ISSN 1870-4069

Una evaluación comparativa de modelos de Deep Learning para el reconocimiento de emociones a partir del habla

Luis Bernal, Álvaro Cuno, Wilber Ramos Lovón

Universidad Nacional de San Agustín de Arequipa,
Departamento Académico de Ingeniería de Sistemas,
Perú

{lbernal, acunopa, wramos}@unsa.edu.pe

Resumen. Diversos modelos de reconocimiento de emociones a partir del habla vienen siendo propuestos en los últimos años. Sin embargo, las evaluaciones de desempeño de algunas propuestas podrían no ser lo suficientemente confiables. Este trabajo tiene como finalidad contribuir con abordar esta problemática, presentando una manera práctica de implementar evaluaciones comparativas cuyos resultados disten de ser atribuidos a la casualidad. La propuesta utiliza las pruebas de significancia estadística no paramétricas en base al test de Wilcoxon para la comparación del desempeño de modelos de *Deep Learning*. Se demostró la utilidad de la propuesta, realizando las comparación del desempeño de cinco modelos convolucionales, dos entrenados con la base de datos RAVDESS y tres entrenados con la base de datos IEMOCAP.

Palabras clave: Emociones, evaluación, habla.

A Comparative Evaluation of Deep Learning Models for Emotion Recognition from Speech

Abstract. Various models of emotion recognition from speech have been proposed in recent years. However, performance evaluations of some proposals may not be reliable enough. The purpose of this work is to contribute to solve this problem, presenting a practical way to implement comparative evaluations whose results are far from being attributed to chance. The proposal uses non-parametric statistical significance tests based on the Wilcoxon test to compare the performance of *Deep Learning* models. The usefulness of the proposal was demonstrated, comparing the performance of five convolutional models, two trained with the RAVDESS database and three trained with the IEMOCAP database.

Keywords: Emotions, evaluation, speech.

1. Introducción

El interés en lograr reconocer las emociones humanas de manera automatizada ha sido motivado por diversas áreas de aplicación. Por ejemplo, para diseñar robots inteligentes que puedan interactuar de manera empática con las personas, para crear anuncios personalizados que tengan en cuenta el estado emocional de potenciales clientes, para mejorar los procesos de enseñanza/aprendizaje, para mostrar contenido más adecuado a una audiencia objetivo, entre otras aplicaciones [4].

Existen varios métodos para medir las emociones humanas, cada una con sus ventajas y desventajas. Para el caso de los sistemas automatizados, podemos clasificarlos en invasivos y no invasivos. Entre los invasivos, tenemos aquellos que utilizan dispositivos pegados al cuerpo para medir alguna señal fisiológica como electrocardiogramas, electroencefalogramas, respuestas galvánicas de la piel, temperatura, respiración, entre otros.

Entre los no invasivos podemos encontrar a los que utilizan como señal de entrada expresiones faciales, gestos, posturas, voz, entre otras. Un sistema de reconocimiento de emociones a partir del habla busca inferir, de manera automatizada, cual es la emoción que una persona está expresando al momento de hablar. Esto se realiza tomando como entrada las ondas de sonido que se producen al hablar y que son capturadas por uno o varios micrófonos.

Tradicionalmente, este problema había sido abordado mediante la selección manual de características y haciendo uso de algoritmos de clasificación basados en aprendizaje de máquina(*machine learning*). Sin embargo, tras el notable incremento del poder computacional en los últimos años, las técnicas basadas en aprendizaje profundo(*deep learning*), que extraen características de manera automática, han superado en precisión a estos métodos tradicionales [1].

Debido a esto, en la actualidad el estado del arte se soporta en modelos de aprendizaje profundo. Para poder comparar el desempeño de los modelos de manera objetiva una o varias métricas de evaluación deben ser utilizadas. Esta tarea está lejos de ser trivial, principalmente, debido a la presencia de algunos procedimientos de carácter aleatorio en los modelos (p.ej. inicialización de pesos, particionamiento de las bases de datos, entre otros).

