ISBN: 978-628-96613-3-0. ISSN: 2414-6390. Digital Object Identifier: https://dx.doi.org/10.18687/LACCEI2026.1.1.2195

Recognition and Classification of Emotions in Driver Voice Data Streams

Gary Reyes, PhD [1,2,3,4], Roberto Tolozano-Benites, PhD[1], Fiorella Achi Limones[2], Gia Zhunio Ramírez [2], Laura Lanzarini, PhD [4], Waldo Hasperué, PhD [4], Dayron Rumbaut [1,3]
[1]Carrera de Sistemas Inteligentes, Universidad Bolivariana del Ecuador, Campus Durán Km 5.5 vía Durán Yaguachi, Durán 092405, Ecuador, gxreyesz@ube.edu.ec, rtolozano@ube.edu.ec, jjbarzolam@ube.edu.ec
[2]Facultad de Ciencias Matemáticas y Físicas, Universidad de Guayaquil, Cdla. Universitaria Salvador Allende, Guayaquil 090514, Ecuador, gary.reyesz@ug.edu.ec, fiorella.achil@ug.edu.ec , gia.zhunior@ug.edu.ec
[3]Artificial Intelligence Research Group, Universidad Bolivariana del Ecuador, Campus Durán Km 5.5 vía Durán Yaguachi, Durán 092405, Ecuador, gxreyesz@ube.edu.ec
[4]Instituto de Investigación en Informática LIDI (Centro CICPBA), Facultad de Informática, Universidad Nacional de La Plata, Buenos Aires CP1900, Argentina, laural@lidi.info.unlp.edu.ar, whasperue@lidi.info.unlp.edu.ar

Abstract*—Voice-based emotion recognition presents a fundamental challenge in automotive driving, as the driver's emotional state directly influences their cognitive abilities and decision-making processes. This study evaluates the Adaptive Random Forest (ARF) algorithm for emotion detection in driving scenarios, using a proprietary dataset of 300 voice recordings across four emotional categories (Anger, Happiness, Sadness, Neutral). The methodology integrates a hybrid strategy of real and semi-synthetic data streams to model dynamically evolving emotional patterns. The ARF is compared to two incremental learning reference models—Hoeffding Tree (HT) and Naive Bayes Gaussian (NB)—under identical streaming conditions. The results confirm the superiority of ARF (Accuracy: 63.67%, F1 macro: 47.65%, Recall: 47.36%) compared to NB (Accuracy: 55.00%, F1: 49.69%) and HT (Accuracy: 36.67%, F1: 5.91%), validating its effectiveness for real-time emotion recognition in intelligent transportation systems and non-intrusive driver monitoring.

Keywords: Speech Emotion Recognition, Speech Processing, Adaptive Random Forest, Speech-Based Emotion Classification, Conceptual Drift, Hoeffding Tree, Naive Bayes.

Reconocimiento y Clasificación de emociones en flujos de datos de voz de conductores

Gary Reyes, PhD [1,2,3,4], Roberto Tolozano-Benites, PhD[1], Fiorella Achi Limones[2], Gia Zhunio Ramírez [2], Laura Lanzarini, PhD [4], Waldo Hasperué, PhD [4], Dayron Rumbaut [1,3]
[1]Carrera de Sistemas Inteligentes, Universidad Bolivariana del Ecuador, Campus Durán Km 5.5 vía Durán Yaguachi, Durán 092405, Ecuador, gxreyesz@ube.edu.ec, rtolozano@ube.edu.ec, jjbarzolam@ube.edu.ec
[2]Facultad de Ciencias Matemáticas y Físicas, Universidad de Guayaquil, Cdla. Universitaria Salvador Allende, Guayaquil 090514, Ecuador, gary.reyesz@ug.edu.ec, fiorella.achil@ug.edu.ec , gia.zhunior@ug.edu.ec
[3]Artificial Intelligence Research Group, Universidad Bolivariana del Ecuador, Campus Durán Km 5.5 vía Durán Yaguachi, Durán 092405, Ecuador, gxreyesz@ube.edu.ec
4Instituto de Investigación en Informática LIDI (Centro CICPBA), Facultad de Informática, Universidad Nacional de La Plata, Buenos Aires CP1900, Argentina, laural@lidi.info.unlp.edu.ar, whasperue@lidi.info.unlp.edu.ar

Resumen– ***El reconocimiento de emociones basado en la voz constituye un reto fundamental en la conducción automotriz, ya que el estado emocional del conductor influye directamente en sus capacidades cognitivas y en sus procesos de toma de decisiones. Este estudio evalúa el algoritmo Adaptive Random Forest (ARF) para la detección de emociones en escenarios de conducción, utilizando un conjunto de datos propio de 300 grabaciones de voz en cuatro categorías emocionales (Enojo, Felicidad, Tristeza, Neutral). La metodología integra una estrategia híbrida de flujos de datos reales y semisintéticos para modelar patrones emocionales que evolucionan dinámicamente. El ARF se compara con dos modelos de referencia de aprendizaje incremental —Hoeffding Tree (HT) y Naive Bayes Gaussiano (NB)— bajo condiciones de streaming idénticas. Los resultados confirman la superioridad del ARF (Exactitud: 63,67 %, F1 macro: 47,65 %, Recall: 47,36 %) frente al NB (Exactitud: 55,00 %, F1: 49,69 %) y al HT (Exactitud: 36,67 %, F1: 5,91 %), validando su eficacia para el reconocimiento de emociones en tiempo real en sistemas de transporte inteligente y monitoreo no intrusivo de conductores.***

Palabras clave: Reconocimiento de Emociones en el Habla, Procesamiento de Voz, Bosque Aleatorio Adaptativo, Clasificación de Emociones Basada en Voz, Deriva Conceptual, Hoeffding Tree, Naive Bayes.

