IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 24, Issue 3, Ser. II (May. –June. 2022), PP 07-12
www.iosrjournals.org

Comparison of Back propagation algorithms: Bidirectional GRU and Genetic Deep Neural Network for Churn Customer

Bhavsar Shachi[*], Ravi Gor
[*]*Research Scholar, Department of Mathematics, Gujarat University, Ahmedabad-380009*
Department of Mathematics, Gujarat University, Ahmedabad-380009

Abstract
Costumer segmentation will be beneficial for service provider to predict category of customers which might discontinue services. The customer discontinuing is also termed as customer churning. Churning of customer can be predicted by analyzing behavior and patterns from data. There are many machine learning techniques which can be used for classification such as Artificial Neural Networks, Evolutionary algorithms, etc. This study focuses on classification of discontinuation of costumers of telecom company. Here, two deep learning models Bidirectional Gated Recurrent Unit and Genetic Deep Neural Network are used for prediction. Also, results are compared with the vanilla neural network.

Keywords: *Machine Learning, Artificial Neural Networks, Deep Learning, Bidirectional GRU, Evolutionary algorithms, Churn Customer*

Date of Submission: 29-05-2022 Date of Acceptance: 10-06-2022

I. Introduction

Costumer maintaining is a huge problem for any service providing company. Research articulates that, compared to generating new customers it is difficult to maintain the current customers as market is highly competitive (Fridrich, 2017). Therefore, it will be beneficial for service provider to predict category of customers which might discontinue. Churning of customer can be predicted by analyzing behavior and patterns from data. So, for the set of targeted customers which are likely to be churn, customer retention programs are organized to maintain relationship. This Customer Relationship Management (CRM) program is mainly based on selective marketing focused on categories of customers (Pendharkar, 2009). For customer segmentation there are many machine learning techniques which can be used such as Artificial Neural Networks, Deep Learning, Evolutionary algorithms, etc.

Machine learning models are very efficient for predicting task and giving promising results. While working with big and abstract dataset Deep Learning gives better results. Deep Neural Network (DNN) and Recurrent Neural Network (RNN) are part of deep learning in machine learning made up of Artificial Neural networks. Both can be used for both classification and regression purpose. Here for customer segmentation, two models are proposed based on Deep Learning, Bidirectional Gated Recurrent Unit (BIGRU) which is a part of RNN and Genetic Deep Neural Network (GDNN) which is a combination of evolutionary algorithm and deep learning. BIGRU attempts to increase the efficiency of model by using leaky relu activation function and GDNN attempts to decrease the computation cost by initially generalizing weights for the DNN. Here target variable is customer churned or not, so it is binary-class classification model.

II. Literature review

Pendharkar (2009) used genetic algorithm based neural network model for customer churning in cellular wireless network services. The attributes were subscription plan, monthly total peak usage in minutes, promotional mailing variable, and churn indicator. The model outperforms compared to the classic Genetic Algorithm and Neural Network models with 0.97 accuracy. The model is also compared with the statistical z-score model, which shows that genetic algorithm based neural network gives better results. One of the drawbacks of the model is it takes 134 to 346 minutes to compute. Also, author suggested that rank relevant inputs and fitness function can be modified.

Obiedat et al. (2013) used Genetic Algorithm and K-means approach for customer churn prediction in telecommunication industry. In the first stage the K-means algorithm was used to reduce the data set by clustering. From small clusters generated by K-means algorithm, two clusters with upmost number of churner and non-churner were chosen for modeling. In the second stage, GP is used to build classification model. The

churner rate obtained by model in fold-3 was 70.2% and accuracy was 91.4%. The model is applied on one cluster instead of whole dataset. So, results might differ when ratio of churn and non-churn customers changes. Obeidat et al. has clustered the data for modeling, here in author's own work, whole data is applied on model which gives more general results. Also, in author's own work, number of parameters were decreased which reduces the computation cost.

