Red Neuronal Perceptrón Multicapa para la Clasificación de Malware Mediante la Extracción de Características MFCC

Brandon A. Herrera L., Jorge A. Alcántara A., Rogelio Reyes R., Clara Cruz R.
Instituto Politécnico Nacional, ESIME Unidad Culhuacán
Av. Santa Ana No. 1000, Col. San Francisco Culhuacán, CP. 04440, México, Ciudad de México
Tel. (55) 5729-6000, Fax (55) 56562058
e-mail: bherreral1601@alumno.ipn.mx, jalcantaraa@alumno.ipn.mx, rreyesre@ipn.mx, ccruzra@ipn.mx

***Resumen*- Actualmente los Malware (malicious software) son uno de los mayores problemas en el área de la cyber-información, generalmente son programas cuya función es dañar de manera premeditada el correcto funcionamiento de un sistema, ya sea a nivel de software o hardware. Todos los malware comparten ciertas similitudes, pueden ser sus objetivos dentro del sistema, el cómo es que logran ingresar al sistema, etc. Esta información se encuentra dentro de la estructura del malware y puede ser utilizada para su clasificación, ya que aquellos que compartan similitudes especificas se pueden catalogar como familias de malware. En este documento se propone desarrollar e implementar un sistema que permita abstraer las características esenciales del malware por medio de los Coeficientes Cepstrales en la Frecuencia de Mel (MFCC), de tal manera que se pueda clasificar en sus distintas familias empleando una red neuronal perceptrón multicapa, ya que conociendo la familia a la que pertenecen es más sencillo contar con protocolos de contención y eliminación. Los resultados presentados por el sistema, tomando como métrica de evaluación F1 Score demuestran una eficiencia en la mayoría de las familias entre el 90% y 100%.**

Keywords*—*Malware, MFCC, red neuronal perceptron multicapa, F1 Score.

I. Introducción

Desde la documentación de los primeros malware conocidos para ordenadores en los años 80's este tipo de programas han estado en un crecimiento exponencial y no sólo afectando en un ambiente personal, sino también dentro de áreas corporativas. Se considera malware a todo aquel software con propósitos malignos, como lo es el robo de información, secuestro de sistemas o equipo, saturación de la red, robo de identidad, espionaje, falsificación de documentación, entre otros. Estos se pueden clasificar en distintas familias como lo son: Virus, troyanos, software espía o Spyware, gusanos, Ransomware, adware, Botnets, entre otros [1].

Si bien ya existen métodos de clasificación efectivos, diariamente se crean aproximadamente 350,000 programas maliciosos nuevos [2], por lo tanto, se debe estar a la vanguardia en la clasificación del malware ya que esto nos permitirá tener protocolos de seguridad en caso de contingencia.

Existen 2 tipos de técnicas para el análisis de malware, el dinámico y estático [3].

- *Dinámico*. Los métodos dinámicos se caracterizan por ejecutar el código malicioso en un entorno controlado, esto permite supervisar los cambios realizados en sistema, archivos, comunicación, procesos, memoria.
- *Estático*. El método estático, no ejecuta el archivo malicioso, se analiza su código con herramientas de debugging, editores HEX, búsqueda de cadenas, entre otros.

II. Antecedentes

En [4] se presenta un método de clasificación de malware que representa los archivos binarios de malware como imágenes en escala de grises, donde se emplean dos descriptores de patrones binarios locales (LBP), uno basado en distancia y el otro basado en orientación, así como una red neuronal convolucional para la clasificación. El método propuesto aprovecha el hecho de que la mayoría de las variantes de malware de la misma familia tienen estructuras locales y globales similares, por lo que el cálculo de los descriptores binarios locales de las imágenes en escala de grises facilita la detección y clasificación de las variantes de un malware. Sus resultados experimentales demuestran una precisión aproximada al 98% para la detección de diferentes familias de malware.

En [5] se propone un método basado en técnicas de procesamiento de señales de audio. Se representan bytes binarios como señales de audio, del resultado anterior se le aplican técnicas de recuperación de información musical (MIR) y con ello detectar patrones utilizando coeficientes ceptrales en la frecuencia de mel (MFCC) además de características en cromagrama. Luego se crea un modelo de aprendizaje automático basado en las características extraídas anteriormente y así clasificar nuevas muestras. Para la evaluación de este método se realizaron varios experimentos los cuales mostraron una buena precisión y bajo uso de recursos como memoria y tiempo.