Debido a esta cuota de aleatoriedad, los resultados de las evaluaciones suelen ser no reproducibles y pueden variar entre una u otra implementación o incluso entre una u otra ejecución. La complejidad es mayor cuando se trata de comparar una nueva propuesta con modelos existentes en el estado del arte. Esto se debe, principalmente, a que los modelos han sido configurados para obtener los mejores resultados, utilizando bases de datos, formas de entrenamiento, métricas y evaluaciones, de carácter específico y particular a un modelo.

Si se cambia el particionamiento de la base de datos o la manera de inicializar los pesos de los modelos, los resultados podrían variar. A esto se suma la indisponibilidad de código fuente de los modelos, que dificulta replicar las evaluaciones y hacer comparaciones justas [6]. Una implementación o configuración errada de modelos propuestos por terceros podría sesgar los resultados de las comparaciones de manera voluntaria o involuntaria.

Una alternativa que se suele utilizar para facilitar las comparaciones entre modelos es la realización de pruebas comparativas (*benchmarks*), que consisten en distribuir una base de datos de manera pública y desarrollar competiciones por alcanzar la mejor predicción posible. Sin embargo, esto también puede llevar a conclusiones erróneas. Se ha visto casos donde es posible vulnerar los *benchmarks*, permitiendo a un modelo alcanzar el podio sin siquiera haber sido entrenado, solamente observando los resultados de exactitud (*accuracy*) obtenidos [15].

Otra manera de demostrar que un modelo presenta una mejora frente a otros modelos, es mediante la realización de pruebas de significancia estadística. Esta alternativa utiliza la estadística para determinar que un modelo es diferente y superior a otro fuera de los márgenes de la casualidad. Sin embargo, una implementación deficiente de las pruebas de significancia estadística podría llevarnos a conclusiones erróneas. Por lo tanto, en esta investigación se presenta una manera práctica de implementarlas. Se demostró la utilidad de la propuesta, realizando la comparación estadística del desempeño de tres modelos convolucionales entrenados con la base de datos RAVDESS y tres modelos para la base de datos IEMOCAP.

2. Trabajos relacionados

La comparación del desempeño de modelos de clasificación basados en aprendizaje de máquinas no es una tarea trivial [12]. Para el caso de modelos de reconocimiento de emociones de aprendizaje profundo, la mayoría de las revisiones de literatura, por ejemplo [11,14,1,8], se limitan a realizar análisis descriptivos de los métodos de reconocimiento de emociones sin llegar a replicar (diferente equipo, mismo diseño experimental) o reproducir (diferente equipo, diferente diseño experimental) las investigaciones del estado del arte.

Pero si hablamos de comparaciones específicas, Fayek et al. [6], presentaron una comparación de algunos modelos de aprendizaje profundo, siendo esta forma (replicar por uno mismo varios modelos y evaluarlos) la más común a la que los autores recurren para comparar una propuesta contra el estado del arte. Este problema se extiende a otras áreas donde se ha observado problemas de rigurosidad, replicabilidad y falta de estándares para realizar comparaciones.

Por ejemplo, en el área de sistemas de recomendación, Zun et al. [13] ponen en cuestión las evaluaciones realizadas y proponen una herramienta para hacerlo de manera adecuada. De la misma forma en el área de redes neuronales basadas en grafos, Errika et al. [5] proponen dos fases para conseguir comparaciones justas y reproducibles: selección de modelos y evaluación de modelos. También explora los fallos más comunes para lograr reproducibilidad como la ausencia de información sobre el preprocesamiento y el particionamiento de los datos.

Este tipo de investigaciones nos demuestran lo lejos que estamos de tener comparaciones confiables. Los mayores esfuerzos están sumados en la realización de desafíos (*challenges*), que son competencias donde se distribuye una base de datos etiquetada. Los competidores entrenan sus modelos y finalmente los modelos deben ser evaluados contra una base de datos de prueba.