Rana et al. (2016) compared LSTM and GRU model for classification of emotion from speech. Data with six different emotions based on the speech of actors were collected. With Stochastic Gradient different impact such as number of GRU/LSTM cells and absence or presence of bias were distinguished. Compared results showed that both models perform well. LSTM preforms 4.6% better, but GRU takes 18.16% less time to compute.

Fridrich (2017) used artificial neural network model for customer churn prediction of an e-commerce retail company. Z-score and Principal Component Analysis are used for feature selection. Also, genetic algorithm was used for hyperparameter selection such as number of epochs, neurons, etc; where neural network with 7-4-2 neurons in layers with 350 epochs was considered suitable for this task.

Vijaya and Sivasankar (2017) used Particle Swarm Optimization model for customer churn prediction in telecommunication industry. PSO outperforms decision tree, naive bayes, K-nearest neighbor, support vector machine, random forest and three hybrid models. Though it was observed that this technique is more conservative in identifying classes. All the services provided by company was considered as an attribute, which increases the computation cost of model. Also, model was used to perform on different dataset to check strength of model.

Gruber and Jockisch (2020) studied Recurrent neural network models for classification and observed their performance. The dataset having picture stories were used and classified in 11 different categories. The results shows that GRU had 0.85 of accuracy and LSTM had 0.82 accuracy. Also, the observation showed that GRU model had higher true negative rate and can learn less prevalent content. Whereas LSTM model had higher positive rate and learn better high prevalent content.

Zhong and Li (2020) used deep learning Recurrent Neural Networks such as Gated Recurrent Unit and Long Short Term Memory for detecting churn customer signals. Model is applied on phone call transcripts dataset of customers. Word embeddings is used on Gated Recurrent Unit and Long Short Term Memory with assumption of 700 words per transcript to speed up the training. Results showed that Gated Recurrent Unit outperforms the Long Short Term Memory, Convolutional Neural Network models with 0.86 accuracy for phone transcript dataset.

Dilhara (2021) used deep learning based model for detecting harmful camouflaged websites. Deep Learning based 7 models were proposed with LSTM, GRU and CNN as non hybrid models and BIGRU-LSTM, LSTM-GRU, LSTM-BILSTM and LSTM -LSTM as hybrid models. The target variable is to filter out malicious websites in form of binary classification with 0 as genuine and 1 as malicious website. The concatenation of LSTM-BIGRU model outperformed other models and CNN model showed lower performance.

Seymen et al. (2021) used different machine learning techniques for retail supermarket customer churning prediction. They included information such as customers identification number, purchasing information on different categories of items, and transaction details. The major city customers were used for training purpose and further city customers were used for test purpose in modeling. The model was divided in two parts: customers purchasing with promotion discounts and non-promotional purchase. Regression, Neural network, Logistic Regression, Neural Network and Deep Learning techniques were used where Deep Learning model with promotional dataset preforms best with 0.90 accuracy.

Ghosh and Gor (2022) used K-means clustering and Random Forest Regression algorithms for sales prediction. They used clustering methods for ad campaigning analysis. First, ad groups are created using the K-Means clustering algorithm then Random Forest Regressor algorithm is used to optimize sales conversion and predict future sales. Impressions, clicks, and spent are used as independent variables to predict total number of people that asked about the product after viewing the ad on Facebook. They also calculated Mean Absolute Error and Root Mean Square Error. The integration of two algorithms K-means clustering and Random Forest regressor gives permissive result with 75% accuracy.

III. Data Collection

Here, curated telecom company data is obtained, containing information of 3333 customers (from Kaggle.com). The given telecom dataset includes charges, number of calls, time duration of day, evening and night calls respectively, international calls detail, area of customer, length of the account and churned customers information.

- Data pre-processing: Parameters such as state of customer, voice mail, voice mail messages were excluded from the dataset due to less relevance. High correlation filter is applied to check the correlation between every parameter. Here, parameters such as time duration of day, evening, night and international call

found highly corelated with charges of day, evening, night and international call respectively. Hence only one of two i.e. call charges are chosen. As it is considered to have alike information and follow same movement.

