Alteración de datos E2E: impacto de un ataque de envenenamiento y evasión en una red celular

Hao Qiang Luo-Chen, David Segura, Carlos Baena, Emil J. Khatib, Sergio Fortes, Raquel Barco.
hao@ic.uma.es, dsr@ic.uma.es, jcbg@ic.uma.es, emil@uma.es, sfr@ic.uma.es, rbm@ic.uma.es
Instituto de Telecomunicación (TELMA), Universidad de Málaga. Bulevar Louis Pasteur 35, 29010 Málaga (España)

***Resumen*—The evolution of mobile networks is currently going through a stage of opening up the infrastructure, known as O-RAN, a paradigm that also proposes providing more intelligence to the Radio Access Network (RAN). The key element that allows this change is the RAN Intelligent Control (RIC). Possible service improvements to customers are affected by new security breaches that may occur on the network. This paper analyses the impact of poisoning and evasion attacks, where training and testing data, respectively, are altered on Machine Learning (ML) algorithms. To this end, an E2E scenario has been analysed, in which the direct effects on users' perception are studied.**

I. Introducción

La Open Radio Access Network (O-RAN) [1] promueve la evolución de las redes mediante la apertura de las especificaciones de la infraestructura de la red de acceso celular, estableciendo nuevos elementos lógicos y definiendo interfaces entre ellas, así como la dotación de inteligencia a la RAN. Para esto último, el elemento encargado es el *RAN Intelligent Control* (RIC), que hace uso de algoritmos de Inteligencia Artificial (IA), y que trae consigo retos a nivel de seguridad [2]. En este artículo se centrará en aquellos ataques a la integridad de los datos de los modelos de Aprendizaje Automático (AA), llamados envenenamiento y evasión; que tienen como consecuencia el empeoramiento el servicio de los usuarios. El envenenamiento se enfoca en modificaciones de los datos de entrenamiento, haciendo que los modelos establezcan relaciones no esperadas entre los parámetros que analiza. Y, en el caso de evasión, las alteraciones se realizan sobre los datos que se desean evaluar después de entrenar.

Existe una literatura extensa sobre métodos para defenderse ante ataques de envenenamiento, como se recoge en [3]. Sin embargo, esta tiene un enfoque genérico, faltando estudios que ayuden a conocer el impacto que dichos ataques pueden tener sobre la gestión de redes O-RAN. Asimismo, habitualmente todos los casos estudiados se centran en problemas de clasificación [4]. No obstante, como se indica en [5], las métricas que permiten caracterizar el comportamiento *End-to-End* (E2E) de una red celular son continuas, las conocidas como *Key Quality Indicators* (KQIs); siendo su estimación basada en regresores [6]. Cuando se atacan los modelos usados para la regresión, mediante envenenamiento o evasión, se puede producir deterioro de los servicios recibidos por los usuarios. Algunos ejemplos de envenenamiento son el de [7], donde se centra en las portadoras radio de la red, o el de [8], en el que se falsea la localización de los usuarios con nodos de referencia incorrectos. En ambos casos, se ofrece una visión limitada del impacto a un tramo en particular de la red. En cuanto a los recursos utilizados en la literatura para el estudio de envenenamiento, el enfoque seguido suele basarse en simulación [9]. Por otra parte, hasta donde conocen los autores, la evasión suele ser más ignorado; a veces aparece acompañando al envenenamiento [10] o, en cambio, es aplicado a otro campo [11], y una vez más centrado en los problemas de clasificación.

En este artículo, se proporciona un estudio del impacto del envenenamiento y la evasión en los datos de diferentes algoritmos de AA aplicados a la predicción de KQIs, para la optimización E2E de la experiencia de usuario. Considerando que la literatura se enfoca en problemas de clasificación, se proponen métricas que ayudan a cuantificar el impacto de los dos ataques mencionados en un contexto de regresión. Este estudio se ha llevado a cabo usando equipamiento capaz de recrear las condiciones de comunicación de usuarios reales.

El resto del documento se organiza de la siguiente manera. En primer lugar, en la Sección II se describe el contexto del estudio; que es la importancia de una visión E2E para considerar el impacto en la percepción del usuario y dos tipos de ataques diferentes (envenenamiento y evasión). En la Sección III, se muestra el testbed utilizado y los resultados que obtenidos. Finalmente, en la Sección IV, se presentan las conclusiones y líneas futuras que derivan de este trabajo.