Este trabajo utiliza los MFCC de una señal generada a partir del código binario de un malware para su clasificación en familias utilizando una red neuronal perceptrón multicapa y demuestra una precisión arriba del 90% en la base de datos utilizada por [6].

II. SISTEMA PROPUESTO

Fig. 1. Diagrama a bloques de la propuesta de solución

Este sistema consta de dos fases principales; la fase de entrenamiento y la fase de pruebas como se muestra en la Fig. 1.

La etapa de entrenamiento servirá para obtener como resultado un modelo y los pesos adecuados de la red neuronal, mientras que en la etapa de pruebas se hará uso de los parámetros obtenidos en la etapa de entrenamiento, a lo que, teniendo un modelo y pesos fijos se puede probar y evaluar con cualquier muestra o muestras de malware para obtener su familia correspondiente.

A. Fase de Entrenamiento

La obtención de la señal de audio forma parte del diagrama a bloques general de la propuesta de solución como se puede observar en la Fig. 1, de acuerdo con éste entra una muestra de malware [7] al bloque "Obtención de la señal de audio digital" para realizar un preprocesamiento el cual se presenta en la Fig. 2.

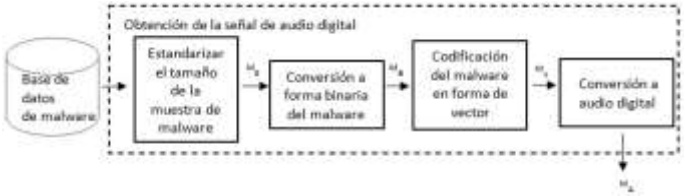

Fig. 2. Diagrama a bloques de la etapa de obtención de la señal de audio digital

a) Extraccion de coeficientes MFCC

Los MFCC son coeficientes para la representación del habla basados en la percepción auditiva humana, este método de obtención de características de una señal permite obtener los coeficientes más representativos de una señal de audio para su posterior análisis.

El proceso para la obtención de los MFCC es: ventaneo, transformada de Fourier (sin embargo, se emplea un algoritmo conocido como Transformada Rápida de Fourier o FFT por sus siglas en ingles), aplicación de banco de filtros de mel, logaritmo de la señal transformada y transformada coseno discreto. A continuación, se explicará cada uno de estos pasos más detalladamente.

b) Ventaneo

En el procesamiento de señales se recomienda trabajar en términos cortos, ya que se entiende que las muestras no serán tan cambiantes [7], por ello se hará uso de una función de ventana $W(n)$, en este caso será la ventana de Hamming, la cual es una función matemática para evitar discontinuidades durante el análisis de señales. Con esta función se pretende fragmentar en tramas o segmentos, denominados ventanas la señal. Se escogió que estas ventanas sean de *25 ms* con un traslape de *1ms* para evitar la pérdida de información.

c) FFT.

La transformada de Fourier es la manera en que se cambia una señal en el dominio del tiempo al dominio de la frecuencia. Esto se realiza ya que en el dominio del tiempo no hay información relevante para nuestro propósito, sin embargo, con la DFT (Transformada Discreta de Fourier) es posible pasar al dominio de la frecuencia donde se podrá analizar de manera óptima la señal. La DFT de una señal $x[n]$ definida en el rango $0 \leq n \leq N-1$ se define como:

$$X(k) = \sum_{n=0}^{N-1} X(n) e^{-j\frac{2\pi}{N}nk} \quad , \tag{1}$$

donde $W_N = e^{-j\frac{2\pi}{N}nk}$ y los valores espectrales $X[k]$ se evalúan en $0 \leq k \leq N-1$.