- Data Cleaning: The null/Nan values from dataset are removed. Final dataset contains of 3324 columns. Also, no outliers are removed from dataset.
- Train and test data: 80% of data is used for training purpose and 20% of data is used for testing purpose.

Stat	Accou	Area code	Internatio	Voice mai	Number v	Total day	Total day	Total day	Total eve	Total eve	Total eve	Total nigh	Total nigh	Total nigh	Total intl	Total intl	Total intl	Customer	Churn
KS	128	415	No	Yes	25	265.1	110	45.07	197.4	99	16.78	244.7	91	11.01	10	3	2.7	1	FALSE
OH	107	415	No	Yes	26	161.6	123	27.47	195.5	103	16.62	254.4	103	11.45	13.7	3	3.7	1	FALSE
NJ	137	415	No	No	0	243.4	114	41.38	121.2	110	10.3	162.6	104	7.32	12.2	5	3.29	0	FALSE
OH	84	408	Yes	No	0	299.4	71	50.9	61.9	88	5.26	196.9	89	8.86	6.6	7	1.78	2	FALSE
OK	75	415	Yes	No	0	166.7	113	28.34	148.3	122	12.61	186.9	121	8.41	10.1	3	2.73	3	FALSE
AL	118	510	Yes	No	0	223.4	98	37.98	220.6	101	18.75	203.9	118	9.18	6.3	6	1.7	0	FALSE
MA	121	510	No	Yes	24	218.2	88	37.09	348.5	108	29.62	212.6	118	9.57	7.5	7	2.03	3	FALSE
MO	147	415	Yes	No	0	157	79	26.69	103.1	94	8.76	211.8	96	9.53	7.1	6	1.92	0	FALSE
LA	117	408	No	No	0	184.5	97	31.37	351.6	80	29.89	215.8	90	9.71	8.7	4	2.35	1	FALSE
WV	141	415	Yes	Yes	37	258.6	84	43.96	222	111	18.87	326.4	97	14.69	11.2	5	3.02	0	FALSE
IN	65	415	No	No	0	129.1	137	21.95	228.5	83	19.42	208.8	111	9.4	12.7	6	3.43	4	TRUE
RI	74	415	No	No	0	187.7	127	31.91	163.4	148	13.89	196	94	8.82	9.1	5	2.46	0	FALSE
IA	168	408	No	No	0	128.8	96	21.9	104.9	71	8.92	141.1	128	6.35	11.2	2	3.02	1	FALSE
MT	95	510	No	No	0	156.6	88	26.62	247.6	75	21.05	192.3	115	8.65	12.3	5	3.32	3	FALSE
IA	62	415	No	No	0	120.7	70	20.52	307.2	76	26.11	203	99	9.14	13.1	6	3.54	4	FALSE

Figure 1: Dataset description with all parameters

IV. Model Description of Bidirectional GRU With Dense Layers (BIGRU)

Bidirectional Gated Recurrent Unit (BIGRU) can be considered as an upgraded or improvised version of standard recurrent neural network. It was introduced by Cho, et al. in 2014 and known as variation of LSTM. GRU's main aim is to solve vanishing gradient problem which occurs in standard Recurrent neural networks. And BIGRU's main aim is to solve sequential data but it provides notable results for non-sequential data too. BIGRU is made-up forward and backward directional hidden layer. Hidden layer contains mainly two gates Update gate and Reset gate that helps in how much or which information to be passed in future and which past information to be forget shown in figure 2 (A). In Bidirectional GRU the information obtained in time step t is passed on to the time step t+1, in a similar way information is passed on in opposite direction also i.e. in direction from t+1 to t as shown in figure 2 (B). For every time step, information obtained from both directions are combined and output is computed using hard sigmoid activation function.

