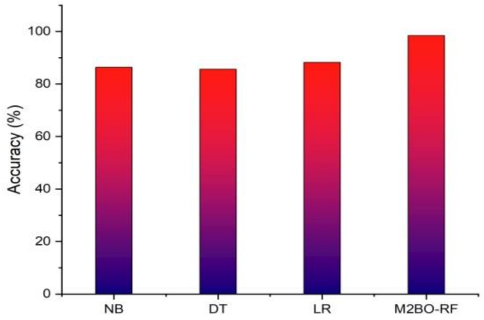

Fig 1: Result for accuracy

Precisions are the expected results or information a statistical model produces based on the parameters or observable data. According to connections or trends identified from the existing data, predictions are used to forecast or infer unknown or future values. Given a data collection with specific results or target elements, a statistical model is trained in the context of prediction. The outcome or target variable is then predicted using this framework for new or unforeseen observations. Figure 2 shows the results of precision for various methods. The M2BO-RF model achieved the highest precision at 98%, followed closely by LR (Logistic Regression) at 90%. NB (Naive Bayes) had a precision of 89.04%, and DT (Decision Tree) reached 88.47%.

$$Precision \ = \frac{True\ Positives}{(True\ Positives + False\ Positives)} \qquad (14)$$

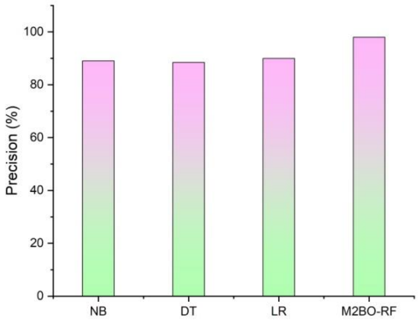

Fig 2: Result for Precision

The ratio of real positives to the total of true and false negatives is known as Recall, also known as sensitivity or true positive frequency. It evaluates how well the algorithm can recognize each good case in the information set. Figure 3 shows the results of Recall for various methods. The model for M2BO-RF achieved the highest recall rate of 96%, indicating its ability to identify positive instances effectively. LR followed closely with a recall rate of 90.45%, demonstrating good performance in correctly identifying positive samples. DT achieved a recall rate of 89.01%, slightly lower than LR but still indicating satisfactory performance. NB had a recall rate of 87.40%, showing relatively more bass performance in correctly identifying positive instances than the other models. M2BO-RF exhibited the highest recall rate, followed by LR, DT, and NB.

$$Recall \ = \frac{True\ Positives}{(True\ Positives + False\ Negatives)} \qquad (15)$$

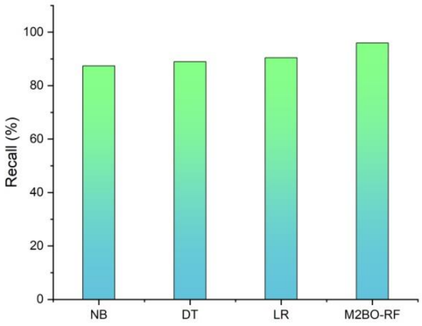

Fig 3: Result of Recall

An evaluation of an algorithm's performance in binary classification tasks involves the F1 score. It provides a fair assessment of the effectiveness of a model by combining precision and recall into a single score. The F1 score is especially helpful when the dataset is unbalanced, which happens if a particular class has a disproportionately large number of occurrences compared to another. False positives and false negatives are both taken into account, and both types of errors are given equal weight. Figure 4 shows the results of the F1-score for various methods. The M2BO-RF model achieved the highest F1 score at 95%, indicating excellent performance. DT achieved an F1-score of 87.23%, demonstrating strong accuracy. LR performed an F1-score of 85%, slightly lower than DT. NB achieved an F1-score of 85.11%, somewhat higher than LR. M2BO-RF exhibited the highest F1 score, followed by DT, NB, and LR.

$$F1\ Score \ = \frac{2 * (Precision * Recall)}{(Precision + Recall)} \qquad (16)$$