I. Introducción

El reconocimiento automático de emociones a través de la voz se ha convertido en un área de investigación cada vez más activa dentro de la Inteligencia Artificial, con aplicaciones que van desde la salud digital y la educación hasta los sistemas de seguridad vial. A diferencia de los métodos invasivos basados en sensores, el análisis acústico es completamente natural y no intrusivo, lo que lo hace especialmente adecuado para entornos vehiculares [1], [2].

Las condiciones del tráfico —congestión, horas pico, fallas en infraestructura o accidentes— generan con frecuencia diversas respuestas emocionales en los conductores. Estados como el estrés, la ira, la fatiga o la ansiedad pueden alterar significativamente el desempeño al volante e incrementar el riesgo de accidentes. Si se detectan a tiempo, estos estados podrían activar sistemas inteligentes que alerten al conductor o desencadenen medidas de seguridad preventivas [11], [27].

El reconocimiento tradicional de emociones ha dependido de modelos estáticos de aprendizaje automático (SVM, k-NN, CNN, LSTM) que asumen condiciones de datos invariantes en el tiempo. Esta suposición no se sostiene en escenarios reales de conducción, donde los datos llegan de forma continua y pueden presentar deriva conceptual (concept drift) [10]. Por ello, se requieren algoritmos adaptativos que procesen datos de forma incremental y se ajusten a los cambios en la distribución.

El algoritmo Adaptive Random Forest (ARF) destaca por su capacidad de aprender continuamente, detectar la deriva conceptual y mantener un rendimiento robusto bajo condiciones cambiantes, integrándose de manera natural con características acústicas estándar como MFCC, Chroma y coeficientes de frecuencia Mel [10].

Este artículo aborda las siguientes preguntas de investigación: (1) ¿Puede el ARF adaptarse eficientemente a los cambios acústicos en los flujos de voz de conductores? (2) ¿Cómo se comparan los métodos de aprendizaje incremental en términos de exactitud de clasificación? (3) ¿Cómo se desempeña el ARF frente a otros modelos de referencia incrementales (Hoeffding Tree, Naive Bayes) en la detección de emociones en flujos acústicos continuos?

Este artículo se organiza de la siguiente manera: la Sección II revisa el trabajo relacionado; la Sección III describe los materiales y métodos; la Sección IV presenta los resultados; la Sección V discute los hallazgos; y la Sección VI presenta las conclusiones y el trabajo future.

II. Trabajo Relacionados

La conducción es una actividad compleja que requiere la integración constante de procesos cognitivos, atención y toma de decisiones rápidas en un entorno dinámico. Estos procesos se ven modulados por el estado emocional del conductor, el cual puede influir significativamente en su desempeño al volante. Según la RAE, una emoción es una alteración del ánimo

acompañada de conmociones somáticas y una reacción corporal causada por un evento del entorno [11]. Desde una perspectiva más amplia, la psicología de las emociones establece que estas constituyen procesos multidimensionales que involucran componentes cognitivos, fisiológicos y conductuales [8]. En el contexto de la conducción, emociones como la ira, la tristeza, el estrés o la ansiedad son frecuentes, y diversos estudios han documentado que estas se asocian a incrementos en la velocidad, mayor impulsividad y disminución de la atención selectiva, elevando significativamente la probabilidad de incidentes viales [11], [27].

Durante la última década, el reconocimiento de emociones en conductores ha experimentado un crecimiento significativo, consolidándose como un área de investigación fundamental para la seguridad vial [30]. Los investigadores han enfocado su atención principalmente en estados emocionales de alta activación y valencia negativa —estrés, ira y ansiedad— debido a su relación directa con el deterioro de las capacidades cognitivas necesarias para la conducción segura [1], [23]. Los enfoques metodológicos para la detección de emociones en conductores pueden clasificarse en cinco categorías: métodos basados en expresión facial [23], análisis del habla [21], [22], [24], patrones de comportamiento de conducción, señales fisiológicas [26], e integración multimodal [9], [30]. De estos enfoques, el reconocimiento basado en habla ha ganado especial relevancia por su naturaleza no invasiva y su capacidad de integrarse de manera natural en los sistemas de interacción conductor-vehículo existentes [1], [2].

a) Evolución de las Técnicas de Reconocimiento de Emociones en el Habla

El reconocimiento de emociones en el habla ha evolucionado significativamente durante las últimas dos décadas [25]. Los enfoques iniciales se apoyaron en modelos estadísticos como las Máquinas de Vectores de Soporte (SVM), desarrolladas en los laboratorios AT&T Bell Labs como técnica de clasificación basada en kernels [18], [28]. Estas demostraron eficacia en el reconocimiento de locutores e idiomas mediante bandas espectrales [20], sentando las bases metodológicas para sistemas posteriores de reconocimiento emocional.

Una limitación clave de estos enfoques era la extracción manual de características acústicas de alta calidad, como los Coeficientes Cepstrales en Frecuencia Mel (MFCC), introducidos por Davis y Mermelstein [5] como representación compacta del espectro vocal. La transformada wavelet emergió como herramienta complementaria para la representación tiempo-frecuencia de señales, ofreciendo mayor flexibilidad que los MFCC tradicionales [13], y trabajos posteriores exploraron su aplicación en reconocimiento de comandos de voz y emociones [14].

