Dispositivo IoT para prevenir la violencia de género usando TinyML

Mónica T. Avila Rodríguez[1], Elsa M. Quizphe Buñay[1], Wilson G. Chango Sailema[2], Stalin M. Arciniegas[3*]

[1]Pontificia Universidad Católica del Ecuador Sede Esmeraldas
[2]Escuela Superior Politécnica de Chimborazo
[3]Pontifica Universidad Católica del Ecuador Ibarra
*Autor para correspondencia: smarciniegas@pucesi.edu.ec
Recibido: 2023/07/31 Aprobado: 2023/11/27
DOI: https://doi.org/10.26621/ra.v1i29.920

RESUMEN

El estudio se enmarca en el desarrollo de una solución basada en el Internet de las cosas (IoT) y el aprendizaje automático para prevenir y detectar situaciones de peligro relacionadas con la violencia basada en el género (VBG). El objetivo es proporcionar una herramienta útil y accesible para las mujeres en riesgo, contribuyendo así a la prevención y reducción de la VBG. El problema que aborda el estudio es la violencia basada en el género, un tema de gran relevancia social y humanitaria. Se busca utilizar tecnologías digitales y aprendizaje automático para detectar palabras asociadas con situaciones de peligro y prevenir la VBG en tiempo real. Para abordar el problema, se utiliza un *data set* público creado por Microsoft que contiene muestras de audio de diferentes palabras, incluyendo palabras asociadas con situaciones de peligro como "*yes*" y "*no*", así como otras palabras y ruido estático. Se utilizan datos de audio en formato WAV, divididos en ventanas de un segundo con una frecuencia de muestreo de 16000 Hz. Se selecciona una ventana de datos homogénea con una duración de un segundo y se utiliza el coeficiente cepstral de frecuencia (MFCC) para resaltar la voz humana y reducir el ruido de fondo. El modelo desarrollado mostró un buen desempeño en general, con una eficiencia promedio del 91.3 % en el conjunto de entrenamiento y del 85.83 % en el conjunto de evaluación. Se obtuvo una alta precisión en la detección de palabras asociadas con situaciones de peligro, como "yes" y "no". Se reconoce que la tecnología tiene un papel significativo para abordar la VBG, pero también se enfatiza la necesidad de un compromiso de la sociedad y los gobiernos para lograr un cambio duradero y significativo en la erradicación de este problema a nivel mundial.

Palabras clave: aprendizaje automático, reconocimiento automático de voz, computación de borde, Internet de las cosas

ABSTRACT

TThe study is framed within the development of a solution based on the Internet of things (IoT) and machine learning to prevent and detect dangerous situations related to gender based violence (GBV). The goal is to provide a useful and accessible tool for women at risk, thus contributing to the prevention and reduction of GBV. The problem addressed by the study is gender-based violence, an issue of great social and humanitarian relevance. It seeks to use digital technologies and machine learning to detect words associated with dangerous situations and prevent GBV in real time. To address the problem, a public *data set* created by Microsoft containing audio samples of different words, including words associated with dangerous situations such as "*yes*"; and "*no*"; as well as other words and static noise, is used. Audio data in WAV format is used, divided into one-second windows with a sampling rate of 16000 Hz. A homogeneous data window with a duration of one second is selected and the frequency cepstral coefficient (MFCC) is used to highlight the human voice and reduce background noise. The developed model showed good overall performance, with an average efficiency of 91.3 % in the training set and 85.83% in the evaluation set. High accuracy was obtained in the detection of words associated with danger situations, such as "*yes*"; and "*no*". It is recognized that technology has a significant role to play in addressing GBV, but it also emphasizes the need for a commitment from society and governments to achieve lasting and significant change in eradicating this problema worldwide.

Keywords: entrepreneurship, agribusiness, organizational performance, local development, primary sector

Mónica T. Avila Rodríguez orcid.org/0009-0001-9862-3892
Elsa M. Quizphe Buñay orcid.org/0009-0009-1527-04251
Wilson G. Chango Sailema orcid.org/0000-0003-3231-0153
Stalin M. Arciniegas Aguirre orcid.org/0000-0001-9535-6058

INTRODUCCIÓN

La violencia basada en el género (VBG) es un problema social de gran magnitud que afecta los derechos de las mujeres a nivel mundial (Enaifoghe et al., 2021). Las estadísticas revelan que un número alarmante de mujeres ha experimentado violencia física o sexual, y que muchas de ellas no denuncian estos actos debido al temor a sus agresores o a la vergüenza social. La VBG afecta a mujeres de todas las edades y contextos sociales, y su erradicación representa un desafío que requiere acciones coordinadas a nivel global (John et al., 2021).

Para abordar esta difícil situación y garantizar la seguridad y bienestar de las mujeres, es esencial adoptar enfoques integrales y utilizar todas las herramientas disponibles. En este sentido, las nuevas tecnologías digitales y la ciencia de la computación desempeñan un papel fundamental, al ofrecer posibilidades para desarrollar soluciones innovadoras en la lucha contra la VBG.