Para realizar de manera más sencilla este proceso, ya que la DFT exige una gran cantidad de cálculos, se emplea el algoritmo de la Transformada rápida de Fourier (FFT por sus siglas en inglés) la cual aprovecha las características de periodicidad y simetría de los coeficientes dados por:

$$W_N = e^{-j\frac{2\pi}{N}nk} \quad , \quad (2)$$

Por lo que la DFT en términos de la FFT se puede escribir como

$$X(k) = \sum_{n=0}^{N-1} X(n) W_N^{kn} \ \text{con } k = 1, \dots, N-1 \quad (3)$$

d) Banco de filtros de mel

La escala de mel es una forma de representar visualmente como el oído humano percibe las diferencias en la frecuencia de una señal, esto se debe a que los humanos somos mucho mejores para identificar cambios en las frecuencias bajas que en las altas [7]. En una escala normal de frecuencias no es posible ver este cambio como es percibido por el oído humano promedio, sin embargo, con la escala de mel, si se puede observar el cambio como lo identifica el oído humano.

Para convertir f hercios en Mel se emplea (4)

$$Mel(f) = 2595\ log_{10}(1 + \frac{f}{700}), \quad (4)$$

donde *f* corresponde con la frecuencia en Hertz.

Este reescalado, servirá como primer filtro ya que representará de manera más exacta los coeficientes que estamos buscando.

e) Logaritmo de señal transformada

Con la señal ya representada en el dominio de la frecuencia y en la escala apropiada de mel ahora se debe aplicar una función logarítmica. De esta manera se pasa al dominio de la potencia espectral logarítmica. Se realiza este paso, ya que aplicar el filtro de la escala de mel nos dará valores muy grandes y pequeños, lo cual dificultaría su análisis ya que los valores pequeños se verían "opacados" por aquellos valores mucho más grandes; al obtener el logaritmo de la señal transformada, todos los valores tendrán una relevancia suficiente para el análisis.

f) DCT

La transformada coseno discreta es altamente usada en temas de compresión de imágenes y señales, ya que trabaja de tal manera que concentra la mayor cantidad de energía en una pequeña cantidad de coeficientes. Al resultado de esta transformada es lo que llamamos MFCC, el cual se calcula de la siguiente manera:

$$C_n = \sum_{k=1}^{K} (logS_K) cos\left\{n(k - {}^1/_2)\frac{\pi}{K}\right\} \ con\ n = 1, \dots, K-1, \quad (5)$$

donde k es la banda de frecuencias, n es el coeficiente MFCC en cuestión, $logS_k$ es el logaritmo de los coeficientes de Mel y K es el número total de bandas o filtros.

Posteriormente se calculan de la misma manera el resto de las ventanas obtenidas en la etapa de ventaneo, de esta manera se obtendrán los coeficientes de toda la señal.

g) Entrenamiento de la red neuronal

En esta etapa se hace uso de datos obtenidos en la etapa previa de extracción de características MFCC, como se observa en la figura 1; el sistema hace uso de una red neuronal perceptrón multicapa para la clasificación y el algoritmo Backpropagation como algoritmo de entrenamiento.

Se propone emplear el 70% de los patrones de malware de la base de datos para el entrenamiento de la red neuronal, un 15% se utilizará para validación, que servirá para detener el entrenamiento cuando la red esté sobre ajustando los datos y finalmente, un 15% para la etapa de pruebas, tal como se muestra en la Fig. 3.

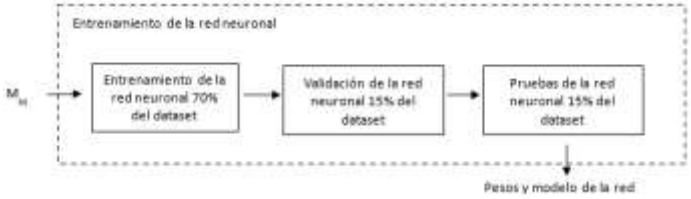

Fig. 3. Diagrama a bloques de entrenamiento de la red neuronal

B. Fase de pruebas

La Figura 4 describe a la red neuronal perceptrón multicapa, entrenada (como se describió en la sección 2.3) y lista para clasificar; donde M_{PM} representa los MFCC obtenidos de una muestra de malware que se desea clasificar.