Los métodos de aprendizaje profundo —CNN, RNN, Transformers y Autoencoders— demostraron posteriormente una capacidad superior para capturar patrones temporales y espectrales complejos en señales de audio [15]. En particular, los transformers de audio han alcanzado un rendimiento excepcional al explotar mecanismos de atención para dependencias acústicas de largo alcance, especialmente en tareas con hablantes desconocidos [22]. Los modelos híbridos CNN-BiLSTM, combinados con técnicas de aumento de datos como la adición de ruido y el desplazamiento de espectrogramas, han logrado precisiones superiores al 70 % en experimentos dependientes e independientes del hablante [3], [29]. Investigaciones recientes han complementado estos avances mediante la combinación de múltiples bases de datos para aumentar la diversidad del entrenamiento [4], [19] y el uso de características espectrales avanzadas más allá de los MFCC, con especial atención al español, cuyas características prosódicas y fonéticas varían significativamente respecto a otros idiomas [2], [7].

b) Aprendizaje Incremental y Adaptación en Tiempo Real

A pesar de los avances en precisión, los métodos de aprendizaje profundo enfrentan desafíos importantes en interpretabilidad, fiabilidad, costo computacional y adaptabilidad incremental [15]. En entornos de datos en continua evolución, como el monitoreo de vehículos en tiempo real, se requieren métodos capaces de procesar flujos de datos continuos sin necesidad de reentrenamiento completo del modelo [21], [24].

El modelo Adaptive Random Forest (ARF) resulta particularmente adecuado para este tipo de escenarios debido a sus propiedades de ensemble y sus capacidades de adaptación dinámica [10]. Esto es especialmente relevante en entornos donde los hablantes —con distintos géneros, acentos y estilos de habla— y el ruido ambiental cambian constantemente. A diferencia de intentos previos de adaptar Random Forest a flujos de datos, ARF incorpora un método de remuestreo efectivo y operadores adaptativos que gestionan distintos tipos de deriva conceptual sin optimizaciones complejas por conjunto de datos, y su escalabilidad ha sido validada en múltiples dominios [10], [12].

Sus tres mecanismos principales son: (1) agregación de diversidad mediante bagging en línea, donde las instancias son seleccionadas siguiendo una distribución binomial aproximable a Poisson; (2) selección aleatoria de subconjuntos de características para divisiones de nodos; y (3) detectores de deriva ADWIN por árbol base, que permiten reinicios selectivos ante cambios en la distribución de los datos [10]. Cabe destacar que la estrategia de adaptación no reinicia inmediatamente los modelos base ante un cambio detectado; en su lugar, entrena un árbol de fondo a partir de la advertencia y solo sustituye el modelo primario si la deriva se confirma, lo que otorga al sistema mayor estabilidad y precisión ante fluctuaciones transitorias.

c) Aplicaciones Industriales y Sistemas Integrados

Un ejemplo destacado de aplicación industrial es el sistema MBUX (Mercedes-Benz User Experience), que integra un asistente de voz con reconocimiento de emociones en tiempo real mediante análisis rítmico, señales fisiológicas y expresiones faciales [17]. El sistema procesa flujos continuos de voz del conductor extrayendo características acústicas como

el tono, el timbre y la intensidad, lo que le permite inferir estados emocionales relevantes —estrés, fatiga o frustración— y generar respuestas adaptativas multimodales que pueden sugerir ajustes en el sistema de conducción [17]. Complementariamente, la implementación de aprendizaje federado en sistemas vehiculares permite personalizar estos modelos para cada conductor preservando la privacidad de los datos, procesando la información en el borde (edge computing) sin necesidad de transmitir datos sensibles a servidores centrales.

III. Materiales y Métodos

Este estudio propone una metodología basada en aprendizaje adaptativo que emplea el algoritmo Adaptive Random Forest (ARF) para identificar continuamente las emociones en la voz del conductor a partir de flujos de datos en tiempo real. El análisis de los estados emocionales en el entorno de la conducción resulta complejo debido a la naturaleza dinámica del habla humana, los modismos, las variaciones individuales y los cambios bruscos en las condiciones ambientales propias de la conducción.

a) Fase 1: Adquisición de Datos

Esta sección comprende una serie de muestras de voces sin procesar grabadas en entornos de conducción reales y simuladas, que constituyen la base de la metodología propuesta en el reconocimiento de emociones de voces en conductores.

Todo el proceso comienza con la fase de análisis, que solo requiere una única ejecución por muestra con la captura de grabaciones de voz en entornos de conducción reales y simulados. Para garantizar la representatividad de los datos, se definen previamente los escenarios de grabación considerando variables contextuales como las condiciones ambientales (ruido del tráfico, clima), el momento del día y los estados emocionales que se desean capturar, ya sean inducidos experimentalmente o naturales. Las grabaciones se realizan buscando maximizar la estabilidad de la señal y replicar fielmente las condiciones acústicas del interior del vehículo, minimizando interferencias externas que puedan comprometer la calidad de los datos. En total, se recopilaron 300 grabaciones distribuidas en cuatro categorías emocionales: Enojo (14,5 %), Felicidad (46,3 %), Tristeza (30,4 %) y Neutral (8,8 %), seleccionadas siguiendo estos criterios de representatividad para asegurar que cada muestra refleje fielmente el entorno vehicular.