Según los resultados obtenidos, de acuerdo a alguna métrica elegida, se ubica a los competidores en una tabla de posiciones. De esta forma se busca garantizar que las comparaciones se están realizando en las mismas condiciones con respecto al conjunto de datos proporcionado. En el 2011 se presentó el primer AVEC (Desafío de emociones audiovisuales - *Audio/Visual Emotion Challenge*) [10], que incluyó un sub-desafío utilizando solamente audio. En este desafío se utilizó la base de datos SEMAINE y se proporcionó 3 particiones de datos, evaluándose la exactitud. Esta competencia se llevó a cabo hasta el año 2015 [9], cambiando posteriormente los desafíos a detección de emociones específicas como afecto o emoción.

Aunque AVEC está enfocado en el estudio de detección de emociones multimodal, sentó las bases para evaluar los modelos de manera más confiable, proponiendo bases de datos, métricas e incluso un punto de referencia (*baseline*). EmotiW (Desafío de reconocimiento de emociones - *Emotion Recognition in the Wild Challenge*) [3], de manera similar, propone cada año un desafío de detección de emociones audiovisuales. En este evento se propone el uso de una base de datos dividida en tres particiones (entrenamiento, validación y pruebas) y de seis emociones categóricas (enojo, digusto, miedo, neutral, tristeza y sorpresa).

Aunque EmotiW está enfocado en videos como datos de entrada, propuso un punto de referencia basado en el aprendizaje profundo, lo cual es innovador en comparación a AVEC donde se utilizaba características manuales para el entrenamiento de los modelos. Sin embargo, cuando se realizan desafíos no se tiene certeza de que los resultados no hayan sido fruto de la casualidad, o de un intento de ataque por fuerza bruta como lo demuestra Whitehill J. [15]. Es por eso que se requiere explorar estrategias más robustas, como la validación cruzada por k-fold. La cual permite hacer varias pruebas sobre la misma base de datos, permitiendo detectar alteraciones de los resultados por causa del particionamiento de la base de datos.

3. Materiales y métodos

3.1. Materiales

Los materiales utilizados en esta investigación están conformados por dos bases de datos de emociones en el habla, tres modelos de aprendizaje profundo por cada una de ellas y un computador donde se ejecutan los experimentos. A continuación se detallan las bases de datos utilizadas:

- RAVDESS (*Ryerson Audio-Visual Database of Emotional Speech and Song*) [7]: Esta base de datos cuenta con 1441 segmentos de audios, que suman cerca de una hora y media de duración. Los audios están etiquetados con las emociones: calma, felicidad, tristeza, enojo, miedo, sorpresa y disgusto.
- IEMOCAP (*Interactive Emotional Dyadic Motion Capture*) [2]: Es una base de datos de emociones categóricas y dimensionales, representadas en audio y video, con cerca de 10000 muestras.

CNN1D	M11
Conv1d(kernel_size=80) BatchNorm1d(n_channel) MaxPool1d(4) Dropout() Conv1d(kernel_size=3) BatchNorm1d() MaxPool1d(4) Dropout() Conv1d(kernel_size=3) BatchNorm1d() MaxPool1d(4) Dropout() Conv1d() BatchNorm1d() MaxPool1d(4) Linear()	Conv1d(kernel_size=80) BatchNorm1d() ReLU() Dropout(0.15) MaxPool1d(4) Conv1d(kernel_size=3) ReLU() Dropout(0.15) Conv1d(kernel_size=3) ReLU() Dropout(0.15) MaxPool1d(4) Conv1d(kernel_size=3) ReLU(), Dropout(0.15), Conv1d(kernel_size=3) ReLU(), Dropout(0.15), MaxPool1d(4), Conv1d(kernel_size=3) ReLU() Dropout(0.15) Conv1d(kernel_size=3) ReLU() Dropout(0.15), Conv1d(kernel_size=3) ReLU() Dropout(0.15) MaxPool1d(4) Conv1d(kernel_size=3) ReLU() Dropout(0.15) Conv1d(kernel_size=3) ReLU() Dropout(0.15) Linear()

Fig. 1. Arquitectura de los modelos utilizados para la base de datos RAVDESS.