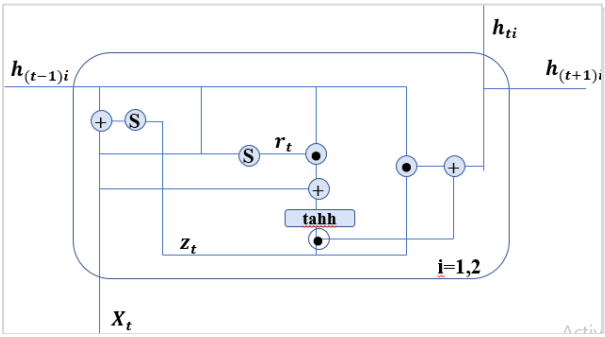

Figure 2(A): GRU structure for update and reset gate

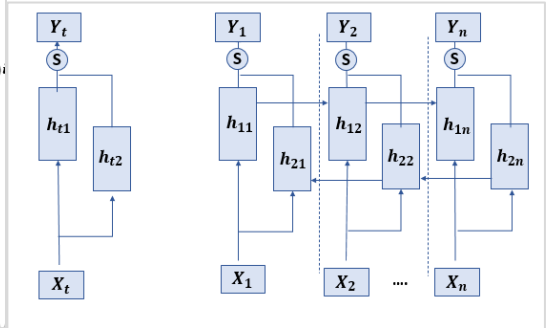

Figure 2(B): Working of Bidirectional GRU

For time step t, update gate denoted as z_t and defined as $z_t = \sigma(W_z x_t + U_z h_{t-1})$. Similarly reset gate denoted by r_t and defined as $r_t = \sigma(W_r x_t + U_r h_{t-1})$. Equivalent manner hidden state computed as $h'_t = tanh\,(W x_t + r_t \odot U h_{t-1})$. And the final output computed as $h_t = z_t \odot h_{t-1} + (1 - z_t) \odot h'_t$. Where, x_t is an input of current state, h_{t-1} is an output of previous hidden state and W_z, W_r, U_z, U_r, W, U are weights. $\odot$ this notation shows the pairwise multiplication or a Hadamard product. (Rajpurohit et al. 2021)

Adaptive moment estimation – Adam optimizer

$w'_t = w_t - \frac{\propto V'_t}{\sqrt{S'_t + e}}$ where $V'_t = \frac{V_t}{1-\beta_1^t}$, $S'_t = \frac{S_t}{1-\beta_2^t}$, $V_t = \beta_1 V_{t-1} + (1-\beta_1)\frac{\partial error}{\partial w_t}$ and $S = \beta_2 S_{t-1} + (1-\beta_2)\left[\frac{\partial error}{\partial w_t}\right]^2$. Where V exponential moving average and S squared moving average of gradients are initially taken as 0. Default values taken as $\propto = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $e = 10^{-8}$. (Kingma and Ba, 2015)
Since our dataset does not have same number of churn and non-churn customers it is likely for a model to be biased. So, to decrease bias, leaky relu and hard sigmoid are used as an activation function.

V. Proposed Methodology (BIGRU)

After several experimental results, an optimal model with three dense hidden layers, each having (10,50,50,10,2) neurons respectively is developed. Same as the number of parameters in the model, 10 neurons are taken in input and 2 neurons in output layer.

- Activation function 'leaky relu' is used for hidden layers and 'hard sigmoid' for output layer as there are 2 different categories to classify.
- Adam optimizer is used for backpropagation to select the best fit model by comparison. (The optimizer is considered in model because it has ability to update learning rate as well as gradient at every step.)
- Truncated normal is taken as kernel initializer, batch size=32, dropout=0.5, leaky relu alpha=0.001 and epochs=50 are considered for model.
- Model is developed in Python language.