Paralelamente, se recopilan los escenarios y condiciones asociados a cada grabación, documentando metadatos relevantes como condiciones de tráfico (alta congestión horas pico, flujo moderado y flujo libre), hora del día (mañana 07:00 a 09:30, y por la tarde-noche, de 17:00 a 20:00). Esta información contextual resulta fundamental para comprender las variaciones en los patrones emocionales y para posibles análisis posteriores de correlación entre contexto y estado emocional.

Una vez capturados los datos de audio, se procede a la clasificación emocional, donde cada grabación es etiquetada según el estado emocional manifestado. Esta clasificación puede basarse en taxonomías establecidas dentro del algoritmo ARF, cada fragmento de voz es etiquetado con una emoción específica ya que transforma las señales de voz crudas en datos estructurados permitiendo que el modelo aprenda a asociar patrones acústicos. Finalmente, todas las grabaciones clasificadas se consolidan en un Dataset de muestras de audio, que integra tanto las señales de voz con sus respectivas etiquetas emocionales y metadatos asociados a los escenarios. Este conjunto de datos se organiza de forma que facilite su uso en los procesos de entrenamiento, validación y pruebas, manteniendo la estabilidad entre las distintas clases emocionales detectadas evitando sesgos en el modelo ARF.

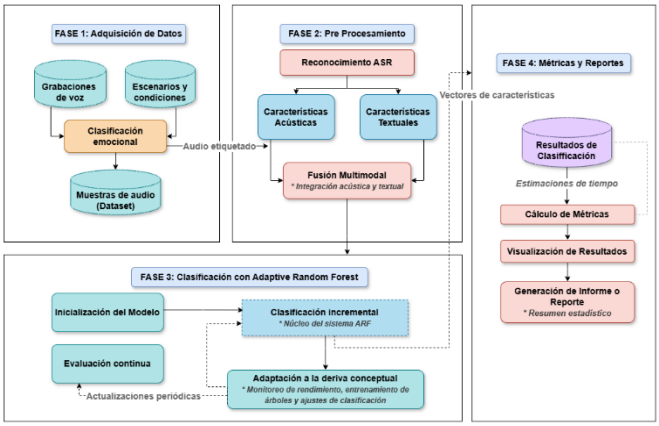

Fig 1. Metodología propuesta basada en ARF para reconocimiento de emociones en voz de conductores

b) Fase 2: Preprocesamiento

Esta fase transforma las grabaciones de voz en representaciones numéricas estructuradas que capturan tanto las características acústicas como textuales del habla, dado que el modelo no opera directamente sobre archivos de audio en formato .wav. El proceso inicia con el Reconocimiento Automático del Habla (ASR, *Automatic Speech Recognition*), que convierte las señales de audio en texto transcrito. Esta transformación no solo facilita el análisis del contenido lingüístico del discurso del conductor, sino que también permite identificar patrones en el vocabulario, la estructura sintáctica y el contenido semántico indicativos de estados emocionales específicos, proporcionando información complementaria a las características acústicas. La incorporación de ASR dentro del modelo ARF resulta especialmente relevante, dado que ciertos estados emocionales se manifiestan no solo en la forma de hablar, sino también en palabras y expresiones coloquiales.

Paralelamente, se utilizó la biblioteca Librosa de Python para la extracción de características acústicas, calculando coeficientes MFCC, Chroma y de frecuencia Mel a una

frecuencia de muestreo de 16 kHz. Posteriormente, las características acústicas y textuales son integradas mediante un proceso de fusión multimodal, cuyo propósito es combinar ambas fuentes de información en un vector de características unificado por instancia. Esta integración permite aprovechar las fortalezas de cada modalidad, mejorando la capacidad del algoritmo para detectar emociones de forma más robusta y precisa. La fase concluye con la generación del conjunto de datos preprocesados para el entrenamiento y la evaluación del modelo, garantizando la coherencia y practicidad de los datos.

c) Fase 3: Clasificación con Adaptive Random Forest

En la tercera fase, el algoritmo Adaptive Random Forest representa el núcleo de la metodología propuesta para el reconocimiento de emociones en flujos de datos de voz de conductores. En esta etapa, el modelo ARF clasifica de forma incremental los estados emocionales a partir de los vectores de características generados en la fase de preprocesamiento. A diferencia de los algoritmos de aprendizaje por lotes tradicionales, que requieren un reentrenamiento completo ante nuevas muestras, ARF procesa cada nueva instancia una sola vez sin almacenar ni reprocesar datos históricos, actualizando sus estructuras internas de manera progresiva y sin interrupciones.

El mecanismo de clasificación se apoya en el aprendizaje en conjunto (*ensemble learning*), donde múltiples árboles de decisión contribuyen conjuntamente a la predicción final de la emoción mediante votación mayoritaria ponderada por la exactitud estimada de cada árbol. Este enfoque permite que cada árbol capture distintos aspectos del espacio de características, mejorando la robustez del modelo frente al ruido, la variabilidad individual de los conductores y las condiciones cambiantes del entorno.

Durante el proceso de clasificación, el algoritmo realiza una evaluación continua del rendimiento analizando métricas como accuracy, precisión, recall y F1-score. Esta evaluación constante permite identificar degradaciones en el desempeño asociadas a cambios en el comportamiento emocional del conductor o en los escenarios de conducción, activando cuando es necesario los mecanismos de detección y adaptación a la deriva conceptual (*concept drift*).