El problema de investigación se centra en encontrar formas efectivas de prevenir y detectar situaciones de peligro vinculadas a la VBG. Por ello, el objetivo principal de este trabajo es desarrollar una aplicación basada en el Internet de las cosas (IoT, por sus siglas en inglés) y el aprendizaje automático (ML, por sus siglas del inglés) que pueda identificar palabras o sonidos asociados con situaciones de peligro en tiempo real (Mishra y Tyagi, 2022). Esta aplicación busca ser una herramienta útil y accesible para las mujeres en riesgo, que permita así contribuir a la prevención y reducción de la VBG.

Para enfrentar este desafío, se utiliza el enfoque de IoT y aprendizaje automático en el desarrollo de la aplicación. Se emplean dispositivos pequeños con capacidad de cálculo para realizar análisis de borde (edge analytics) en los datos de voz recopilados. Se implementa una red neuronal convolucional (CNN) como modelo de aprendizaje automático para identificar palabras o comandos específicos que puedan denotar situaciones de peligro (Kamalraj et al., 2021).

Así, se espera que la aplicación logre identificar y reconocer con precisión los sonidos y palabras asociados con situaciones de peligro en tiempo real, lo que representa un avance significativo en la prevención de la violencia basada en el género (VBG). Al utilizar un modelo ligero de redes neuronales convolucionales (CNN) y la analítica de borde, la herramienta será capaz de ejecutarse eficientemente en dispositivos de baja potencia, permitiendo una detección más rápida y eficiente de estas situaciones. Esto se traduce en una atención oportuna de la VBG, ya que se podrá responder de manera inmediata ante situaciones de peligro, brindando apoyo y asistencia a las mujeres en riesgo.

La implementación del Internet de las cosas (IoT) en la aplicación desempeña un papel fundamental, al permitir la conexión de dispositivos y sensores, facilitando la recopilación de datos en tiempo real y habilitando una respuesta automatizada. Los sensores, como el micrófono en los dispositivos TinyML, capturan y procesan los datos de audio, los cuales son analizados directamente en el dispositivo a través de la analítica de borde. Esta capacidad de procesamiento local disminuye la necesidad de enviar grandes cantidades de datos a servidores externos, lo que a su vez reduce la latencia y garantiza una mayor privacidad y seguridad de los datos recopilados.

La aplicación inteligente busca ser una herramienta efectiva y accesible para prevenir y atender situaciones de peligro relacionadas con la VBG. Al reconocer patrones de audio asociados con palabras como *"yes" o "no"*, que pueden indicar una situación de riesgo, la aplicación podrá alertar a las personas cercanas o a las autoridades correspondientes, lo que contribuirá a una respuesta más rápida y efectiva ante potenciales peligros.

El artículo se estructura en diferentes secciones, con el fin de presentar de manera sistemática y organizada el desarrollo de la aplicación y sus resultados. Se comienza con una revisión de trabajos relacionados con la prevención de la VBG utilizando tecnologías digitales y aprendizaje automático. Esta revisión proporciona un contexto sólido y una comprensión del estado actual de la investigación en este campo, destacando la relevancia y la novedad de la aplicación propuesta.

Posteriormente, se presenta la metodología empleada para el procesamiento y entrenamiento del modelo de aprendizaje automático. En esta sección se detallan los pasos y técnicas utilizados para transformar los datos de audio en imágenes y se indica cómo se implementó el modelo de CNN. También se describen los hiperparámetros utilizados y la optimización del modelo para adaptarse a los dispositivos de baja potencia.

En la sección de resultados, se exponen los hallazgos obtenidos del entrenamiento del modelo y su discusión. Se muestran las eficiencias alcanzadas en la clasificación de las diferentes clases asociadas con las palabras clave, y se realiza un análisis detallado del rendimiento del modelo en cada una de ellas. Asimismo, se discuten posibles limitaciones y áreas de mejora identificadas durante el proceso.

Finalmente, se presentan las conclusiones derivadas de este trabajo y se discute el potencial impacto de la aplicación en la prevención de la VBG. Se resaltan los logros obtenidos y la relevancia de utilizar tecnologías digitales y aprendizaje automático en la lucha contra la violencia de género. Además, se enfatiza la importancia de seguir trabajando en la mejora y optimización del modelo para aumentar su precisión y eficacia en la detección de situaciones de peligro.

Estado del arte

En la literatura, existen diversos trabajos sobre el reconocimiento automático de la voz (ASR, por sus siglas en inglés) y el reconocimiento de las emociones del habla (SER, por sus siglas en inglés). El ASR se enfoca en detectar y comprender las palabras que una persona emite durante su discurso, mientras que los sistemas SER se dedican a identificar el estado emocional en el que se encuentra una persona al hablar (Mrozek et al., 2021).

El objetivo del sistema de aprendizaje consiste en encontrar la etiqueta cuyo vector de entrada tenga la mayor probabilidad de corresponder al sonido

$$L = arg \max_{l \in V} P(X|L) \quad (1)$$

Existen diferentes algoritmos para el reconocimiento de voz, como el modelo oculto de Markov (HMM), el modelo mixto gaussiano (GMM), el modelo de red neuronal profunda (DNN) para HMM y el modelo *end-to-end* (E2E). Los modelos híbridos basados en DNN han mejorado la evaluación de la probabilidad acústica, y, en la actualidad, se observa una tendencia hacia el desarrollo de modelos *end-to-end* (Lo et al., 2020).