Para la base de datos RAVDESS se implementaron dos modelos de redes neuronales basados en convoluciones 1D, que toman como entrada un vector de audio plano. La arquitectura de los modelos (m11, m5) se detallan en la Figura 1.Mientras que para la base de datos IEMOCAP se implementaron tres modelos que utilizan una arquitectura de Red Neuronal Convolucional (Convolución + MaxPool + Convolución + MaxPool + ... + Flatten + Dense).

Cada modelo recibe como entrada una matriz que representa un espectrograma MFCC de un segmento de audio. La arquitectura de cada modelo (A, B , C) se detallan en la Figura 2. Para la implementación de los algoritmos y modelos se utilizó las librerías Pytorch y Sklearn. El entorno computacional utilizado en el entrenamiento de los modelos y la ejecución de los experimentos estuvo

conformado por un computador personal con procesador Core i5 9400F, 16 GB de memoria RAM, una tarjeta NVidia RTX 2060 (6GB Vram) y el Sistema Operativo Linux.

Modelo A	Modelo B	Modelo C
Input(shape=(20, window)) Conv2D(64, 3, "relu") MaxPool2D((1, 2)) Conv2D(64, 3, "relu") MaxPool2D((1, 2)) Conv2D(64, (1, 3), "relu") MaxPool2D((1, 2)) Conv2D(64, (1, 3), "relu") MaxPool2D((2, 1)) Conv2D(128, 3, "relu") Dropout(0.3) Flatten() Dense(1024, "relu") Dropout(0.4) Dense(8, "softmax")	Sequential(Input(shape=(20, window)) Conv2D(64, 3, "relu") MaxPool2D((1, 2)) Conv2D(64, 3, "relu") MaxPool2D((1, 2)) Conv2D(128, 3, "relu") MaxPool2D((1, 2)) MaxPool2D((2, 1)) Conv2D(128, 4, "relu") Dropout(0.3) Flatten() Dense(1024, "relu") Dropout(0.4) Dense(8, "softmax")	Sequential(Input(shape=(20, window)) Conv2D(64, 3, "relu") MaxPool2D((1, 2)) Conv2D(64, (1, 3), "relu") MaxPool2D((1, 2)) Conv2D(64, (1, 3), "relu") MaxPool2D((2, 1)) Conv2D(128, 3, "relu") Dropout(0.3) Flatten() Dense(1024, "relu") Dropout(0.4) Dense(1024, "relu") Dropout(0.4) Dense(8, "softmax")

Fig. 2. Arquitectura de los modelos utilizados para la base de datos IEMOCAP.

3.2. Métodos

El método utilizado en la presente investigación consiste en la implementación de los siguientes procedimientos:

1. **Pre-procesamiento:** Todos los segmentos de audios son remuestrados a 8khz. Y en el caso de IEMOCAP transformados a su representación MFCC (Coeficientes Cepstrales en las Frecuencias de Mel - *Mel Frequency Cepstral Coefficients*). Los parámetros utilizados para esta transformación fueron: $WindowSize = 55$ y $HopLength = 10$.
2. **Selección de arquitectura de modelos:** La arquitectura de los modelos a comparar es determinada empíricamente buscando maximizar la métrica exactitud (*accuracy*) en la validación.
3. **Recolección de muestras:** Una vez que los modelos a ser comparados han sido determinados, se procede a recolectar muestras de su comportamiento con diferentes particiones de la base de datos. Para esto, se utilizan iteraciones sucesivas de validaciones por k-fold de la manera presentada en el Algoritmo **??**. Se definen los modelos a evaluar M_A y M_B y en cada iteración

se recolecta un conjunto C de validaciones k-fold con longitud k:

$$C(D, M_A, M_B) = \{(S_A^j, S_B^j), (S_A^{j+1}, S_B^{j+1}), ..., (S_A^k, S_B^k))\}, \quad (1)$$

donde S_A^j se refiere al puntaje o *score* (métrica seleccionada a utilizar, por ejemplo, exactitud) obtenido por el modelo M_A sobre la partición de datos D_j generado por el método k-fold. Se repite N veces este procedimiento, obteniendo N conjuntos C_i:

$$G = \{C_i, C_{i+1}, ..., C_N\}. \quad (2)$$

4. **Prueba de significancia estadística:** Con las muestras recolectadas, se realizan las pruebas de significancia estadística. Para cada $C_i \in G$, se calculan los valores p-value y t-value mediante la prueba estadística de Wilcoxon. El valor t indica cuanta diferencia existe entre los resultados obtenidos por los dos modelos. Mayores valores de t implican menores valores de p. Cuando $p < 0{,}05$ podemos asumir que existe diferencia estadísticamente significativa. Como puede verse en el Algoritmo **??**, este procedimiento se realiza para cada $C_i \in G$.

Si analizamos el subconjunto $SG = \{p | p \in P \wedge p < 0{,}05\}$ y observamos la proporción de las cardinalidades $n(SG)/n(G)$, es decir, el porcentaje de muestras en las que se evidencia diferencia significativa, se puede apreciar el comportamiento general comparativo de estos modelos. El código fuente de esta implementación se encuentra disponible en el siguiente repositorio[1].

4. Resultados

Para la base de datos RAVDESS se compararon los modelos CNN1D Short, M11 (batch_size = 128) y M11 (batch_size = 128) utilizando **10 iteraciones** del algoritmo propuesto y un k-fold de 6.

Cuando observamos la comparación de los modelos CNN1D y M11_128 (Fig. 3), encontramos 4 resultados por debajo del umbral $p < 0{,}05$. Lo que indica diferencia significativa. Pero también podemos observar 6 resultados por encima del umbral.

Es decir, bajo ciertas condiciones hay evidencia para afirmar que existe diferencia significativa entre los modelos. Pero, en otras condiciones podríamos tener evidencia de lo contrario. En la comparación de los modelos CNN1D y M11_256 (Figura 4) observamos el mismo comportamiento que la comparación anterior.

La comparación entre los modelos M11_128 y M11_256, (Fig. 6) muestra que en todos los resultados no existe diferencia significativa entre las dos configuraciones del modelo M11. El principal limitante para obtener más iteraciones

[1] https://github.com/luantber/ser-benchmark

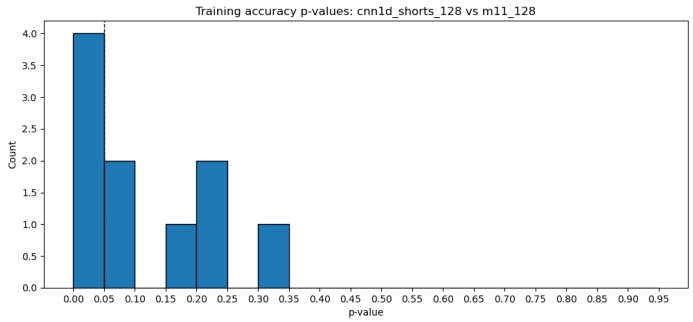

Fig. 3. Comparativa de los modelos CNN1D vs M11_128.

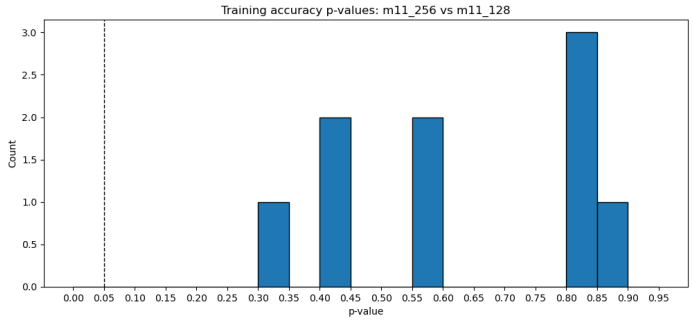

Fig. 4. Comparativa de los modelos CNN1D vs M11_256.

es el alto costo computacional del algoritmo, pues requiere reentrenar todos los modelos $N \times K$ veces.