Cada árbol incorpora un detector de deriva ADWIN de dos niveles: un nivel de advertencia que indica una posible desviación, y un nivel de deriva que confirma un cambio significativo. Al alcanzar el nivel de advertencia, un árbol de fondo comienza a entrenarse simultáneamente con nuevas instancias mientras el árbol principal continúa proporcionando predicciones. Si la deriva se confirma, el árbol de fondo reemplaza al árbol principal, garantizando una transición fluida sin pérdida de rendimiento. Si la advertencia resulta ser una falsa alarma, el árbol de fondo se descarta y el árbol principal continúa funcionando. Este enfoque dinámico permite que el modelo ARF se mantenga actualizado ante cambios abruptos o graduales en los datos, preservando su capacidad de clasificación a lo largo del tiempo.

Configuración de Hiperparámetros

Los hiperparámetros se seleccionaron mediante búsqueda en cuadrícula sobre un conjunto de validación correspondiente al 15 % de las instancias totales, manteniendo la restricción de procesamiento en una sola pasada. Los valores finales se eligieron para maximizar el F1 macro en el conjunto de validación, priorizando la estabilidad ante cambios de distribución (Tabla 1).

Tabla 1. Configuración de hiperparámetros para los modelos ARF, HT y NB. Elaboración propia.

Modelo	Parámetro	Valor	Justificación
ARF	n_models	10	Balance precisión/costo
ARF	grace_period (τ)	50	Detección sensible de deriva
ARF	delta ADWIN (δ)	0,001	Equilibrio deriva/ruido
HT	grace_period (τ)	150	Menor sensibilidad, menos FP
HT	delta (δ)	1×10^{-5}	Divisiones de alta confianza
HT	leaf_prediction	Clase mayoritaria	Estabilidad en hojas
NB	Distribución	Gaussiana	Características continuas

d) Fase 4: Métricas y reportes

En la etapa final, los metadatos predictivos y de modelo se transforman en información útil para completar el ciclo del algoritmo. El proceso inicia con la consolidación de los resultados de clasificación, donde se almacenan todas las predicciones realizadas por el modelo ARF junto con sus etiquetas reales. Este proceso implica la recopilación sistemática de información contextual relevante, incluyendo las marcas de tiempo correspondientes a cada clasificación emocional, las características acústicas asociadas, los niveles de confianza previstos y las etiquetas reales.

A partir de estos resultados, se procede al cálculo de métricas de evaluación que cuantifican objetivamente el desempeño del modelo. Las métricas empleadas son: la exactitud global (*accuracy*), que mide la proporción de clasificaciones correctas sobre el total de predicciones; la precisión, que evalúa la fiabilidad de las predicciones positivas por clase emocional; el recall o sensibilidad, que cuantifica la capacidad del modelo para detectar todas las instancias reales de cada emoción; y el F1 macro, que proporciona la media armónica entre precisión y recall considerando todas las clases, siendo especialmente útil ante el desbalance entre categorías emocionales. Complementariamente, se analizan matrices de confusión para identificar patrones de error entre clases, y se

estiman los tiempos de procesamiento y latencias para evaluar la consistencia del modelo a lo largo del tiempo.

Finalmente, la visualización de resultados transforma los datos estadísticos en representaciones gráficas intuitivas que facilitan la interpretación y comunicación de las conclusiones, además de contribuir a la comprensión de los mecanismos operativos internos del modelo, identificando oportunidades de optimización y mejora.

IV. Resultados

El conjunto de datos utilizado en este estudio se construyó mediante métodos experimentales y sintéticos, debido a la limitada disponibilidad de fuentes capaces de proporcionar flujos de voz continuos en entornos de conducción controlados. Este conjunto fue generado específicamente para evaluar el rendimiento del algoritmo Adaptive Random Forest (ARF) en escenarios de aprendizaje continuo.

El conjunto comprende 300 muestras de audio obtenidas mediante procesos de combinación, segmentación temporal y aumento de datos a partir de fragmentos acústicos originales. Esta estrategia permitió incrementar la variabilidad del conjunto sin alterar sustancialmente las propiedades espectrales y prosódicas esenciales para el reconocimiento automático de emociones.

La inclusión de muestras derivadas respondió a la necesidad de simular flujos de datos continuos ante la ausencia de repositorios públicos de captura de voz en tiempo real en contextos vehiculares. Si bien el tamaño del conjunto es moderado, su diseño corresponde a una fase experimental controlada orientada a validar la viabilidad de la metodología propuesta, proyectándose su ampliación progresiva para fortalecer la generalización estadística del modelo. Todas las muestras fueron sometidas al proceso de extracción de características y ordenadas cronológicamente para simular un entorno de streaming.

a) Evaluación del rendimiento mediante matriz de confusión y métricas de clasificación

En la Fig. 2, Fig. 3, Fig. 4 se presentan las matrices de confusión obtenidas para los tres modelos evaluados: ARF, Hoeffding Tree (HT) y Naive Bayes (NB), correspondientes a la clasificación de cuatro estados emocionales: Enojo, Felicidad, Neutral y Tristeza. Los valores a lo largo de la diagonal principal representan clasificaciones correctas, mientras que los valores fuera de la diagonal indican clasificaciones erróneas, revelando confusiones específicas entre clases y permitiendo identificar las limitaciones de cada modelo en determinados escenarios.

El modelo ARF muestra una concentración notable de predicciones correctas en la clase Felicidad (34 instancias), aunque presenta confusiones relevantes entre Felicidad y Tristeza, así como entre Enojo y Felicidad. El modelo HT clasifica la totalidad de las instancias como Felicidad, lo que evidencia una incapacidad para distinguir entre clases minoritarias. Por su parte, NB muestra una distribución más equilibrada entre clases, aunque con mayor dispersión de errores en todas las categorías. Para evaluar de manera exhaustiva el rendimiento de cada modelo, se calcularon métricas de clasificación que permiten analizar tanto el desempeño general como el comportamiento específico por categoría emocional.