II. Vulnerabilidades en redes celulares

A. Contexto

El RIC es la función lógica de O-RAN que permite optimizar las funciones y recursos de la red, mediante la recolección de datos de la red celular y los usuarios. Para lograr esto, existen dos componentes: el *non-real-time RIC* (Non-RT RIC) y *near-real-time RIC* (Near-RT RIC), que se aplican con resolución temporal de s y ms, respectivamente. Gracias a la actuación de aplicaciones dentro de ellos, llamadas rApps y xApps respectivamente, se mejora el servicio provisto por la red; por ejemplo, mediante la distribución de carga de las celdas, para evitar la congestión en alguna celda con exceso de tráfico, o la asignación inteligente de recursos de usuarios. La consecución de las funciones anteriores se logra mediante algoritmos de AA; como el uso de regresores que predigan la carga de celdas vecinas para distribuir los usuarios; o, mediante estimadores de las necesidades de recursos radio de los usuarios, según su posición geográfica y servicio utilizado.

Entre los tipos algoritmos de AA, el foco central de este artículo son los métodos supervisados; que son aquellos con una etiqueta que desea predecir, y el algoritmo identifica patrones de los datos de entrada asociados a cada una de ellas. La alta dependencia hacia dichos métodos de AA, abre un abanico de posibilidades para alterar el comportamiento

deseado de estos. Estas alteraciones son conocidos como ataques, y se clasifican en:

- **Envenenamiento**: se alteran las muestras de entrenamiento; esto podría ser, en una red celular, enmascarar usuarios con un alto SNR por otro de menor calidad, para que la red le destine más recursos de los necesarios.
- **Evasión**: se modifican los datos de evaluación; como el caso de usuarios con bajo RSRP, cuyo valor es falseado para mostrarse como mayor, y, por ello, se intenta transmitir más datos de lo que el usuario es capaz de recibir adecuadamente.
- **Inversión de modelo**: se hallan las características y arquitectura del modelo de AA; como la identificación de la ponderación de redes neuronales de un sistema propietario de ajuste de potencia de antenas.
- **Inferencia de miembros**: se identifican las muestras usadas para crear el modelo de AA; por ejemplo, la identificación de usuarios específicos del escenario de entrenamiento.

Ante la importancia de los algoritmos de AA en las redes celulares, se analizará la robustez de diferentes métodos ante el envenenamiento y evasión. En concreto, de cómo estos afectan a la regresión de variables para la estimación de KQIs.

B. Envenenamiento

Durante el entrenamiento de un algoritmo de AA, se analizan las relaciones entre los parámetros. Un ataque se basa en inyectar nuevas muestras que siguen patrones diferentes del resto, como el cambio de etiquetas para asociarla a otro caso. En caso de no re-entrenar con frecuencia los modelos de AA, el impacto en un operador puede perdurar mucho tiempo. Una forma básica de detectar un envenenamiento, es el estudio de datos anómalos dentro de los datos de entrenamiento; para lo cual se pueden aplicar desde métodos estadísticos como otros basados en algoritmos de AA para identificar dichas muestras.

C. Evasión

Una vez que se posee un modelo de AA entrenado, este predictor es aplicado para estimar las etiquetas de muestras no vistas anteriormente, y que son el objetivo de este ataque. Se alteran las muestras de los usuarios, y se ocultan dichas modificaciones para que pasen desapercibidas entre aquellas que son legítimas, razón por la que el impacto se limita solo a las muestras modificadas. No obstante, el error de predicción puede ser potencialmente más significativo cuando la gestión de la red dependa de esas predicciones. El enfoque para detectar la presencia de este problema se basa también, frecuentemente, en identificar la presencia de datos cuyas características no concuerdan con respecto al resto del conjunto.