Bahar et al. (2019) utilizan una red neuronal recurrente (RNN) para realizar la predicción de secuencias a nivel de caracteres. La alineación entre las entradas y la secuencia de caracteres deseada se aprende mediante un mecanismo de atención integrado en la RNN.

Por otro lado, Chang et al. (2021) emplean varios enfoques de aprendizaje permanente (*LifeLong learning* o LLL, por sus siglas en inglés) en ASR utilizando un modelo *end-to-end* (E2E). Además, proponen métodos para almacenar datos de dominios anteriores con el fin de reducir el problema de olvido y mantener el aprendizaje continuo a medida que se incorporan nuevos datos.

Chang et al. (2020) utilizan un transformador, un modelo de aprendizaje profundo con mecanismo de autoatención, para desarrollar un sistema ASR multicanal que recolecta información espacial y espectral mediante la integración de diferentes micrófonos. Para lograrlo, emplean varias capas de atención. En su trabajo, logran reducir la latencia del proceso de transmisión en los sistemas ASR al obtener respuestas tempranas basadas en el reconocimiento parcial (*prefetching*). Si el resultado parcial coincide con el resultado del reconocimiento final, la respuesta obtenida anticipadamente se entrega al usuario, reduciendo la latencia que ocurre después de completar el reconocimiento. Para ello, utilizan una RNN-T (transductor de red neuronal recurrente) y una red neuronal o LAS (*Listen Attend Spell)*.

Una característica de los sistemas ASR y los sistemas SER es que las señales acústicas pueden variar cuando el hablante se encuentra en situaciones de estrés (Zhang et al., 2021). Para abordar este problema durante la fase de entrenamiento, se utilizan técnicas de aumento de datos, específicamente bajo condiciones de estrés, ya que a menudo hay una falta de datos para esta condición.

Rituerto-González et al. (2019) llevan a cabo un aumento de muestras utilizando métodos estadísticos y generando datos de habla de forma sintética bajo condiciones de estrés. Para equilibrar la complejidad computacional y la precisión, consideran crítico el consumo de batería y optan por utilizar una red neuronal de multi-capas perceptron (MLP, por sus siglas en inglés) con una sola capa oculta, debido a su sencillez, rapidez y buen desempeño.

En el trabajo de Wang et al. (2020) se utilizan modelos generativos profundos para generar sonidos más diversos durante el entrenamiento de un análisis discriminante lineal probabilístico (PLDA, por sus siglas en inglés). Esto permite obtener datos más variados y enriquecidos, mejorando así la capacidad de reconocimiento en condiciones de estrés.

Reconocer emociones a partir del habla, como el miedo o el estrés, causadas por situaciones de VBG, ha sido objeto de estudio con el fin de desarrollar dispositivos inteligentes capaces de detectar posibles casos de esta violencia. Por ejemplo, Binni es un sistema autónomo que combina el aprendizaje automático y el Internet de las cosas (IoT) para identificar y reportar automáticamente situaciones de riesgo de VBG (Miranda et al., 2022).

Por su parte, Zhang et al. (2018) abordan el problema de la brecha existente entre la naturaleza subjetiva de las emociones y el bajo nivel de características disponibles para identificarlas. Proponen un enfoque en capas que utiliza una red neuronal convolucional profunda (DCNN) para el aprendizaje automático de características, utilizando como entrada la salida del espectrograma. Posteriormente, proponen un algoritmo de coincidencia de pirámide temporal discriminante (DTPM) que recibe como entrada las características aprendidas por las DCNN. Este algoritmo permite agrupar funciones para el reconocimiento de emociones del habla; finalmente, se utiliza una máquina de soporte vectorial (SVM) para identificar la emoción. Estos enfoques requieren cómputos pesados, por lo que su ejecución en un entorno de analítica de borde es actualmente imposible (Guo, 2022).

Por esta razón, en el presente trabajo se utilizó un modelo ligero de redes neuronales convolucionales (CNN) que permite realizar la analítica de borde, lo que resulta en una solución óptima para la implementación y ejecución del reconocimiento de emociones en dispositivos con recursos computacionales limitados.

La elección de un modelo ligero de CNN es fundamental para abordar el desafío de llevar a cabo el reconocimiento de emociones en dispositivos de baja potencia y recursos limitados, como los dispositivos TinyML mencionados anteriormente (Arduino Nano 33 BLE Sense, Seeed Studio XIAO nRF52840 Sense, Seeed Studio XIAO ESP32S3 Sense). Estos dispositivos cuentan con restricciones significativas en términos de memoria y capacidad de procesamiento, por lo que el uso de modelos complejos es inviable.

Al emplear un modelo ligero de CNN, se logra una solución altamente eficiente y eficaz, ya que este tipo de modelos están diseñados para reducir la cantidad de parámetros y capas, lo que a su vez disminuye considerablemente el consumo de recursos. Esto permite que el modelo pueda ejecutarse de manera óptima en dispositivos con capacidades computacionales limitadas, sin sacrificar significativamente su rendimiento.