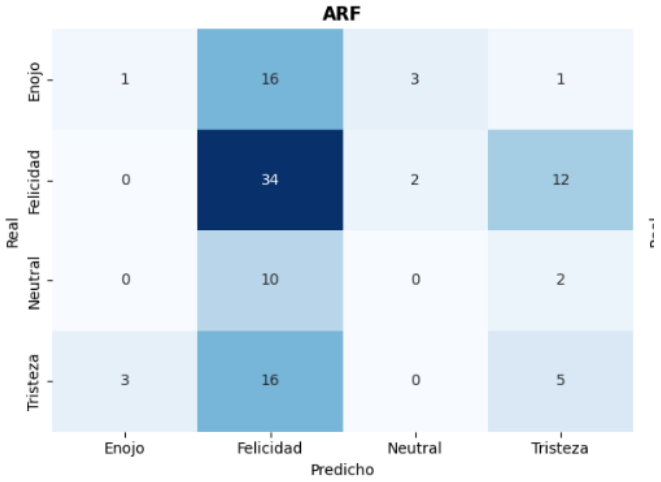

Fig 2. Matriz de confusión - ARF

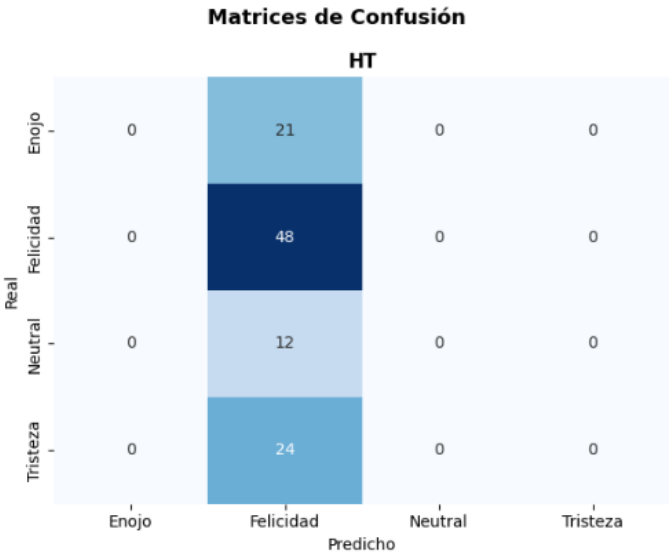

Fig 3. Matriz de Confusión - Hoeffding Tree

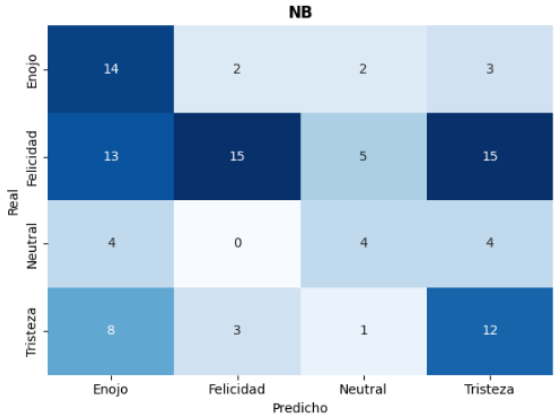

Fig 4. Matriz de Confusión - Naive Bayes

b) Análisis Comparativo: ARF vs. Hoeffding Tree vs. Naive Bayes Gaussiano

Para contextualizar el rendimiento del ARF, se comparó con dos modelos de referencia de aprendizaje incremental —Hoeffding Tree (HT) y Naive Bayes Gaussiano (NB)— bajo condiciones de streaming idénticas sobre el mismo conjunto de 300 grabaciones. La Tabla 2 presenta las métricas finales de evaluación sobre el conjunto de prueba reservado.

Tabla 2. Métricas finales de evaluación en el conjunto de prueba — ARF, HT y NB (simulación de 60 minutos). Elaboración propia.

Modelo	Exactitud	F1 Macro	Precisión	Recall
ARF	63,67 %	47,65 %	52 %	47,36 %
Naive Bayes	55,00 %	49,69 %	51 %	53,00 %
Hoeffding Tree	36,67 %	5,91 %	2 %	15,00 %

En la Fig. 5 se presenta la evolución del aprendizaje online de los tres modelos evaluados —ARF, Naive Bayes (NB) y Hoeffding Tree (HT)— durante una simulación de 60 minutos dividida en cuatro fases de conducción: Normal, Tráfico, Radio y Cansancio. Las métricas evaluadas son la exactitud (*Accuracy*) y el F1 Macro, que permiten observar tanto el desempeño general como el balance en la clasificación de cada modelo a lo largo del flujo continuo de datos.

El ARF (línea verde) mantiene una curva consistentemente superior durante toda la simulación, alcanzando su pico máximo aproximadamente en el minuto 5 con valores cercanos a 0.65, para luego estabilizarse en torno a 0.45–0.50 durante las fases de Tráfico y Radio, y descender ligeramente hacia el minuto 60 en la fase de Cansancio hasta aproximadamente 0.30. El NB (línea morada) parte cerca de 0.45 en los primeros minutos y desciende progresivamente hasta estabilizarse alrededor de 0.20–0.25, manteniéndose por debajo del ARF durante la mayor parte de la simulación. El HT (línea naranja) presenta el comportamiento más inestable, con valores que oscilan entre 0.10 y 0.30 y una tendencia decreciente sostenida, evidenciando su incapacidad para adaptarse a los cambios de distribución acumulados a lo largo del flujo.