D. Propuesta de métricas

En la literatura no se definen métricas específicas para los problemas anteriores, ya que se centra en la comparativa de medidas como la precisión o exactitud en clasificaciones. La falta de métricas específicas para cuantificar la variación producida por los ataques dificulta un consenso en la comparativa de resultados. Por ello, se definen en este artículo diferentes propuestas. En primer lugar, la métrica habitual para medir el error en las predicciones es la siguiente,

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\mathrm{KQI}_i - \widehat{\mathrm{KQI}}_i)^2}; \tag{1}$$

siendo KQI la métrica que se quiere predecir, $\widehat{\mathrm{KQI}}$ el valor predicho y N el número total de muestras de la métrica. Para el estudio del impacto de los ataques, se proponen las siguientes métricas, generalizables más allá del RMSE:

$$\epsilon_j = \frac{\mathrm{RMSE}_j^{\mathrm{Ataque}} - \mathrm{RMSE}_j^{\mathrm{Base}}}{\mathrm{RMSE}_j^{\mathrm{Base}}} 100; \tag{2}$$

donde $\mathrm{RSME}_j^{\mathrm{Ataque}}$ y $\mathrm{RSME}_j^{\mathrm{Base}}$ son el RSME del algoritmo i de AA donde las muestras han sido atacadas o no, respectivamente. Esta medida del error relativo permite resaltar la variación después de un ataque, independientemente de la magnitud de la variables bajo estudio. Además, se proponen diferentes métricas cruzadas, aplicadas sobre los valores de la Eq. 1, que facilita el estudio del impacto, siendo el vector s_X

$$s_X = \frac{\mathrm{RMSE}^{\mathrm{Ataque}} - \mu_{\mathrm{RMSE}^{\mathrm{Base}}}}{\sigma_{\mathrm{RMSE}^{\mathrm{Base}}}}; \tag{3}$$

donde $\mu_{\mathrm{RMSE}^{\mathrm{Base}}}$ y $\sigma_{\mathrm{RMSE}^{\mathrm{Base}}}$ es la media y la desviación típica, respectivamente, del RMSE de las técnicas de AA estudiadas sin estar bajo el efecto de un ataque. La obtención de la media μ_X y la desviación típica σ_X de s_X permite evaluar la variación respecto al caso base; es decir, si $\mu_X \neq 0$ significa que la media del RMSE de las técnicas de AA tras el ataque es diferente. De igual forma, la interpretación de $\sigma_X \neq 1$ es análogo a lo anterior. Finalmente, se ha diseñado una métrica que recoge en un único valor, permitiendo una comparativa directa del conjunto de diferentes técnicas de AA, la variación tras un ataque[1],

$$\rho_X = \mathrm{sgn}(\sum_{j=1}^{M}\Delta_j)\frac{1}{\mu_{\mathrm{RMSE}^{\mathrm{BASE}}}}\sqrt{\frac{1}{M}\sum_{j=1}^{M}(\Delta_j)^2}; \tag{4}$$

donde M es el número de algoritmos que están bajo estudio y $\mathrm{sgn}()$ es el operador signo; por lo tanto, el $\mathrm{sgn}(\rho_X)$ permite conocer la tendencia del cambio de predicción entre antes y después del ataque; y $|\rho_X|$ es la desviación que existe entre un caso y otro. Y, siendo $\Delta_j = \mathrm{RMSE}_j^{\mathrm{Ataque}} - \mu_{\mathrm{RMSE}^{\mathrm{Base}}}$.

III. Estudio de impacto

A. Testbed

Con el fin de recrear un escenario realista, se ha planteado un servicio de vídeo bajo demanda, de Youtube, solicitado por un cliente. Esto tiene como objetivo plantear un escenario equivalente a que un operador recogiese los datos de un usuario para modelar la relación de los parámetros radio y la posterior predicción de KQIs, mediante algoritmos de AA. Se ha tomado de base lo que está desarrollado en [5]. En dicho trabajo, se planteó un testbed de red 5G SA para distintos servicios multimedia, en el que se establece un único usuario que reclama recursos de la red, automatizado mediante Selenium WebDriver. En este estudio, se expande esta propuesta,

[1] ρ_X ha sido diseñado basándose en la formulación del coeficiente de variación [12], pero centrado en la media del RMSE sin ser atacado.

incluyendo usuarios que generan tráfico de fondo, para añadir realismo, y un usuario atacante. Por ello, los elementos de los que constan el escenario son, representados en la Fig. 1:

- **Portátil + CPE**: es el usuario principal, y a través de un CPE, un router avanzado con conexión a redes celulares, se recogen las métricas de la comunicación.
- **AMARI UE Simbox**: equipo encargado de desplegar una celda móvil 5G SA, y que permite la reconfiguración de distintas características de la red.
- **AMARI Callbox**: es un emulador de usuarios, para generar el tráfico de fondo. Específicamente, se han establecido 5 servicios que están en uso: 2 de vídeo de Youtube, 2 de FTP y 1 de ping.
- **Terminal móvil**: usuario conectado a la red, solicitando recursos, pero solo con interés de empeorar el servicio provisto al resto de los usuarios.

Fig. 1: Arquitectura física del set-up utilizado en este estudio.

Este setup permite obtener métricas que relacionan las condiciones radio y los KQIs. Como entrada a los algoritmos de AA, se han considerado los parámetros radio descritos en [6]; tales como RSRP, SINR, CQI y métricas relacionadas con la tasa de bits, entre otros. El servicio elegido para este estudio es el de vídeo bajo demanda, por lo que los KQIs seleccionados son:

- **Tiempo de inicio**: tiempo necesario para empezar a reproducir el vídeo.
- **Salud promedio del búffer**: cantidad de vídeo almacenado en el cliente.
- **Frames promedios descartados**: número de frames que han sido descartados por problemas con la red.
- **Eventos de congelación**: congelación de la reproducción de vídeo.

Asimismo, los algoritmos de AA seleccionados son los siguientes, e implementados con [13]:

- **Linear Regressor** (LR): modelo lineal en el que se minimiza el error cuadrático medio.
- **Random Forest** (RF): conjunto de árboles de regresión que son entrenados en subconjuntos aleatorios de los datos, para crear complementariedad entre los árboles.
- **K-Nearest Neighbors** (KNR): predicción basada en los K vecinos más cercanos al dato de entrada.
- **Support Vector Regressor** (SVR): regresión basada en la búsqueda de los vectores que maximicen la separación entre muestras.
- **Multi Layer Perceptron** (MLP): redes neuronales con arquitectura de multicapas.

Esta selección se basa en que son los algoritmos más habituales en el campo de predicción de KQI, como se muestra en [6]. Se han aplicado los algoritmos con la configuración por defecto de sus parámetros, ya que la optimización de hiperparámetros se considera fuera del alcance de este estudio.

B. *Descripción del experimento*

Con el fin de estudiar ataques de envenenamiento y evasión, el usuario principal solicitará contenido de forma continua y, al mismo tiempo, el tráfico de fondo también será constante. Además, dada la capacidad de configuración del AMARI UE Simbox, se ha modificado la cantidad de *Physical Resource Blocks* (PRBs) (5, 10, 15, 20, 25, 50, 75, 100) y la ganancia del canal (0, -10, -20dB), dando lugar a un total de 24 combinaciones entre ambos parámetros, para recrear un escenario realista con diferentes condiciones. Del usuario bajo estudio se recogen métricas radio que son utilizadas para entrenar los algoritmos de AA de predicción de los KQIs, ambos mencionados en la sección anterior. El atacante es un usuario con interés en crear patrones irregulares en los datos, mediante solicitudes de recursos en momentos específicos; haciendo que el sistema no asigne adecuadamente los recursos y crear una estimación incorrecta del estado del usuario principal.

Con todo este escenario, por una parte, se recolectarán muestras sin ninguna actividad por parte del usuario atacante (caso base) y, por otra, se recogerán también muestras con datos con un comportamiento anómalo (caso con ataque), generado por la presencia del atacante.

En la Tabla I se resume el número de muestras tomadas para cada tipo de ataque. En ambos casos se ha procurado seguir la regla de Pareto para la separación de los datos; es decir, un reparto de 80 % y 20 % para el entrenamiento y evaluación de los algoritmos. La presencia de muestras atacadas se considera aproximadamente equivalente en proporción, cercano al 25 %, aunque no son exactamente iguales por que se ha procurado mantener la misma cantidad de muestras atacadas para cada combinación de ataque, evitando la aparición de sesgo por una presencia desbalanceada de muestras en cada una.