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 bidirectional (Bidirectiona  (None, 10, 20)           780
 l)

 dense (Dense)               (None, 10, 50)            1050

 dense_1 (Dense)             (None, 10, 50)            2550

 dense_2 (Dense)             (None, 10, 10)            510

 bidirectional_1 (Bidirectio  (None, 20)               1320
 nal)

 leaky_re_lu (LeakyReLU)     (None, 20)                0

 dense_3 (Dense)             (None, 2)                 42

=================================================================
Total params: 6,252
Trainable params: 6,252
Non-trainable params: 0
_________________________________________________________________
```

Figure 3: Summary of the purposed model

VI. Model Description of Deep Neural Network + Genetic Algorithm (GDNN)

Deep Neural Network model is initialized, these two neural networks are generated with random weights, all weights are gathered from models to create a population. Then neural networks are forward propagated to pass fitness score which is set to 0.9 (Exit if optimal model matches the fitness score). From these top 5 values (fitness score), randomly 2 parents are selected for offspring generation and iterated through all weights for crossover. Weights are added to generated offspring for mutation. These weights are updated by back propagation algorithm to append optimal solution. Here, weights for hidden layer neurons are updated by $h_{311}{}'= h_{311} - l(2(\mathrm{p}-\mathrm{a})h_{31})$, $h_{321}{}' = h_{321} - l(2(\mathrm{p}-\mathrm{a})h_{32})$, $h_{211}{}' = h_{211} - l\,(2(\mathrm{p}-\mathrm{a})h_{311}h_{21})$, $h_{111}{}' = h_{111} - l\,(2(\mathrm{p}-\mathrm{a})h_{311}h_{211}h_{21})$ where p is predicted value, a is actual value, l is learning rate.

VII. Proposed Methodology (GDNN)

- After several experimental results, an optimal model with three dense hidden layers, each having (14,2,2,2,1) neurons respectively is developed. Same as the number of parameters in the model, 14 neurons are taken in input and 1 neuron in output layer.
- Activation function 'relu' is used for hidden layers and 'sigmoid' for output layer as there are 2 different categories to classify.
- Adam optimizer is used for backpropagation to select the best fit model by comparison. (This optimizer is considered in model because it has ability to update learning rate as well as gradient at every step.)
- Model is developed in Python language.

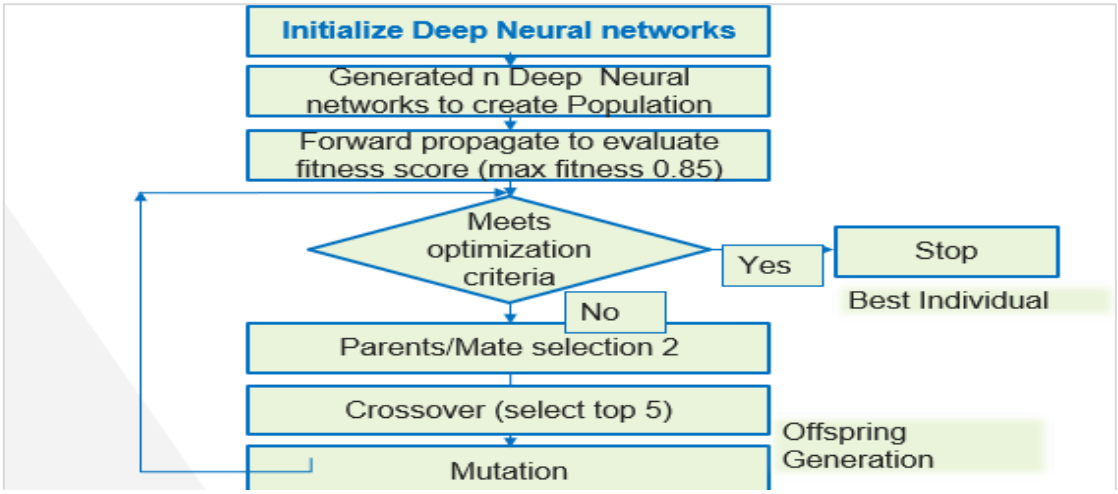

Figure 4: Flow chart of Deep neural networks and Genetic algorithm

VIII. Result and discussion

BIGRU and GDNN models are compared with vanilla artificial neural network. Various experiments were carried out on every model to find the optimized solution. Experimental results,

- For BIGRU, activation function such as relu and tanh were used which were not sufficient to give promising result due to the imbalance in churn and non-churn ratio. So, leaky relu is used with Adam optimizer, where accuracy is 78.13 with recall is 86.46 and F_1 score is 87.36 for training dataset and 78.07 accuracy with recall is 83.88 and F_1 score is 87.47.