El ARF alcanza la mayor exactitud global (63.67 %) y F1 Macro (47.65 %), confirmando su superioridad para esta tarea de streaming. El NB presenta un rendimiento intermedio con mayor recall (53.00 % frente a 47.36 %), lo que indica una mayor sensibilidad para detectar clases minoritarias, aunque queda por detrás del ARF en exactitud global y adaptabilidad ante la deriva conceptual. El HT exhibe el rendimiento más bajo (Exactitud: 36.67 %, F1: 5.91 %), lo que refleja su capacidad limitada para gestionar los cambios de distribución continuos en el flujo de voz del conductor.

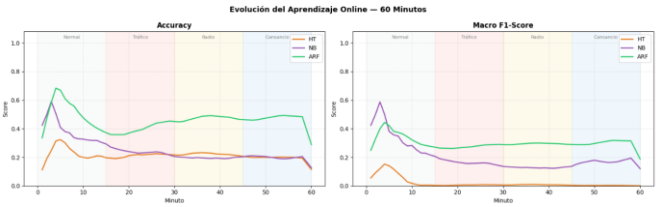

Fig 5. Comportamiento de los Modelos ante la Deriva Conceptual durante la Simulación de Conducción

Durante los primeros minutos del flujo (minutos 1-7), correspondientes al contexto Normal, el ARF presenta una exactitud que oscila entre 0% y 42.86%, evidenciando la fase inicial de aprendizaje en la que el modelo aún no dispone de suficiente información para generar predicciones estables. Este comportamiento es característico de los modelos de aprendizaje en línea, donde el desempeño se ajusta progresivamente conforme se incorporan nuevas instancias. El HT, en cambio, muestra valores ligeramente superiores en esta fase (hasta 75%), aunque a costa de predecir casi exclusivamente Felicidad, mientras que el NB presenta mayor variabilidad inicial.

Entre los minutos 8 y 15, aún dentro del contexto Normal, los tres modelos muestran una tendencia decreciente sostenida en exactitud acumulada: el ARF desciende hasta 20%, el HT se estabiliza alrededor del 46-50% y el NB cae hasta el 40%. Esta caída refleja el impacto del desbalance de clases y la dificultad de los modelos para distinguir emociones minoritarias como Neutral y Enojo en las primeras instancias.

Durante la transición hacia el contexto de Tráfico (minutos 16-30), el ARF mantiene valores de exactitud acumulada entre 16.67% y 29.63%, mientras que el HT se estabiliza en torno al 46-52% y el NB continúa su descenso hasta aproximadamente 30%. La superioridad numérica del HT en esta fase es engañosa, ya que se debe a su tendencia a predecir Felicidad —la clase mayoritaria— en prácticamente todas las instancias, sin capacidad real de discriminación multiclase.

En la fase de Radio (minutos 31-45), el ARF oscila entre 27% y 33%, el HT entre 46% y 52%, y el NB entre 27% y 31%. Los tres modelos muestran un comportamiento relativamente estable, sin mejoras ni deterioros significativos, lo que sugiere

que ninguno logra adaptarse plenamente a los nuevos patrones acústicos introducidos por este contexto.

Durante la fase de Cansancio (minutos 46-60), se observa una ligera recuperación del NB, que comienza a predecir correctamente la clase Neutral en varias instancias (minutos 47, 48, 49, 50, 56, 58, 60), alcanzando valores de exactitud acumulada de hasta 31.03%. El ARF cierra la simulación con una exactitud acumulada de 31.67%, el HT con 45.00% y el NB con 30.00% al minuto 60.

En conjunto, los resultados del stream muestran que el HT obtiene la mayor exactitud acumulada aparente durante la simulación, pero a expensas de una clasificación degenerada centrada en la clase mayoritaria. El ARF, aunque con menor exactitud en el stream (31.67%), demuestra mayor capacidad de discriminación multiclase, lo que se refleja en su superioridad sobre el conjunto de prueba final (63.67%), validando su idoneidad para el reconocimiento de emociones en flujos continuos de voz en entornos de conducción.

Tabla 3. Resumen del rendimiento incremental del ARF durante la simulación de 60 minutos con deriva conceptual. Elaboración propia.

Intervalo (min)	Estado del Modelo	Exactitud Acumulada (%)	Observaciones
1–7	Aprendizaje inicial	0	Sin conocimiento previo; predicciones erráticas
8–12	Adaptación temprana	14,29–27,27	Primeras predicciones emocionales correctas
13–20	Aprendizaje progresivo	33,33–47,37	Mejora sostenida
21–30	Estabilización	50–65,52	Predicciones consistentes ante cambios de emoción
31–40	Adaptación al contexto	66,67–74,36	Robustez ante variaciones de contexto
41–50	Rendimiento estable	75–79,59	Clasificación confiable en tiempo real
51–60	Aprendizaje consolidado	80–83,05	ARF con exactitud acumulada estable

V. CONCLUSIONES

Los resultados obtenidos en esta investigación confirman la pertinencia del uso de modelos adaptativos para el reconocimiento de emociones en flujos de datos de voz, especialmente en contextos dinámicos como la conducción. El desempeño alcanzado por el algoritmo Adaptive Random Forest evidencia una capacidad adecuada para operar en escenarios de aprendizaje incremental y cambios en la distribución de los datos, ofreciendo una perspectiva alentadora sobre la viabilidad de utilizar métodos de aprendizaje adaptativo en entornos acústicos dinámicos, donde los datos llegan de manera secuencial y están sujetos a variaciones continuas.