Por otro lado, utilizando el dataset IEMOCAP, y tras recolectar datos de la métrica de *accuracy* durante 30 iteraciones y k-fold igual a 5 por un periodo de tiempo de cerca de 41 horas, se obtuvieron los resultados que se presentan en la Figura 5. Como se puede observar, todos los valores p-value obtenidos están por encima de 0.05, y dado que definimos como umbral para rechazar la hipótesis nula valores menores a 0.05, no podemos afirmar que exista alguna diferencia significativa entre los tres modelos para este dataset.

5. Discusión

Los resultados obtenidos en el experimento utilizando el dataset RAVDESS demuestran que, incluso, realizando una estrategia de k-fold y una prueba estadística, los resultados podrían cambiar e inclusive contradecirse. Pues con el mismo modelo se podría aceptar y rechazar la hipótesis H_0 si se repitiera el experimento con la misma base de datos, pero diferentes particiones en la validación k-fold.

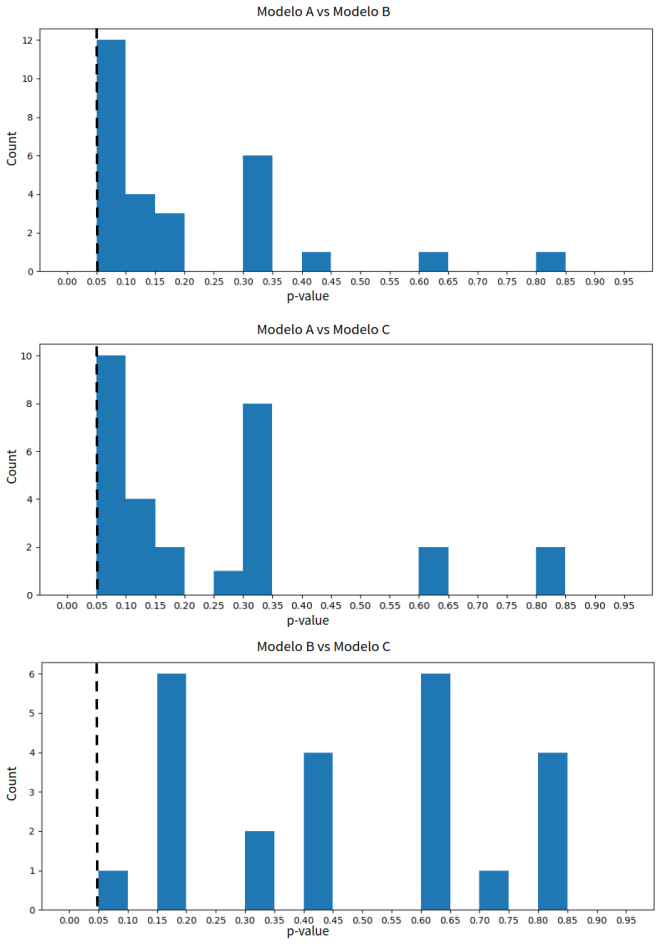

Fig. 5. *p*-values del modelo A vs. el modelo B (arriba), del modelo A vs. el modelo C (medio) y del modelo B vs el modelo C (abajo) utilizando el dataset IEMOCAP.

Por tal motivo, se podría llegar a conclusiones erróneas o sin la suficiente evidencia para afirmar que un modelo es superior a otro. Esta situación podría calificarse como producto de la casualidad o del azar. Como puede verse en los resultados, queda a criterio de los investigadores definir un umbral para declarar a un modelo superior a otro. Por ejemplo, en nuestro caso podría ser 30 %.