- For GDNN, different combination of activation functions is used but relu provided best result with 0.88 accuracy and 0.84 fitness score. Also, nadam optimizer with relu gives accuracy of 0.85 with 0.85 fitness score which is close enough to the results of adam optimizer.

- For ANN, activation function relu is used with adam optimizer. Also, two hidden layers with 10 and 25 neurons are used in architecture of the model. Model gives 0.56 accuracy with train set and 0.64 accuracy with test set.

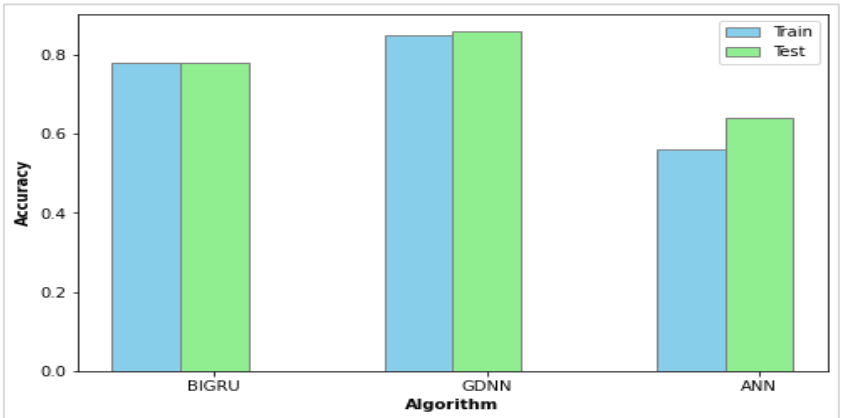

Figure 5: Accuracy of train and test dataset calculated by algorithms

Table 1: Accuracy of train and test dataset

Algorithm	Accuracy	
	Train	Test
BiGRU	0.7813	0.7807
GDNN	0.85	0.86
ANN	0.56	0.64

IX. Conclusion

Here customer segmentation in form of churn and non-churn customers is carried out. Two models are proposed based on Deep Learning, BIGRU and GDNN. The reason of selecting BIGRU is due to its faster and better performance with small dataset as compared to other RNN models. Also, it deals with under fitting problem due to outliers. And reason of selecting GDNN is that it decreases the computation cost by initially generalizing optimal weights for the DNN with the help of Genetic Algorithm. The model for churn customer is developed to overcome the drawbacks such as computation cost. Here model is applied on whole data which gives more general results. Also, number of parameters in this proposed model were decreased which reduces the computation cost. Results indicates that BIGRU and GDNN performed well as compared to the vanilla neural network. Accuracy of GDNN is high as compared to BIGRU. But BIGRU performed equally well as the value of recall and F1 score is high. Further, same model is capable for including more complex information. Future work can be done by applying different real life classification problems on the model.

References

[1]. B.A.S. Dilhara, "Phishing URL Detection: A novel hybrid Approach using Long Short-Term Memory and Gated Recurrent Units", *International Journal of Computer Applications*, ISSN: 0975 – 8887, Volume 183 – No. 44, December 2021.

[2]. Buscema, Massimo, "Back Propagation Neural Networks, Substance Use & Misuse", Volume 33, Issue 2, Print ISSN: 1082-6084 Online ISSN: 1532-2491, pp. 233–270, 1998.

[3]. Diederik P. Kingma and Jimmy Ba, "Adam: A Method for Stochastic Optimization", *3rd International Conference for Learning Representations*, San Diego, 2015.

[4]. Rajib Rana, Julien Epps , Raja Jurdak , Xue Li , Roland Goecke , Margot Breretonk and Jeffrey Soar, "Gated Recurrent Unit (GRU) for Emotion Classification from Noisy Speech", arXiv:1612.07778v1, pp 1-9, Dec 2016. Available online: http://arxiv.org/ pdf/1612.07778v1

[5]. S. Venkatesh and Dr.M. Jeyakarthic, "Adagrad Optimizer with Elephant Herding Optimization based Hyper Parameter Tuned Bidirectional LSTM for Customer Churn Prediction in IoT Enabled Cloud Environment", *Webology*, ISSN: 1735-188X, Volume 17, Number 2, December, 2020.