TABLA I: Datos aplicados para el envenenamiento y evasión.

	Envenenamiento	**Evasión**
Muestras totales	3000	3000
Total entrenamiento/ muestras alteradas (% del entrenamiento)	2400/600 (25 %)	2400/0 (0 %)
Total evaluación/ muestras alteradas (% del evaluación)	600/0 (0 %)	600/288 (24 %)

C. *Análisis de resultados*

En primer lugar, se muestra el impacto en las métricas, recogido en la Fig. 2, donde se muestra primero el RMSE del caso base, tras el ataque y finalmente el error relativo. El caso base ya muestra que la salud promedio de búffer y los frames promedios descartados son los KQIs que mejor y peor se estiman, respectivamente. Además, se puede contemplar que el algoritmo MLP presenta el peor rendimiento en sus estimaciones. Continuando con el envenenamiento, del análisis de los valores de ϵ_j, se puede observar que el impacto tenido el mayor impacto en LR, habiendo un caso mayor al 90 %, mientras que ha sido poco significativo en RF y MLP, en los que se han producido casos de reducción

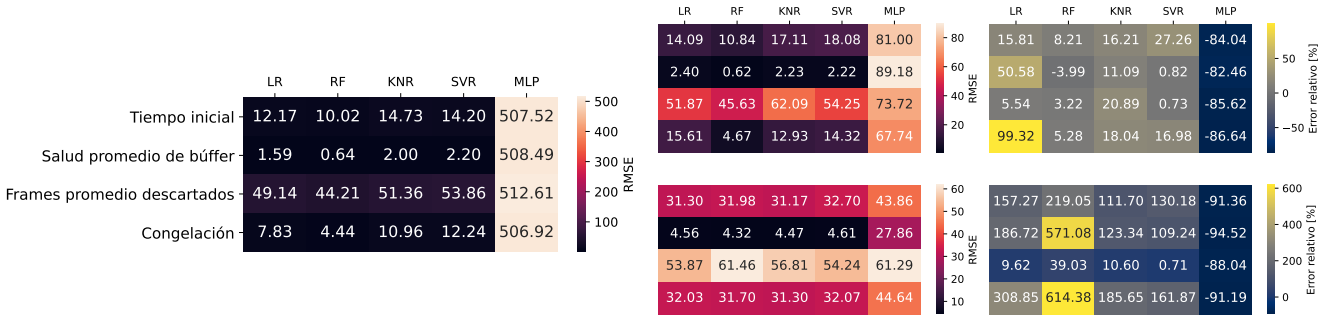

Fig. 2: Análisis del RMSE: caso base, con muestras alteradas y error relativo (arriba = envenenamiento, abajo = evasión).

del error. La presencia de muestras de ataque en el entrenamiento ha ayudado a crear diversidad en los datos, que ayudan a la predicción de muestras no vistas. Pasando al caso de la evasión, ϵ_j son de varias órdenes de magnitud mayores que su contraparte de envenenamiento, alcanzando errores de más de $500\,\%$. Y, al igual que antes, MLP es el más robusto, por un infrajuste del algoritmo debido a la falta de concreción de su arquitectura, otorgando una mayor capacidad de generalización en muestras nuevas. En cuanto al operador, estos resultados indican que un ataque de evasión puede repercutir más en las predicciones de KQIs, asignando recursos radio incorrectamente a los usuarios. En cambio, el ataque de envenenamiento puede tener una duración más larga, puesto que el problema afecta al modelo que se ha creado, y no solo las muestras que se están evaluando.

Después, en la Tabla II, se recogen los resultados de las variables cruzadas. Como se puede observar por ρ_X, existe mucha variación respecto a la media; ya que al basarse en el coeficiente de variación, cualquier valor mayor que 0.3 es considerado como una alta variación. Además, otro fenómeno que se puede observar aquí es que σ_X es menor que 0 en todos los casos, así que se puede afirmar que la presencia de ataque puede reducir la variación de RMSE de los algoritmos AA, que concuerda con la reducción de μ_X.

TABLA II: Resultados de las métricas cruzadas.

