# Critque of “A Machine Learning Approach To Prevent Malicious Calls Over Telephony Networks”

ECE 535 Final Project
Master’s in data science,
University of Victoria, Canada
Mentored by,
Dr. Stephen W. Neville

Submitted by,
Arshiya Gulati
V00949938
Rajneesh Gulati
V00949939
Spring 2020
Email-id: arshiyagulati@uvic.ca
rajneeshgulati@uvic.ca

*Abstract*— **In this report, we are critiquing the paper “A Machine Learning Approach to Prevent Malicious Calls Over Telephony Networks” [1], talking about the approach taken by the authors of the mentioned paper and whether those approaches hold generally, or certain conditions are required. We will also talk about the ways those approaches can be improved.**

***Keywords — machine learning, malicious, scam, neural networks and random forest***

## I. Introduction

Aforementioned paper is trying to solve the problem of malicious calls by the proposition of detection and notifying the user that caller is a malicious caller and giving user options to take required steps, therefore, the paper attempts to be preemptive in its approach whereby the call can be detected as malicious beforehand. The paper builds on the previous approaches where the malicious calls were detected based on the statistics of the call. Therefore, we see the paper making use of previous research in this area and coming up with new methodologies to enhance previous solutions.

The data analysis used by the paper follows a two-step approach whereby the initial details of a call record are taken which are user ID which are basically Touch Pal user ID, call type which is a flag demonstrating whether it is either an incoming or outgoing call with the additional flag indicating whether the phone number is a Touch Pal user or not among other details. The paper uses these details as a base to build up an additional dataset based on two approaches which are historical information and cross-referencing information. These two are then used with eight other parameters to be fed into the machine learning models that the paper uses which are random forest model, neural networks, logistic regression models, as well as Support Vector Machine with an emphasis that non-historic features possess a static dimension while historic features of a call form a sequence of vectors.

The paper uses a large dataset of call records from the Touch Pal mobile app which comprises over 9 billion call records and these calls are spread in a span of three months through October 2016 to December. Further, the paper does an examination of the various aspects of this data, firstly using the criteria used in previous approaches but now taking advantage of the very large dataset to come up with criteria for coming up with a machine learning solution.

There are several assumptions made by the paper although they are not explicitly stated. The paper does a good job of avoiding previous assumptions of guaranteed initial access to underlying telephony network infrastructures to determine malicious calls. Although supported by empirical inference, the paper states that most numbers that are used to make the malicious call make over 91 calls each within the given period. Also, most malicious calls are made in the working hours and seldom over the weekends or lunch hour. All through the month, malicious calls are least made following these patterns and they also fall if the period is a holiday. Also, many malicious calls are not within the contact list of those called but those that are existent in the contacts are those of sale professionals and real estate agents. The assumptions are valid to a large extent because these conclusions are drawn from an analysis of a very large dataset of 9 billion call records. We see that these assumptions were not the same for previous works related to this subject, but it explicitly stated that this was since the dataset used in previous work was limited to only 700 records. All these are supported by evidence-based on the location of the dataset which is China and they may be different for another location, for example, the US [2].

The assessment begins with some desirable properties of the model to be implanted and these are that most benign calls ought not to be projected as malicious calls as well as a new malicious call should be identified by observing as a little number of malicious calls as possible and therefore more predictive. The paper used the four machine learning models stated earlier. The model implementation for the three models: SVM, logistic regression as well as a vanilla neural network, the paper uses the built-in implementations from sklearn library (Sci-Kit Learn). For the Random Forest Model, the paper uses two executions with one from the sklearn as well as the other from XGBoost. The evaluation metrics for the assessment involve two metrics, one of them being the AUC score, a standard in the assessment of a machine learning model's performance. This is in preference to other standards because data is skewed and has 100 times more negative examples than the positive ones. AUC score provides better information in case of skewed data.

## II. Assumptions

The authors of the mentioned paper have considered some assumptions for doing their analyses. We have divided this section into three parts where we explore the assumptions made whether they hold true in real world scenarios and if any special conditions are required to make them work.

### A. *Dataset*

As authors stated that they have 73M malicious call records out of total 9B records also out of 500M numbers only 800,000 are malicious numbers. We feel, although the data is large but malicious calls data is not enough and will not cover all kinds of malicious calls. Whatever amount of data is available, it is never enough, as ratio of malicious calls to all call logs is 0.008. We feel like choice of dataset could have been better, where number of malicious calls would be high.

Authors state that out of 73M malicious calls, 9.9M are in contact list of the users. Authors assume that users who want to do business with such numbers, they store them in their contact list, and others tag them as spam or malicious. We think that this is partially true, as users may tag a number on purpose due to some personal circumstances, which leads to type 1 errors during detection.