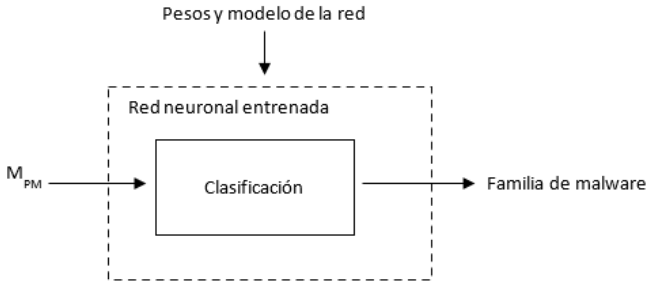

Fig. 4. Diagrama a bloques de la fase de clasificación

Para la etapa de Pruebas, se utilizarán muestras de malware obtenidas de [4], la cual consiste en una base de

datos con 9339 muestras de malware clasificadas en 25 familias, las cuales se muestran en la Tabla 1.

El preprocesamiento para la obtención de la señal de audio digital es aplicado a la base de datos de malware los cuales están codificados como imágenes a escala de grises, para convertirlos en un formato de audio digital definiendo esta salida como M_{PA}, siguiendo el diagrama a bloques general en la extracción de las características MFCC definiendo esa salida como M_{PM}.

TABLA. 1. TABLA MALIMG DATASET

Número	Nombre de la familia	Número de muestras
1	Adalier.C	125
2	Agent.FYI	116
3	Allaple.A	2949
4	Allaple.L	1591
5	Aleuron.gen!j	198
6	Autorun.K	106
7	C2LOP.gen!g	200
8	C2LOP.P	146
9	Dialplatform.B	177
10	Dontovo.A	168
11	Fakerean	381
12	Instantaccesss	431
13	Lolyda.AA1	213
14	Lolyda.AA2	184
15	Lolyda.AA3	123
16	Lolyda.AT	159
17	Malex.gen!j	136
18	Obsfuscador.AD	142
19	Rbot!.gen	158
20	Skintrim.N	80
21	Swizzor.gen!E	132
22	Swizzor.gen!I	128
23	VB.AT	108
24	Wintrin.BX	94
25	Yuner.A	800

III. PRUEBAS Y RESULTADOS

El sistema propuesto fue implementado en el lenguaje de programación Matlab R2020b y las características del equipo son: Procesador Intel® Core i5® 2.9 GHz, memoria RAM de 16 GB y sistema operativo Windows 10 de 64 bits.

Las imágenes se vieron sometidas a un preprocesamiento, que incluye un reescalado a tamaño de 512x512 pixeles, posteriormente se obtuvieron los datos necesarios para representarla en forma de señal de audio digital como se muestra en la Fig. 6, y finalmente el proceso de extracción de características MFCC obtenido como resultado las señales que se muestran en la Fig. 7. Se muestran los resultados obtenidos para dos familias distintas las cuales son la familia Adalier.C y Dontovo.A (familia 1 y familia 10 según; Tabla 1. Tabla Malimg dataset).

Para la evaluación de los resultados de clasificación, se utilizaron las métricas llamadas recall, precisión y F1 score para cada una de las familias. El cálculo de la precisión se obtiene usando (6)

$$Precisión = \frac{TP}{TP+FP}, \tag{6}$$

donde *TP* es la cantidad de verdaderos positivos que existen y *FP* es la cantidad de Falsos positivos, en términos sencillos, este valor indica la cantidad de veces que el modelo acertara al momento de realizar la clasificación de determinada familia.

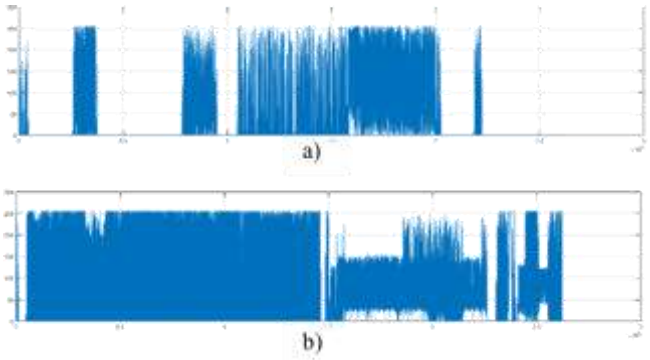

Fig. 6. Malware representados como señales de audio digital. a) Malware de la familia Adalier.C y b) Malware de la familia Dontovo.A.