La analítica de borde se refiere al procesamiento y análisis de datos directamente en el dispositivo o en la periferia, en lugar de enviar los datos a un servidor o nube para su procesamiento. Al realizar la analítica de borde, se reduce la carga en la red y se minimiza la latencia; así, se consigue una respuesta más rápida y una mayor privacidad y seguridad de los datos, ya que estos no se transmiten a servidores externos.

En consecuencia, la combinación de un modelo ligero de CNN con la analítica de borde se convierte en una solución altamente efectiva para llevar a cabo el reconocimiento de emociones en tiempo real en dispositivos de baja potencia y recursos limitados. Esto es especialmente relevante para su aplicación en la detección de situaciones de peligro asociadas con la violencia basada en el género (VBG), ya que permite que la herramienta sea accesible y útil para las mujeres en riesgo, brindando una respuesta inmediata.

Así, esta combinación de tecnologías digitales y aprendizaje automático representa una solución prometedora para abordar problemas sociales complejos y mejorar la seguridad y bienestar de las personas vulnerables.

MÉTODOS

El despliegue de nuestra solución se detalla en el flujo de la Figura 1, el cual es típico en proyectos de ML; sin embargo, se añadió un paso adicional al final, el despliegue del modelo en un dispositivo de bajo consumo. El modelo se optimiza en términos de consumo de memoria para que pueda ejecutarse dentro del dispositivo, dado que éeste cuenta con recursos limitados.

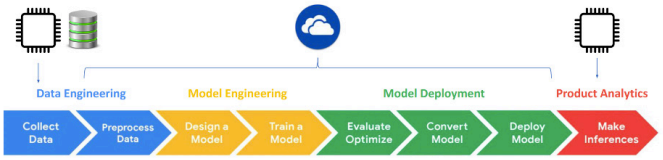

Figura 1. Diagrama de flujo de proyectos de TinyML, incluido el despliegue en dispositivos para la optimización de la memoria y consumo de baja energía.

Captura de datos

El proyecto se inicia con la selección y captura de datos. Inicialmente, sTe propusieron un conjunto de palabras para identificar una situación de violencia; se estableció que las palabras tenían que ser monosílabas, ya que su registro analógico contiene un solo sonido y, por tanto, son más sencillas de identificar. Además, era deseable que las palabras no formaran parte del vocabulario común (en español) de las personas. Por ello, se decidió, para esta primera versión del dispositivo, que la palabra *"yes"* sirviera de palabra clave para indicar una situación de violencia.

Inicialmente, se utilizó un *data set* público creado por Microsoft (Edge Impulse, s.f). Este *data set* contiene 25 minutos de datos por clase en formato WAV, divididos en ventanas de un segundo con un muestreo de 16000 Hz. El conjunto de datos incluye lo siguiente:

- Yes: muestras de un segundo con la palabra *"yes".*
- No: muestras de un segundo con la palabra *"no".*
- Unknown: muestras de un segundo de otras palabras.
- Noise: muestras de un segundo de ruido estático.

Los datos, posteriormente, se aumentaron con 10 minutos adicionales de grabación; estos nuevos datos registraban el sonido de la palabra *"yes"* pronunciada por personas que simulaban una situación de peligro. Este aumento perseguía que el entrenamiento identificara con mayor precisión la palabra *"yes"* que efectivamente describiera una situación de violencia.

Preprocesamiento

En la segunda parte, durante el procesamiento de datos, se selecciona una ventana de datos homogénea con una duración de un segundo para todos ellos. Además, se recomienda una frecuencia de muestreo de 16000 Hz para los audios de forma predeterminada. A continuación, se realiza un preprocesamiento utilizando el coeficiente cepstral de frecuencia, más conocido como MFCC, por sus siglas en inglés. Este preprocesamiento es apropiado para resaltar la voz humana y reducir al mínimo el ruido que pueda estar presente en el entorno. En la Figura 2 se muestra un ejemplo de la palabra *"yes"* en una representación temporal de un segundo.

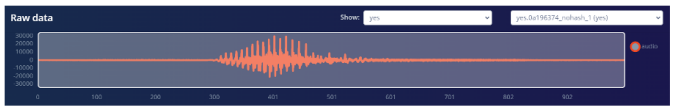

Figura 2. Captura de audio de la palabra "*yes*" en ventana de un segundo. Procesamiento para resaltar la voz humana y reducir el ruido.

A continuación, el audio se transforma en una imagen, utilizando tratamiento digital de señales (DSP). En la Figura 3 se puede observar dicha transformación. Debido a que se utilizan pocos colores y una baja resolución temporal, el tiempo de entrenamiento es considerablemente menor en comparación con el reconocimiento de imágenes. Asimismo, el algoritmo se beneficia de la reducción de la cantidad de entradas para la red neuronal.

En total, se generan 50 slots en el eje de las x y 13 coeficientes en el eje de la y. Como resultado de esta transformación, la imagen tiene una resolución de 40x13, lo que da un total de 650 características (features). Estas características serán las entradas de nuestra red neuronal.

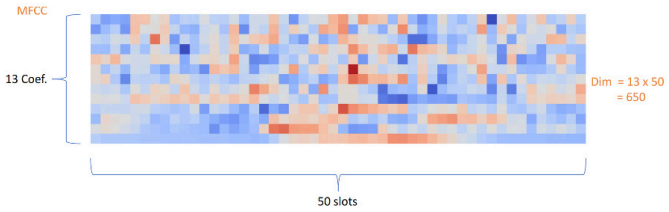

Figura 3. Coeficiente cepstral de frecuencia (MFCC) de la palabra *"yes"*. Transmisión de la imagen según sus características de entrada.