	Evenenamiento			**Evasión**		
	ρ_x	μ_x	σ_x	ρ_x	μ_x	σ_x
Tiempo inicial	-1.71	-0.42	0.13	-1.86	-0.39	0.02
Salud promedio de búffer	-1.82	-0.41	0.17	-2.09	-0.46	0.05
Frames promedios descartados	-1.38	-0.46	0.05	-1.42	-0.46	0.02
Eventos de congelación	-1.81	-0.43	0.11	-1.92	-0.37	0.03

IV. CONCLUSIONES

La aparición del paradigma de O-RAN ha supuesto una evolución hacia nuevas oportunidades de apertura de la red y de introducción de inteligencia en el acceso radio de los usuarios a las celdas móviles. Al mismo tiempo, esto supone un nuevo punto de origen para ataques hacia la red. En este artículo, se ha realizado un estudio del impacto de ataques de envenenamiento y evasión de los datos en algoritmos de AA, con objetivo la predicción de métricas E2E; en el que se ha podido observar que el envenenamiento ha tenido un efecto menor en las métricas analizadas, con respecto a la evasión, que puede impactar significativamente en la gestión de la red, en cuanto a la asignación de recursos al usuario final. Asimismo, se proponen distintas métricas de cuantificación del error de predicción, con el propósito de rellenar una falta de medidas específicas para estos problemas en la literatura. Como líneas futura se propone extender este estudio con la aplicación de las métricas propuestas sobre medidas como el R^2, MASE, MAPE o similares.

AGRADECIMIENTOS

Ministerio de Asuntos Económicos y Transformación Digital y la Unión Europea - NextGenerationEU, en el marco del Plan de Recuperación, Transformación y Resiliencia y el Mecanismo de Recuperación y Resiliencia bajo el proyecto MAORI. Además, también está parcialmente financiado por la Universidad de Málaga, a través de II Plan Propio de Investigación y Transferencia.

REFERENCIAS

[1] O.-R. Alliance, "-RAN WhitePaper - Building the Next Generation RAN," https://mediastorage.o-ran.org/white-papers/O-RAN.White-Paper-2018-10.pdf, (acceso Marzo 2024).

[2] M. Liyanage and et al., "Open RAN security: Challenges and opportunities," *Journal of Network and Computer Applications*, vol. 214, p. 103621, 2023.

[3] Z. Tian and et al., "A Comprehensive Survey on Poisoning Attacks and Countermeasures in Machine Learning," *ACM Computing Surveys*, vol. 55, no. 8, pp. 1–35, 2022.

[4] N. Pitropakis and et al., "A taxonomy and survey of attacks against machine learning," *Computer Science Review*, vol. 34, p. 100199, 2019.

[5] C. Baena and et al., "Video Streaming and Cloud Gaming services over 4G and 5G: a complete network and service metrics dataset," *IEEE Communications Magazine*, vol. 61, no. 9, pp. 154–160, 2023.

[6] ——, "Measuring and estimating key quality indicators in cloud gaming services," *Computer Networks*, vol. 231, p. 109808, 2023.

[7] S. Soltani and et al., "Poisoning Bearer Context Migration in O-RAN 5G Network," *IEEE wireless communications letters*, vol. 12, no. 3, pp. 401–405, 2022.

[8] Y. Li and et al., "Secure 5G positioning with truth discovery, attack detection, and tracing," *IEEE Internet of Things Journal*, vol. 9, no. 22, pp. 22 220–22 229, 2021.

[9] S. A. Khowaja and et al., "Spin: Simulated poisoning and inversion network for federated learning-based 6g vehicular networks," in *ICC 2023-IEEE International Conference on Communications*. IEEE, 2023, pp. 6205–6210.

[10] J. Liu and et al., "Adversarial machine learning: A multilayer review of the state-of-the-art and challenges for wireless and mobile systems," *IEEE Communications Surveys Tutorials*, vol. 24, no. 1, pp. 123–159, 2022.

[11] W. Jiang and et al., "Poisoning and evasion attacks against deep learning algorithms in autonomous vehicles," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 4, pp. 4439–4449, 2020.

[12] H. Abdi, "Coefficient of variation," *Encyclopedia of research design*, vol. 1, no. 5, 2010.

[13] F. Pedregosa and et al., "Scikit-learn: Machine learning in Python," vol. 12, pp. 2825–2830, 2011.