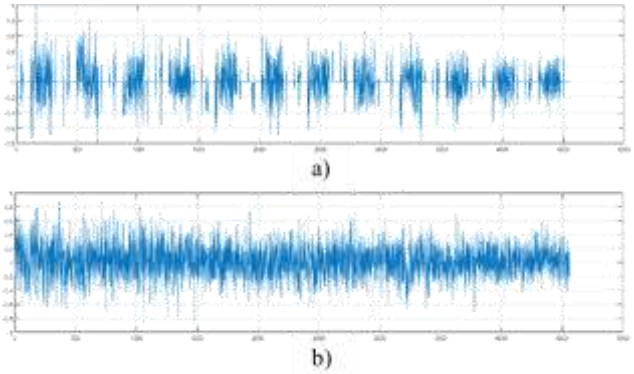

Fig. 7. MFCC obtenidos de las muestras de las familias: a) Adalier.C y b) Dontovo.A

Para obtener el valor de recall (también llamado exhaustividad), se obtiene usando (7)

$$Recall = \frac{TP}{TP+FN}, \tag{7}$$

donde *FN* es la cantidad de Falsos negativos, el valor obtenido de recall se explica como la cantidad de muestras

correctamente clasificadas por el sistema. Finalmente, el valor *F1 Score* es un valor que relaciona los valores anteriores, de esta manera resulta más sencillo interpretar los resultados obtenidos ya que se obtiene un valor entre 0 y 1 donde 1 es una clasificación ideal, este se obtiene mediante (8)

$$F1 = 2 * \frac{precisión * recall}{precisión + recall}, \quad (8)$$

Los resultados obtenidos para las métricas de Recall, Precisión y F1 Score se muestran en la Tabla 2, como se puede observar, se tiene una precisión de entre 90% al 100% en 20 de las 25 familias, en 3 familias (7, 8 y 24) se tiene una precisión de entre 72% a 88%, es decir un poco más bajo que el resto y están los casos de las familias 21 y 22 donde se tiene una precisión de 30% y 20% respectivamente, esta baja precisión se justifica ya que las muestras de estas 2 subfamilias son bastante similares y podrían pertenecer a una misma familia, por lo tanto, las muestras son muy semejantes entre sí.

TABLA 2. EVALUACIÓN DE LA CLASIFICACIÓN.

Familia	Recall	Precisión	F1 Score
1	1.0	1.0	1
2	1.0	0.92	0.95
3	0.99	1.0	0.99
4	1.0	1.0	1
5	0.89	0.94	0.91
6	1.0	1.0	1
7	0.75	0.71	0.72
8	0.75	0.75	0.75
9	1.0	1.0	1
10	1.0	0.96	0.97
11	1.0	0.96	0.97
12	1.0	0.98	0.99
13	1.0	0.96	0.99
14	1.0	1.0	1
15	0.91	1.0	0.95
16	0.95	1.0	0.97
17	1.0	1.0	1
18	1.0	1.0	1
19	0.96	0.96	0.96
20	1.0	1.0	1
21	0.26	0.38	0.30
22	0.25	0.17	0.20
23	1.0	1.0	1
24	0.92	0.85	0.88
25	1.0	1.0	1

IV. CONCLUSIONES

La evidencia presentada demuestra cómo el objetivo del proyecto se cumplió de manera satisfactoria gracias al correcto uso de la información obtenida a partir de una señal de audio y poder utilizar las características propias, tal es el caso de los MFCC´s. De esta manera, al obtener las características esenciales del malware con el sistema propuesto, se evita la necesidad de ejecutar el malware para determinar las acciones que realizaría en el sistema, de igual manera las probabilidades de clasificar correctamente variaciones de estas muestras son altas, gracias a que todas ellas comparten características esenciales. Tomando en consideración los resultados obtenidos en la métrica F1 score, se puede determinar que el sistema tuvo un buen desempeño en la mayoría de las familias, teniendo resultados entre el 0.90 y el 1; sin embargo, el caso de las familias 21 y 22 que tuvieron una baja precisión, se debe a que estas 2 subfamilias son bastante similares y podrían pertenecer a una misma familia, por lo tanto, tienen características similares causando confusión entre ellas para la clasificación.