Diseño y entrenamiento del modelo

Cada una de las 8408 instancias es una imagen unidimensional que se crea a partir de una ventana de un segundo, generando 650 características. Estas instancias se dividen en datos de entrenamiento y datos de prueba, en una proporción de 80/20, lo que resulta en 7207 datos de entrenamiento y 3603 datos de prueba. Cada instancia tiene 4 salidas, representadas por el siguiente vector.

Vector_clase = ["yes", "no", "noise", "unknnown,"]

Los datos de las imágenes fueron procesados utilizando una red neuronal convolucional (CNN). Se seleccionó este tipo de red porque son las redes más utilizadas para trabajos de clasificación y procesamiento de imágenes . Las red se diseñó con una primera capa con 8 neuronas y una segunda capa con 16 neuronas, ambas activadas por la función de activación ReLU (Oh et al., 2021). Finalmente, tenemos una capa final de 4 neuronas correspondientes al vector de salida, las cuales están activadas por la función de activación Softmax.

Para el entrenamiento se utilizó Edge Impulse, una plataforma de desarrollo de aplicaciones TinyML que simplifica la creación de aplicaciones embebidas; ofrece herramientas de alto nivel (utiliza TensorFlow) para todas las fases de desarrollo, y para el entrenamiento del modelo no se requiere escribir código. Los modelos generados se pueden convertir a Java o C++, con el objeto de poder integrarlos en cualquier plataforma. Es importante mencionar que esta plataforma ofrece una opción de autotuner (EON-Tuner), la cual brinda soporte para definir los hiperparámetros de la red neuronal.

Para el entrenamiento del modelo, se utilizaron los siguientes hiperparámetros generados por Edge Impulse:

- Número de ciclos de entrenamiento = 100
- Tasa de aprendizaje = 0.005
- Tamaño de lote = 32
- Porcentaje de validación = 20 %

RESULTADOS Y DISCUSIÓN

Evaluación del modelo

Después de realizar la evaluación de los datos, se obtuvo una eficiencia promedio del 91.3 % para las 4 salidas, con una pérdida del 0.29 %. Los resultados se detallan en la Tabla 1.

Tabla 1. Eficiencia del modelo

	Entrenamiento		Evaluación	
Evaluación / Métricas	Exactitud	Puntaje	Exactitud	Puntaje
Yes	93.40 %	0.94	89 %	0.93
No	91.5 %	0.93	85.1 %	0.88
Noise	91.3 %	0.93	92.4 %	0.94
unknown	89.1 %	0.86	76.7 %	0.85
Global	91.3 %		85.83 %	

El modelo mostró un buen desempeño en general, con una eficiencia promedio del 91.3 % en el conjunto de entrenamiento y una eficiencia del 85.83 % en el conjunto de evaluación.

Análisis y discusión por clase

- Clase "Yes": El modelo logró una precisión del 93.4 % en el conjunto de entrenamiento y del 89 % en el conjunto de evaluación. El puntaje (*score*) también es alto, indicando un buen rendimiento para esta clase.

- Clase "No": El modelo tuvo una precisión del 91.5 % en el conjunto de entrenamiento y del 85.1 % en el conjunto de evaluación. Aunque un poco menor que la clase "Yes", sigue siendo un rendimiento aceptable.

- Clase "Noise": El modelo obtuvo una precisión del 91.3 % en el conjunto de entrenamiento y del 92.4 % en el conjunto de evaluación. Es la clase con mayor precisión en el conjunto de evaluación.

- Clase "Unknown": El rendimiento en esta clase es menor, con una precisión del 89.1 % en el conjunto de entrenamiento y del 76.7 % en el conjunto de evaluación. Es la clase con menor precisión en el conjunto de evaluación.

En resumen, el modelo logró resultados prometedores, pero se observa una ligera degradación del rendimiento en el conjunto de evaluación, especialmente para la clase "Unknown". Sería importante considerar ajustes y optimizaciones para mejorar el rendimiento general del modelo y abordar la clase con menor precisión.

Solución del *hardware*

Para el prototipo, se desea construir un sistema embebido compuesto por un *hardware* que contenga un microprocesador, memoria y un microfono. Además, debe almacenar y ejecutar el *software* que contiene el modelo e identifica las palabras. Por ello, para el despliegue del modelo, se consideró que los dispositivos TinyML debían tener incluido el micrófono. En la Tabla 2 se encuentran las características de los posibles dispositivos a utilizar en este trabajo, que fueron los siguientes: Arduino Nano 33 BLE Sense, Seeed Studio XIAO nRF52840 Sense y Seeed Studio XIAO ESP32S3 Sense.