[6]. Vaidehi Rajpurohit, Shachi Bhavsar and Ravi Gor, A comparision of GRU-based ETH price prediction, Proceeding of International Conference on Mathemaitcal Modelling and Simulation in Physical Sciences (MMSPS-2021), pp 424-431, 2021.

[7]. Venkata Pullareddy Malikireddy and Madhavi Kasa, "Customer Churns Prediction Model Based on Machine Learning Techniques: A Systematic Review, Atlantis Highlights in Computer Sciences", Volume 4, pp 67-174.

[8]. Madhumita Ghosh and Ravi Gor, "Ad-Campaign Analysis and Sales prediction using K-means Clustering and Random Forest Regressor", *IOSR-Journal of Mathematics*, e-ISSN: 2278-5728, p-ISSN: 2319-765X, Volume 18, Issue 2, 10.9790/5728-1802021014 **,** pp 10-14, 2022.

[9]. Homa Meghyasi and Abas Rad, Customer churn prediction in telecommunication industry using data mining methods, *Proceedings of the 3rd International Conference on Integrated Intelligent Computing Communication & Security, Iran. Innovaciencia*; Volume 8, Issue 1, pp 1-8, 2020.

[10]. Martin Fridrich¸ "Hyperparameter Optimization of Artificial Neural Network in Customer Churn prediction using Genetic Algorithm", *Trendy Ekonomiky A Managementu Trends Economics And Management*, ISSN 1802-8527, Volume 28, Issue 1, pp 9-21, 2017.

[11]. Nicole Gruber and Alfred Jockisch, "Are GRU Cells More Specific and LSTM Cells More Sensitive in Motive Classification of Text?", doi.org/10.3389/frai.2020.00040, Volume3, article 40, 30 June 2020.

[12]. Parag C. Pendharkar, "Genetic algorithm based neural network approaches for predicting churn in cellular wireless network services", *Expert Systems with Applications*, Volume 36, pp 6714–6720, 2009.

[13]. Omer Faruk Seymen, Onur Dogan and Abdulkadir Hiziroglu, "Customer Churn Prediction using Deep Learning", DOI: 10.1007/978-3-030-73689-7_50, pp 1-11, 2021.

[14]. Junmei Zhong, William Li, "Detecting Customer Churn Signals for Telecommunication Industry Through Analyzing Phone Call Transcripts with Recurrent Neural Networks", *Research Gate*, pp 1-12, 2020.

[15]. J. Vijaya and E. Sivasankar, "An efficient system for customer churn prediction through particle swarm optimization based feature selection model with simulated annealing", *Cluster Comput*, 22:S10757– S10768, 2017.

[16]. Ruba Obiedat, Mouhammd Alkasassbeh, Hossam Faris and Osama Harfoushi1, "Customer churn prediction using a hybrid genetic programming approach", Volume 8, Issue 27, pp. 1289-1295, 18 July, 2013 ISSN 1992-2248.

[17]. Timothy Dozat, "Incorporating Nesterov Momentum into Adam", 2015.

[18]. Zainuddin Z., P. Akhir E. A., Hasan M. H, "Predicting machine failure using recurrent neural network gated recurrent unit (RNN-GRU) through time series data", *Bulletin of Electrical Engineering and Informatics*, Volume 10, No. 2, pp. 870~878, ISSN: 2302-9285, 2021.

Bhavsar Shachi. "Comparison of Back propagation algorithms: Bidirectional GRU and Genetic Deep Neural Network for Churn Customer." *IOSR Journal of Computer Engineering (IOSR-JCE)*, 24(3), 2022, pp. 07-12.