Con este umbral se podría afirmar que el modelo A es superior al modelo B, dado que existe 30 % de ocasiones en las que existe diferencia significativa. Sin embargo, no se podría afirmar lo mismo para la comparación del modelo B y C, pues solo un 20 % de los entrenamientos mostraron diferencia significativa. También debemos destacar que la propuesta busca mostrar la tendencia de porcentajes que tendrían los modelos si se repitieran más veces.

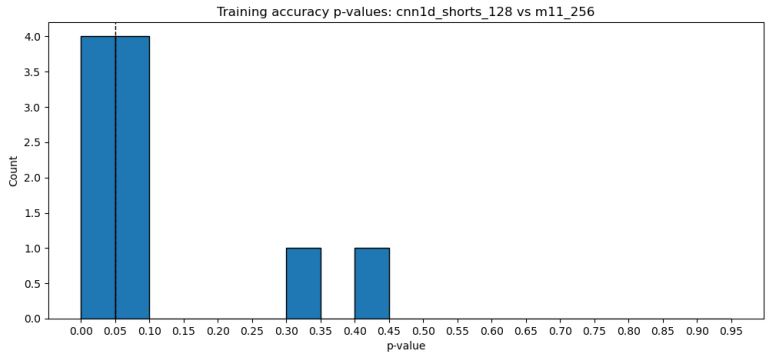

Fig. 6. Comparativa de los modelos M11_128 vs M11_256.

Repetir el experimento una suficiente cantidad de veces puede evidenciar una tendencia clara sobre la superioridad de un modelo en términos de la métrica deseada. Esto nos permitiría aseverar con mayor firmeza que "el modelo X es significativamente superior al modelo Y, en un % de las veces", o en caso contrario, como en el experimento IEMOCAP, que no existe diferencia significativa entre los modelos.

6. Conclusiones

El área de *Speech Emotion Recognition* viene siendo ampliamente estudiada, por lo que se pueden encontrar varios estudios de revisión y trabajos de recopilación, sobre el estado del arte, publicados recientemente. Estas investigaciones muestran la amplia variedad de modelos que se han propuesto para ésta tarea. Sin embargo, un número considerable de propuestas podrían no presentar resultados confiables, que permitan afirmar de manera objetiva que una propuesta es superior a otra.

Esto, entre otras razones, se debe a que las comparaciones con otras propuestas no toman en cuenta los elementos aleatorios que forman parte de los modelos de *Deep Learning*. El presente trabajo busca contribuir en la resolución de esta problemática, presentando una manera práctica de implementar evaluaciones comparativas de modelos de reconocimiento de emociones, mediante el uso de pruebas de significancia estadística.

Se demostró la utilidad de la propuesta, realizando la comparación del desempeño de tres modelos convolucionales entrenados con los dataset RAVDESS y IEMOCAP. Los resultados permitieron determinar el porcentaje de veces en los que los modelos evaluados obtuvieron un desempeño similar y el qué porcentaje de veces donde su desempeño tuvo una diferencia significativa.

Si bien se encontró que en la mayoría de los casos (entre 70 % y 80 %) los modelos tuvieron un desempeño similar, en el restante de los casos se encontró

una diferencia significativa. Esta discrepancia demuestra que si no se hubiese realizado la evaluación propuesta, podríamos llegar a cualquiera de las dos conclusiones (que existe diferencia o que son modelos similares). Con la propuesta presentada podemos identificar la tendencia en porcentajes de diferencia o similitud que tienen los modelos.

7. Trabajo futuro

En una próxima investigación se podrían realizar evaluaciones comparativas con modelos más complejos y bases de datos de mayor tamaño, teniendo en cuenta las restricciones impuestas por los tiempos de entrenamiento y la complejidad computacional. Estas restricciones podrían aliviarse distribuyendo el entrenamiento en múltiples nodos, lo cual permitiría la generación de más muestras que permitirían llegar a conclusiones más acertadas. Una limitación de esta investigación es que está sujeta a una sola estrategia de pre-procesamiento para todos los modelos.