Tabla 2. Características técnicas de dispositivos TinyML

Nombre	MCU	Micrófono	Memoria	Velocidad de reloj	Precio
Arduino nano 33 BLE sense TinyML kit	Cortex M0+	MP34DT05	1 MB	64 MHz	USD 100
Seeed Studio XIAO nRF52840 Sense	Cortex M4	MSM261D-3526H1CPM	2 MB	64 Mhz	USD 30
Seeed Studio XIAO ESP32S3 Sense	Cortex-M4F	MSM261D-3526H1CPM	8 MB	240 Mhz	USD 30

El Arduino Nano 33 BLE Sense TinyML Kit es el dispositivo más costoso en comparación con los otros dos, con un precio de USD 100. Sin embargo, cuenta con una memoria de 1 MB y una velocidad de reloj de 64 MHz, lo que lo convierte en una opción sólida para aplicaciones que requieren mayores capacidades de procesamiento y almacenamiento.

El Seeed Studio XIAO nRF52840 Sense es el dispositivo más económico, con un precio de USD 30. Aunque tiene una memoria de 2 MB y una velocidad de reloj de 64 MHz, es importante tener en cuenta que su MCU es Cortex M4, que es una versión más avanzada que el Cortex M0+ del Arduino Nano 33 BLE Sense. Esto puede significar un mejor rendimiento en ciertas aplicaciones.

El Seeed Studio XIAO ESP32S3 Sense tiene una memoria más amplia, de 8 MB, y una velocidad de reloj de 240 MHz, lo que indica que puede ser más potente y rápido en comparación con los otros dos dispositivos. Aunque tiene un precio similar al XIAO nRF52840 Sense, su MCU es Cortex-M4F, lo que también puede ofrecer ventajas en términos de rendimiento.

En resumen, la elección entre estos dispositivos se basó en las necesidades específicas del proyecto y el presupuesto disponible. Si se desea mayor potencia de procesamiento y almacenamiento, el Arduino Nano 33 BLE Sense TinyML Kit podría ser la opción adecuada, pero si se busca una opción más económica, el Seeed Studio XIAO nRF52840 Sense y el Seeed Studio XIAO ESP32S3 Sense ofrecen características interesantes a un precio más asequible (ver Tabla 3).

Tabla 3. Modelos de optimización y métricas.

Modelo de optimización	RAM [Kb]	ROM [Kb]	Latencia [ms] Nano	Latencia [ms] nRF52840	Latencia [ms] ESP32S3 Sense
Sin optimización	53.2	116.7	626	457	334
Quantized (int 8)	6	47	452	334	71
Quantized + EON	3.7	27	452	334	71

En la Tabla 3 se presentan diferentes modelos de optimización junto con las métricas de RAM utilizada, ROM (memoria de solo lectura) y latencia en milisegundos para tres dispositivos: Nano, nRF52840 y ESP32S3 Sense.

Análisis y discusión por modelo

Sin optimización:
• Para el dispositivo Nano, se utilizan aproximadamente 53.2 Kb de RAM y 116.7 Kb de ROM.
• Para el dispositivo nRF52840, se utilizan aproximadamente 626 Kb de RAM y 457 Kb de ROM.
• Para el dispositivo ESP32S3 Sense, se utilizan aproximadamente 334 Kb de RAM y 71 Kb de ROM.

Optimización Quantized (int 8):
• La optimización Quantized (int 8) reduce significativamente la cantidad de RAM y ROM utilizada en todos los dispositivos.
• Para el dispositivo Nano, la RAM se reduce a 6 Kb y la ROM a 47 Kb.
• Para el dispositivo nRF52840, la RAM se reduce a 452 Kb y la ROM se reduce a 334 Kb.
• Para el dispositivo ESP32S3 Sense, la RAM se reduce a 71 Kb y la ROM se mantiene en 334 Kb.

Optimización Quantized + EON:
• La optimización Quantized + EON proporciona una mayor reducción en la cantidad de RAM utilizada para todos los dispositivos.
• Para el dispositivo Nano, la RAM se reduce a 3.7 Kb y la ROM se mantiene en 27 Kb.
• Para el dispositivo nRF52840, la RAM se mantiene en 452 Kb y la ROM se mantiene en 334 Kb.
• Para el dispositivo ESP32S3 Sense, la RAM se mantiene en 71 Kb y la ROM se mantiene en 334 Kb.

La optimización Quantized (int 8) reduce la cantidad de bits utilizados para representar los valores numéricos en el modelo, lo que conduce a una disminución en la cantidad de memoria necesaria para almacenar los datos. Esto resulta en una reducción significativa de la RAM utilizada en los tres dispositivos.

La optimización Quantized + EON (probablemente EON-Tuner, mencionado anteriormente) mejora aún más la eficiencia, al reducir más la cantidad de RAM utilizada en los modelos. Sin embargo, esta optimización no parece afectar la ROM utilizada en los dispositivos, ya que los valores se mantienen constantes en comparación con la optimización anterior.

En general, estas optimizaciones son valiosas para reducir la carga en la memoria de los dispositivos, lo que permite ejecutar modelos de aprendizaje automático más eficientes en términos de recursos. Este aspecto es especialmente importante en dispositivos de baja potencia y recursos limitados, como los mencionados (Nano, nRF52840 y ESP32S3 Sense).

Después de evaluar los resultados, se seleccionó el dispositivo Seeed Studio XIAO nRF52840, dado que tenía las características que se requerían para soportar el modelo y su precio era menor. La Figura 4 muestra dos imágenes del dispositivo IoT que finalmente se construyó.