Authors make plots showing distribution of malicious calls based on time of the day, days of the week and dates of the month. Intuitively, these graphs look fine as they hold up with real world experiences of the people, where majority of malicious call are made during weekdays and between 9am-5pm which are the work hours. Also, authors see fall in malicious calls during public holidays, which is consistent with real world.

Author plots graph for distribution of active time of the numbers throughout 2016. They see a sharp decline in beginning and a gradual rise towards the end. Authors say that the reason for this to be users using a number for a short time or trial purposes. We think that initial high percentage can attribute to the fact that a number could be recycled that is previously it belonged to someone else that user deactivated, and it got assigned to a new user whose call logs happened to be in Touch pal database. So, we may need another feature in the feature vector to account for that, which is discussed in “Feature Vector” section.

The feature vector used in this paper can’t be used to train models on datasets of other countries, also not likely even other Chinese provinces which are not studied in this paper. For example, in India TrueCaller just like Touchpal is the biggest company that maintains call logs of its millions of users. As the format in which they save the data might be different so the same feature vector can’t be used, also not likely the same machine learning models would give the same performance on new dataset.

Sometimes user tags a call as harassment, spam etc. purposely even though authors said that TouchPal keeps a threshold between 30-100 tags, but what if users tag a number over the threshold and it gets marked as spam. Ultimately, the model would detect it as false positive, even though it is a benign call. Authors also say that users can tag only malicious call, so the source of original tag is unclear. Also, if a spammer registers for a new number and starts spamming with it before getting tagged over the threshold, initially that number will not be marked as spam which is a false negative. Also, authors verify whether a number is truly malicious by calling them, and if the number out of commission is assumed that the tags are correct, which is not a concrete assumption and further proves the point of a number being false positive.

### B. *Feature Vector*

“n_call”, it is a feature numbering how many times call is received from a number. Alone, this feature is not very useful, but authors have backed it up by features like “is_redial” and “gap_to_next”, telling whether a user redials or call back on the number and after many seconds that number called the user again. Which is a good assumption as mostly fraud calls work in a pattern and “gap_to_next” feature captures the pattern in which calls are received. Fact that pattern exists is backed up by the analyses that authors have done on the time of the day malicious calls are received even on the days of the week calls are made. So, this assumption holds in real world scenarios also.

We feel that there should be additional feature in the vector to check whether the number which is tagged as malicious is a personal number of some person, touchpal user or not, because that number will used for malicious activities as well as benign calls. Based on this feature users should be warned. In real world, a person doing private business might use his personal number to call people to sell his product, but users may tag him as malicious. If tags reach the threshold value, calls from that user will be marked spam, even though it is a correct classification, but this affects normal/personal life of that individual. Also, threshold values for such users should be kept high.

We feel that there should be a feature that accounts for a number being recycled. In real world scenarios, unused or deactivated number are reassigned to a new user. This can be known easily if the previous user was touchpal user. There can be serious implications of this as a number previously was used for malicious activities but now got reassigned to a genuine user, other users will see this number still marked as malicious and vice versa.

Authors have a feature “duration” in static features, they claim that short length of a call amounts to it being a malicious call, which is not the case every time although this feature is ed by others, we thought it should be pointed out.

Other features, such as “is_in_contact”, “call_type”, “weekday” etc. selected by the authors totally make sense and would perform good in real world scenarios.

### C. *Models Used*

“FP@(M,p)”, authors have taken a parameter first prediction and precision together. They collectively work to classify a call as malicious or benign. M acts as back up to threshold value and is defined by the authors as the number of times a phone number be tagged as malicious before classifying it as malicious or number of records to be read for that tag. Value of M can be varied as it is an integer value, authors have taken value from {1, 2,….,30}. Higher the value of M accurate the prediction but set the value too high problem of overfitting can be seen in some models. Authors have done experimentation with different values of M and p to choose the best combination.

Authors concluded that xgboost and NN perform the best as number of records checked by these models is 7 and 6 respectively. Which accounts for the time taken for the model to decide, in real world low the latency time better the model. Otherwise, users would have to wait for the model to process the information before they can take any action.

Authors have used model trained on data of the province of Beijing for other provinces and got accurate results. We think this because Beijing has highest number of malicious call records than any other provinces. So, data of other provinces can be said to be the subset of the data from Beijing. Free Lunch theorem and Ugly Duckling theorem still holds. If a new data set comes which is totally different from Beijing's data set then these models might not give the same performance.

Authors have done ablation analysis to find out which features work the best and are necessary. They found out that out of 29 features using 10 of them, similar results were obtained. This is a good thing as using less features makes the processing faster and reduced latency while deciding, which of course is a good thing in real world application. We suggest that PCA (principal component analysis) should also work, more efficiently for some models such as logistic regression, as it orthogonally projects the features avoiding mix up, provided classes are well separated.