AGRADECIMIENTOS

Agradecemos al Instituto Politécnico Nacional, a la COFAA del IPN y a la Beca de Estimulo Institucional de Formación de Investigadores (BEIFI) del IPN por el apoyo otorgado para el desarrollo de este trabajo.

REFERENCIAS

[1] J. Regan (2019, Julio 11), What is malware? How malware Works & how to remove it. [Online]. Available: https://www.avg.com/en/signal/what-is-malware

[2] AV-Test (2020, Octubre 21). Software malicioso [Online]. Available: https://www.av-test.org/es/estadisticas/software-malicioso/

[3] Carlos Z, Ivan C., Maria M., (2015). Técnicas de detección y análisis de malware en entornos corporativos con sistemas operativos Windows. Proyecto presentado para optar al título de Especialista en Seguridad Informática. Universidad de San Buenaventura Seccional Medellin, Medellin, Colombia.

[4] García Daza Cervantes, I., Reyes-Reyes, R., Cruz-Ramos, C., Ponomaryov, V., & Ponomaryov, D. (2019, October). Malware classification using distance and directional local binary patterns. In 2019 IEEE International Scientific-Practical Conference Problems of Info communications, Science and Technology (PIC S&T) (pp. 397-401). IEEE.

[5] M. Farrokhmanesh and A. Hamzeh, "A novel method for malware detection using audio signal processing techniques," *2016 Artificial Intelligence and Robotics (IRANOPEN)*, Qazvin, Iran, 2016, pp. 85-91, doi: 10.1109/RIOS.2016.7529495.

[6] L. Nataraj, S. Karthikeyan, G. Jacob, and B. S. Manjunath. (2011). Malware images: visualization and automatic classification. In Proceedings of the 8th International Symposium on Visualization for Cyber Security (VizSec '11). Association for Computing Machinery, New York, NY, USA, Article 4, 1–7. DOI:https://doi.org/10.1145/2016904.2016908

[7] Mel Frequency Cepstral Coefficient (MFCC) tutorial. [Tutorial de Coeficientes Cepstrales en la Frecuencia de mel (MFCC)] http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/

CURRICULUM VITAE

Jorge A. Alcántara A. estudiante de la carrera de Ingeniera en Computación en el Instituto Politécnico Nacional (IPN). Sus áreas de interés son: redes de computadoras y seguridad de la información.

Brandon A. Herrera L. estudiante de la carrera de Ingeniera en Computación en el Instituto Politécnico Nacional (IPN). Becario del Estímulo Institucional de Formación de Investigadores (BEIFI) del IPN. Sus áreas de interés son: Redes neuronales, procesamiento de señales y software malicioso

Clara Cruz R. recibió el título de Ingeniera en Comunicaciones y Electrónica, el grado de Maestra en Ciencias de Ingeniería en Microelectrónica, y el grado de Doctora en Comunicaciones y Electrónica, por el Instituto Politécnico Nacional en 1999, 2003 y 2009 respectivamente, recibió el Diploma a la Excelencia Académica por sus estudios de Licenciatura y el Diploma como Mejor Promedio de Generación a Nivel Posgrado. Actualmente es profesora titular en el Departamento de Ingeniería en Computación de la ESIME Culhuacán en el IPN. Sus áreas de investigación son reconocimiento de patrones, procesamiento de imágenes y marcas de agua.

Rogelio Reyes R. recibió el título de Ingeniero en Comunicaciones y Electrónica por la opción de Escolaridad, el grado de Maestro en Ciencias de Ingeniería en Microelectrónica, y el grado de Doctor en Comunicaciones y Electrónica, de la ESIME Culhuacán del Instituto Politécnico Nacional en 1999, 2003 y 2009 respectivamente; recibió el Diploma a la Excelencia Académica por sus estudios de Licenciatura y la Presea Lázaro Cárdenas del IPN a Nivel Posgrado en el 2003. Actualmente es profesor titular en el Departamento de Ingeniería en Computación de la ESIME Culhuacán. Sus áreas de investigación son: procesamiento de señales y sistemas embebidos.