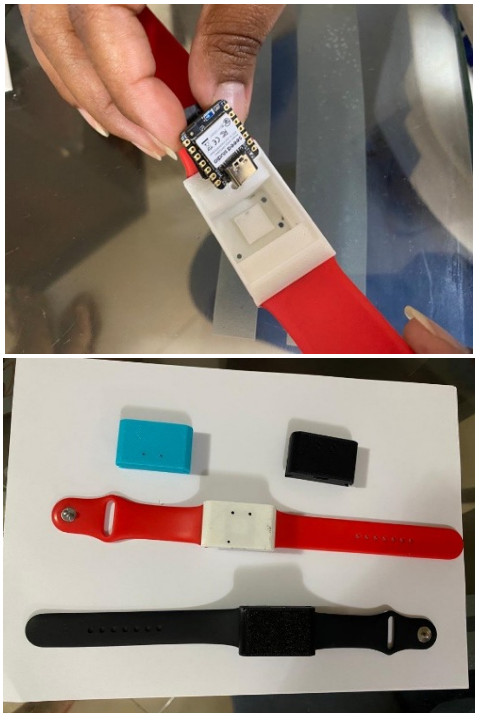

Figura 4. Imágenes del dispositivo IoT construido en este trabajo.

CONCLUSIONES

El objetivo principal de este estudio era desarrollar una aplicación basada en el Internet de las cosas (IoT) y el aprendizaje automático para prevenir y detectar situaciones de peligro relacionadas con la violencia basada en el género (VBG). Así, se buscaba proporcionar una herramienta útil y accesible para las mujeres en riesgo, con el fin de contribuir a la prevención y reducción de la VBG.

El modelo de aprendizaje automático desarrollado mostró un buen desempeño en general, con una eficiencia promedio del 91.3 % en el conjunto de entrenamiento y del 85.83 % en el conjunto de evaluación. Se logró una alta precisión en la detección de palabras asociadas con situaciones de peligro, como *"yes"* y *"no"*. Sin embargo, se observó una ligera degradación del rendimiento en el conjunto de evaluación para la clase "Unknown". A pesar de esto, los resultados fueron prometedores y mostraron la viabilidad de utilizar el Internet de las cosas y el aprendizaje automático para abordar el problema de la VBG.

Los hallazgos indican que la aplicación desarrollada puede ser una herramienta útil para identificar situaciones de peligro en tiempo real. Sin embargo, es importante considerar ajustes y optimizaciones para mejorar el rendimiento general del modelo, especialmente en la detección de situaciones desconocidas. Se destacó la importancia de utilizar tecnologías digitales y el aprendizaje automático para abordar problemas sociales complejos como la VBG.

Una limitación importante del estudio es la ligera degradación del rendimiento en la detección de situaciones desconocidas. Por tanto, es necesario realizar ajustes en el modelo para mejorar su precisión en estas situaciones. Además, se utilizó un conjunto de datos público creado por Microsoft, lo que podría afectar la generalización del modelo a diferentes contextos y culturas.

La aplicación desarrollada puede ser una herramienta valiosa para ayudar a mujeres en riesgo y contribuir a la prevención de la VBG. Se recomienda continuar optimizando el modelo y realizar pruebas en diferentes contextos para mejorar su eficacia y precisión. Además, es importante considerar la privacidad y seguridad de los datos recopilados por la aplicación para garantizar su uso ético y responsable.

El estudio demuestra el potencial de las nuevas tecnologías digitales y del aprendizaje automático para abordar problemas sociales como la VBG. A través de la aplicación desarrollada, se busca brindar una herramienta que pueda marcar una diferencia en la vida de mujeres en riesgo. Sin embargo, se reconoce que es solo un paso en la lucha contra la VBG, y se alienta a seguir trabajando en soluciones integrales y coordinadas para erradicar esta problemática a nivel mundial. La tecnología tiene un papel significativo en este proceso, pero también se necesita el compromiso de la sociedad y de los gobiernos para lograr un cambio duradero y significativo.

Agradecimientos: Quiero expresar mi sincero agradecimiento a la Pontificia Universidad Católica del Ecuador Sede Esmeraldas por el invaluable respaldo brindado durante la elaboración de nuestro trabajo de investigación. El apoyo constante de los profesores, la disponibilidad de recursos y el ambiente académico estimulante fueron fundamentales para alcanzar el éxito en este importante proyecto. Así como a la Dra. Dulce Rivero docente de la PUCE Ibarra por su apoyo y orientación.