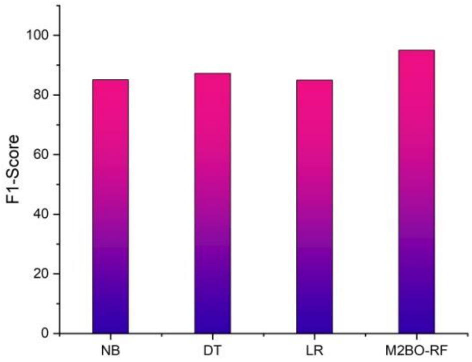

Fig 4: Result for F1-Score

5 Conclusion

To assess the efficiency of the M2BORF (Multiple Model Based on Random Forest) methods in identifying client churn. Our study tested it on a 72,445 dataset from an insurance firm. In terms of precision (98%), Recall (96%), accuracy (98.47%), and F1-score (95%), the trial results unmistakably showed that M2BORF scored better than other algorithms. Due to M2BORF's better performance, insurance companies may use this cutting-edge ensemble technique to identify customers who may churn reliably. Insurance companies can use the insights offered by M2BORF to develop specialized retention plans. Overall, the results of our study demonstrate the potential of the M2BORF technique as a useful tool for predicting customer turnover in the insurance industry, providing major advantages to both firms and their clients. The sequences and traits contained in that particular dataset may impact the results and efficacy of the M2BORF approach. To determine the algorithm's generalizability and robustness across multiple circumstances and client profiles, evaluating the technique's effectiveness on various datasets from various insurance businesses is crucial. The lack of comparisons between the M2BORF method and a broad range of churn forecasting techniques in the study makes it difficult to assess how effective it is compared to alternative strategies. A more thorough investigation of the performance and potential drawbacks of the suggested method would result from further comparison with currently used techniques.

References

[1] Çelik, O. and Osmanoglu, U.O., 2019. Comparing to techniques used in customer churn analysis. *Journal of Multidisciplinary Developments*, *4*(1), pp.30-38.

[2] He, Y., Xiong, Y. and Tsai, Y., 2020, April. Machine learning-based approaches to predict customer churn for an insurance company. In *2020 Systems and Information Engineering Design Symposium (SIEDS)* (pp. 1-6). IEEE.

[3] Stucki, O., 2019. Predicting the customer churn with machine learning methods: case: private insurance customer data.

[4] Li, Y., Wei, J., Kang, K. and Wu, Z., 2019. An efficient noise-filtered ensemble model for customer churn analysis in the aviation industry. *Journal of Intelligent & Fuzzy Systems*, *37*(2), pp.2575-2585.

[5] Sudharsan, R. and Ganesh, E.N., 2022. A Swish RNN-based customer churn prediction for the telecom industry with a novel feature selection strategy. *Connection Science*, *34*(1), pp.1855-1876.

[6] KİLİMCİ, ZH, 2022. The Effectiveness of Homogeneous Classifier Ensembles on Customer Churn Prediction in Banking, Insurance, and Telecommunication Sectors. *International Journal of Computational and Experimental Science and Engineering*, *8*(3), pp.77-85.

[7] Kimura, T., 2022. CUSTOMER CHURN PREDICTION WITH HYBRID RESAMPLING AND ENSEMBLE LEARNING. *Journal of Management Information & Decision Sciences*, *25*(1).

[8] Edwine, N., Wang, W., Song, W. and Ssebuggwawo, D., 2022. Detecting the risk of customer churn in telecom sector: a comparative study. *Mathematical Problems in Engineering*, *2022*.

[9] Devriendt, F., Berrevoets, J. and Verbeke, W., 2021. Why you should stop predicting customer churn and start using uplift models. *Information Sciences*, *548*, pp.497-515.

[10] Jajam, N. and Challa, N.P., 2023. Customer Churn Detection for insurance data using Blended Logistic Regression Decision Tree Algorithm (BLRDT). *International Journal of Intelligent Systems and Applications in Engineering*, *11*(1s), pp.72-83.

[11] Nagaraj, K., GS, S. and Sridhar, A., 2019. Encrypting and preserving sensitive attributes in customer churn data using a novel dragonfly-based pseudonymized approach. *Information*, *10*(9), p.274.

[12] Stucki, O., 2019. Predicting the customer churn with machine learning methods: case: private insurance customer data.

[13] De, S. and Prabu, P., 2022. Predicting customer churn: A systematic literature review. *Journal of Discrete Mathematical Sciences and Cryptography*, *25*(7), pp.1965-1985.

[14] Bharathi S, V., Pramod, D. and Raman, R., 2022. An ensemble model for predicting retail banking churn in the youth segment of customers. *Data*, *7*(5), p.61.

[15] Abdulsalam, S.O., Ajao, J.F., Balogun, B.F. and Arowolo, M.O., 2022. A Churn Prediction System for Telecommunication Company Using Random Forest and Convolution Neural Network Algorithms. *EAI Endorsed Transactions on Mobile Communications and Applications*, *7*(21).

[16] Spiteri, M. and Azzopardi, G., 2018, September. Customer churn prediction for a motor insurance company. In *2018 Thirteenth international conference on digital information management (ICDIM)* (pp. 173-178). IEEE.

[17] Ms. Elena Rosemaro. (2014). An Experimental Analysis Of Dependency On Automation And Management Skills. International Journal of New Practices in Management and Engineering, 3(01), 01 - 06. Retrieved from http://ijnpme.org/index.php/IJNPME/article/view/25

[18] Vijayalakshmi, V., & Sharmila, K. (2023). Secure Data Transactions based on Hash Coded Starvation Blockchain Security using Padded Ring Signature-ECC for Network of Things. International Journal on Recent and Innovation Trends in Computing and Communication, 11(1), 53–61. https://doi.org/10.17762/ijritcc.v11i1.5986