Well, the use of the ROC AUC method to determine the machine learning model performance is insightful given the nature of the dataset where we have 100 times more negative examples than the positive ones. However, it would make more sense and give better outcomes if we combined this with the Brier Score or Brier Loss which helps to determine how close the prediction is to the real case, and in this case due to a very large dataset then we would average the data points.

$$BS = \frac{1}{N}\sum_{t=1}^{N}(f_t - o_t)^2$$

Figure 2.1

In general, this would increase the precision of the ROC AUC outcomes that the pair presents.

## III. INTELLIGENT ADVERSARIAL ENVIRONMENTS

As stated earlier, the presence of intelligent adversaries is very likely to make the results not hold in this case. The paper has quite several limitations especially in the form of haling intelligent adversaries. It is possible for an adversarial model to learn the approach used in the implementation of the proposed solution and trick the system to fail in flagging malicious calls. Major exploitation can be the mimicking of benign calls in such a way that the call numbers are spoofed and thereby trick the machine learning model that the numbers are saved in the users' contacts. This approach alone will render the solution presented by this paper useless. We can see very clearly that the use of adversarial models poses a very great risk to the approach presented by this paper especially in the area of call spoofing. For example, we have many cases where the approach presented would fail and this is in respect to the aspect of call spoofing fraud in the US (H,2017). The model needs to extend and include other aspects such as a determination of several aspects of the dataset such as analyzing data on telephony abuse in such a way that we can stem situational awareness as well as insights on the diverse telephony malicious calls and scams. Also, the model ought to improve in such a way that it can generate relevant and timely intelligence about abuses involving telephony used for easier detection, attribution, as well as mitigation purposes. Additionally, the model ought to be able to analyze in real-time the collected data so as detect the various scamming campaigns on a timely basis especially given the fact that these campaigns take place over a given period of time so that they can scam as many people as possible before they are detected as was the case in the US. This approach would most likely involve a supplementary approach to the one presented which would involve an initiative such a good framework which would be capable and very efficient and effective in collecting near real-time telephony grievances data as well as evaluate it in near real-time so as to come up with near real-time intelligence on the telephony abuse activities. This can be in the form of a web application that runs real-time subroutines that monitor all these aspects of calls [3]. The framework can carry out this analysis with the use of very crucial references such as (in the case of the US or Canada) the Canadian Numbering Administrator database and the North American Numbering Plan. (C.N.A. 2015) (N.A.N.P.A. 2015)

Moreover, the problem of call spoofing can be dealt with by incorporating, in addition to the aforementioned procedure, a mechanism that can leverage an incoming call’s properties. The objective here is to establish a verification process that will determine the connection between the originating number in a call and the call's state in comparison with the anticipated state of that call. This approach thus helps to accurately identify a call spoofing attack by a simple analysis of this information and the necessary steps can be taken to forestall this. CEIVE [4] The proposed approach can take advantage of this kind of an approach to refine its handling of the call records to do away with this identified limitation. Its admissible that this implementation may require additional information not available in the dataset, but the end goal is much worth it. Moreover, it is not very practical to overlook this kind of limitation because it is just so much prevalent. The common trends of malicious calls and telephony spamming and scamming really leverage on this one aspect and therefore it would really go a great way to improve the work proposed in this paper. Additionally, there is some slight oversight in the definition of malicious calls in the paper with a general categorization of all malicious calls. It is important to understand that malicious calls fall under very many auspices such as TDoS - Telephony Denial of Service, call spoofing, call spamming, and scamming amongst others.

## IV. Conclusion

Malicious calls are a serious problem not only in China but in every country, author of the mentioned paper have done a remarkable job in providing a solution to tackle this problem. In our attempt at critiquing their solution, we are looking at ways in which this solution can be applied to a much broader domain.

Inspired from this paper, the solution can be integrated, with some tweaks, into the infrastructure of a different country. Not every country has a helpline for malicious calls reporting as USA, there such solutions become very important.

## References

[1] H. Li *et al.*, "A Machine Learning Approach to Prevent Malicious Calls over Telephony Networks," *2018 IEEE Symposium on Security and Privacy (SP)*, San Francisco, CA, 2018, pp. 53-69.

[2] https://doi.org/10.1016/j.diin.2018.01.016

[3] Bordjiba, Houssem & Karbab, Elmouatez & Debbabi, Mourad. (2018). Data-driven approach for automatic telephony threat analysis and campaign detection. Digital Investigation. 24. S131-S141. 10.1016/j.diin.2018.01.016.

[4] Haotian Deng, Weicheng Wang, and Chunyi Peng. 2018. CEIVE: Combating Caller ID Spoofing on 4G Mobile Phones Via Callee Only Inference and Verification. In The 24th Annual International Conference on Mobile Computing and Networking (MobiCom ’18), October 29-November 2, 2018, New Delhi, India. ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3241539.3241573