Contribución de los autores: "Conceptualización, Mónica T. Avila Rodríguez; metodología, Elsa M. Quizphe Buñay; validación Stalin M. Arciniegas Aguirre; análisis formal, Wilson G. Chango Sailema; investigación, Mónica T. Avila Rodríguez Elsa M. Quizphe Buñay; Stalin M. Arciniegas Aguirre; redacción, revisión y edición Stalin M. Arciniegas Aguirre

Fuente de financiamiento: fondos propios

Conflicto de intereses: "Los autores declaran no tener ningún conflicto de intereses"

REFERENCIAS

Bahar, P., Zeyer, A., Schluter, R. y Ney, H. (2019). On Using 2D Sequence-to-sequence Models for Speech Recognition. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2019-May,* 5671–5675. https://doi.org/10.1109/ICASSP.2019.8682155

Chang, H. J., Lee, H. Y. y Lee, L. S. (2021). Towards lifelong learning of end-to-end ASR. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2,* 1306–1310. https://doi.org/10.21437/INTERSPEECH.2021-563

Chang, S., Li, B., Rybach, D. J., Li, W., He, Y., Sainath, T. N. y Strohman, T. D. (2020). *Low Latency Speech Recognition using End-to-End Prefetching.* https://research.google/pubs/pub49622/

Edge Impulse. (s.f.). *Keyword spotting.* https://docs.edgeimpulse.com/docs/pre-built-datasets/keyword-spotting

Enaifoghe, A., Dlelana, M., Abosede Durokifa, A. y P. Dlamini, N. (2021). The Prevalence of Gender-Based Violence against Women in South Africa : A Call for Action. *African Journal of Gender, Society and Development (Formerly Journal of Gender, Information and Development in Africa), 10*(1), 117–146. https://doi.org/10.31920/2634-3622/2021/V10N1A6

Guo, J. (2022). Deep learning approach to text analysis for human emotion detection from big data. *Journal of Intelligent Systems, 31*(1), 113–126. https://doi.org/10.1515/JISYS-2022-0001/MACHINEREADABLECITATION/RIS

John, N., Roy, C., Mwangi, M., Raval, N. y McGovern, T. (2021). COVID-19 and gender-based violence (GBV): hard-to-reach women and girls, services, and programmes in Kenya. *Https://Doi.Org/10.1080/13552074.2021.1885219, 29*(1), 55–71. https://doi.org/10.1080/13552074.2021.1885219

Kamalraj, R., Neelakandan, S., Ranjith Kumar, M., Chandra Shekhar Rao, V., Anand, R. y Singh, H. (2021). Interpretable filter based convolutional neural network (IF-CNN) for glucose prediction and classification using PD-SS algorithm. *Measurement, 183,* 109804. https://doi.org/10.1016/J.MEASUREMENT.2021.109804

Lo, T. H., Weng, S. Y., Chang, H. J. y Chen, B. (2020). An Effective End-to-End Modeling Approach for Mispronunciation Detection. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2020-October,* 3027–3031. https://doi.org/10.21437/Interspeech.2020-1605

Miranda Calero, J. A., Rituerto-González, E., Luis-Mingueza, C., Canabal, M. F., Barcenas, A. R., Lanza-Gutierrez, J. M., Peláez-Moreno, C. y López-Ongil, C. (2022). Bindi: Affective Internet of Things to Combat Gender-Based Violence. *IEEE Internet of Things Journal, 9*(21), 21174–21193. https://doi.org/10.1109/JIOT.2022.3177256

Mishra, S. y Tyagi, A. K. (2022). The Role of Machine Learning Techniques in Internet of Things-Based Cloud Applications. *Internet of Things,* 105–135. https://doi.org/10.1007/978-3-030-87059-1_4/COVER

Mrozek, D., Kwaśnicki, S., Sunderam, V., Małysiak-Mrozek, B., Tokarz, K. y Kozielski, S. (2021). Comparison of Speech Recognition and Natural Language Understanding Frameworks for Detection of Dangers with Smart Wearables. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 12745 LNCS,* 471–484. https://doi.org/10.1007/978-3-030-77970-2_36/COVER

Oh, S., Shi, Y., del Valle, J., Salev, P., Lu, Y., Huang, Z., Kalcheim, Y., Schuller, I. K. y Kuzum, D. (2021). Energy-efficient Mott activation neuron for full-hardware implementation of neural networks. *Nature Nanotechnology 2021 16:6, 16*(6), 680–687. https://doi.org/10.1038/s41565-021-00874-8

Rituerto-González, E., Mínguez-Sánchez, A., Gallardo-Antolín, A. y Peláez-Moreno, C. (2019). Data Augmentation for Speaker Identification under Stress Conditions to Combat Gender-Based Violence. *Applied Sciences 2019, Vol. 9, Page 2298, 9*(11), 2298. https://doi.org/10.3390/APP9112298

Wang, D., Wang, X. y Lv, S. (2019). An Overview of End-to-End Automatic Speech Recognition. *Symmetry 2019, Vol. 11, Page 1018, 11*(8), 1018. https://doi.org/10.3390/SYM11081018

Wang, S., Yang, Y., Wu, Z., Qian, Y. y Yu, K. (2020). Data Augmentation Using Deep Generative Models for Embedding Based Speaker Recognition. *IEEE/ACM Transactions on Audio Speech and Language Processing, 28*, 2598–2609. https://doi.org/10.1109/TASLP.2020.3016498

Zhang, Q., Wang, D., Zhao, R., Yu, Y. y Shen, J. (2021). Sensing to Hear. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 5*(3). https://doi.org/10.1145/3478093

Zhang, S., Zhang, S., Huang, T. y Gao, W. (2018). Speech Emotion Recognition Using Deep Convolutional Neural Network and Discriminant Temporal Pyramid Matching. *IEEE Transactions on Multimedia, 20*(6), 1576–1590. https://doi.org/10.1109/TMM.2017.2766843