La utilización de diferentes estrategias (p.ej., utilizar diferentes parámetros para generar el espectrograma) es una mejora interesante a ser implementada en trabajos futuros, ya que influye directamente en el desempeño de los modelos. Si bien la presente investigación está enfocada y ha sido realizada en un contexto del *Speech Emotion Recognition*, podría ser replicada en cualquier contexto donde se necesite comparar múltiples clasificadores multiclase.

Agradecimientos. Este artículo es parte de los resultados de la tesis de pregrado del Bachiller en Ciencias de la Computación Luis Bernal Chahuayo, la misma que ha sido financiada por el Proyecto Concytec - Banco Mundial "Mejoramiento y Ampliación de los Servicios del Sistema Nacional de Ciencia Tecnología e Innovación Tecnológica" 8682-PE, a través de su unidad ejecutora ProCiencia [Contrato número 014-2019-FONDECYT-BM-INC.INV].

Referencias

1. Akçay, M. B., Oğuz, K.: Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Communication, vol. 116, pp. 56–76 (2020)
2. Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., Narayanan, S. S.: IEMOCAP: Interactive emotional dyadic motion capture database. Language resources and evaluation, vol. 42, no. 4, pp. 335–359 (2008)
3. Dhall, A., Kaur, A., Goecke, R., Gedeon, T.: Emotiw 2018: Audio-video, student engagement and group-level affect prediction. In: Proceedings of the 20th ACM International Conference on Multimodal Interaction. pp. 653–656 (2018)
4. Dzedzickis, A., Kaklauskas, A., Bucinskas, V.: Human emotion recognition: Review of sensors and methods. Sensors, vol. 20, no. 3, pp. 592 (2020)
5. Errica, F., Podda, M., Bacciu, D., Micheli, A.: A fair comparison of graph neural networks for graph classification. arXiv preprint arXiv:1912.09893, (2019)

6. Fayek, H. M., Lech, M., Cavedon, L.: Evaluating deep learning architectures for speech emotion recognition. Neural Networks, vol. 92, pp. 60–68 (2017)
7. Livingstone, S. R., Russo, F. A.: The Ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PloS one, vol. 13, no. 5, pp. e0196391 (2018)
8. Oh, S., Kim, D. K.: Comparative analysis of emotion classification based on facial expression and physiological signals using deep learning. Applied Sciences, vol. 12, no. 3 (2022) doi: 10.3390/app12031286.
9. Ringeval, F., Schuller, B., Valstar, M., Jaiswal, S., Marchi, E., Lalanne, D., Cowie, R., Pantic, M.: Avec 2015: The first affect recognition challenge bridging across audio, video, and physiological data. In: Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge. pp. 3–8 (2015)
10. Schuller, B., Valstar, M., Eyben, F., McKeown, G., Cowie, R., Pantic, M.: Avec 2011–the first international audio/visual emotion challenge. In: International Conference on Affective Computing and Intelligent Interaction. pp. 415–424. Springer (2011)
11. Schuller, B. W.: Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. Communications of the ACM, vol. 61, no. 5, pp. 90–99 (2018)
12. Stapor, K., Ksieniewicz, P., García, S., Woźniak, M.: How to design the fair experimental classifier evaluation. Applied Soft Computing, vol. 104, pp. 107219 (2021)
13. Sun, Z., Yu, D., Fang, H., Yang, J., Qu, X., Zhang, J., Geng, C.: Are we evaluating rigorously? benchmarking recommendation for reproducible evaluation and fair comparison. In: Fourteenth ACM Conference on Recommender Systems. pp. 23–32 (2020)
14. Swain, M., Routray, A., Kabisatpathy, P.: Databases, features and classifiers for speech emotion recognition: a review. International Journal of Speech Technology, vol. 21, no. 1, pp. 93–120 (2018)
15. Whitehill, J.: Climbing the kaggle leaderboard by exploiting the log-loss oracle. In: Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence (2018)