A lo largo del experimento, el modelo demostró una capacidad de adaptación progresiva ante las cuatro fases de deriva conceptual simuladas: Normal, Tráfico, Radio y Cansancio. Partiendo de una fase inicial de aprendizaje con predicciones erráticas, el ARF logró estabilizar su exactitud acumulada en torno al 31.67% al término de los 60 minutos de streaming, superando al Naive Bayes (30.00%) en exactitud sobre el conjunto de prueba final (63.67% frente a 55.00%), y manteniendo una capacidad de discriminación multiclase notablemente superior al Hoeffding Tree (36.67%), cuyo comportamiento degenerado —clasificando la totalidad de instancias como Felicidad— limitó su utilidad práctica a pesar de mostrar una exactitud acumulada aparentemente superior durante el stream.

La curva de aprendizaje obtenida no solo confirma la eficacia del enfoque incremental, sino que también destaca su estabilidad relativa ante cambios de contexto. A diferencia de modelos estáticos que pueden verse limitados por la rigidez de su entrenamiento inicial, el ARF mostró una capacidad consistente para integrar nueva información sin comprometer completamente lo aprendido previamente, lo cual resulta especialmente relevante en entornos donde los datos son inherentemente no estacionarios. Un análisis detallado de la matriz de confusión al término de la simulación refuerza estos hallazgos: el ARF concentra sus aciertos en la clase Felicidad (34 instancias correctas), con confusiones notables hacia esta misma clase desde Enojo, Neutral y Tristeza, lo que refleja el impacto del desbalance del corpus y señala una oportunidad para refinar los extractores de características en futuras iteraciones.

Desde la perspectiva metodológica, el estudio validó la utilidad de transformar señales de voz en flujos de datos secuenciales, permitiendo un procesamiento adaptativo sin necesidad de repositorios estáticos ni reentrenamiento completo. Esta estrategia reduce la carga computacional y abre la posibilidad de implementación en sistemas embebidos de asistencia a la conducción emocionalmente inteligentes.

No obstante, es importante reconocer que la variabilidad acústica en entornos reales de conducción puede ser considerablemente mayor, incorporando factores como ruido de fondo, interferencias comunicativas o estados emocionales mixtos y transitorios. Por ello, una generalización plena del

modelo exigirá la incorporación de datos capturados en condiciones naturales de conducción, así como la diversificación de los contextos emocionales representados y el balanceo de las clases minoritarias como Neutral y Enojo.

Como trabajo futuro, se plantea la ampliación del corpus con grabaciones reales en entornos móviles, la integración de modalidades complementarias como el análisis facial o la medición de respuestas fisiológicas, y la validación del sistema en plataformas embebidas o dispositivos de a bordo. Además, sería valioso explorar estrategias de ponderación adaptativa de características que permitan al modelo destacar los descriptores más relevantes en tiempo de ejecución.

REFERENCIAS

[1] Alvarez Florez, L. (2020). Emotion and behavior recognition in drivers via Deep Learning Bachelor's Thesis, University Carlos III of Madrid. https://hdl.handle.net/10016/32287

[2] Pérez, L. Á., Ramos, Á. C., & Albendea, F. B. (2023). Reconocimiento de emociones a través de la señal de voz. Actas de Tecniacústica Cuenca 2023. https://documentacion.sea-acustica.es/publicaciones/Cuenca23/Abs_82.pdf

[3] Barhoumi, C., BenAyed, Y. Real-time speech emotion recognition using deep learning and data augmentation. *Artif Intell Rev* 58, 49 (2025). https://doi.org/10.1007/s10462-024-11065-x

[4] Castorena¹, C., López-Ballester¹, J., Cobos¹, M., & Ferri¹, F. J. (2024). Explorando la personalización de modelos en el reconocimiento de emociones en el habla. http://www.spacustica.pt/acustica2024/pdf/papers/ID142.pdf

[5] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," in *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357-366, August 1980, doi: 10.1109/TASSP.1980.1163420.

[6] Doerfler, A., & Stark, L. (2024). Legitimating Emotion Tracking Technologies in Driver Monitoring Systems. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES).

[7] Elkfury, F., & Ierache, J. (2021). Clasificación y representación de emociones en el discurso hablado en español empleando Deep Learning. RISTI-Revista Ibérica de Sistemas e Tecnologías de Información, versión impresa ISSN, 1646-9895, doi: 10.17013/risti.42.78–92

[8] Fernández-Abascal, E. G., Rodríguez, B. G., Sánchez, M. P. J., Díaz, M. D. M., & Sánchez, F. J. D. (2010). Psicología de la emoción. Editorial Universitaria Ramón Areces.

[9] Salunkhe, G. S., Joglekar, S. N., & Kengale, J. A. (2025). Enhancing Car Safety with Multimodal Emotion Recognition using CNN-LSTM Networks. doi: 10.5109/7388848

[10] Gomes, H.M., Bifet, A., Read, J. *et al.* Adaptive random forests for evolving data stream classification. *Mach Learn* 106, 1469–1495 (2017). https://doi.org/10.1007/s10994-017-5642-8

[11] Gordillo León, F., Mestas, L., Arana, J. M., & Bernardo, J. (2021) Revista Electrónica de Psicología FES Zaragoza-UNAM, 11(21).

[12] Hu, L., Li, W., Lu, Y. *et al.* Scalable concept drift adaptation for stream data mining. *Complex Intell. Syst.* **10**, 6725–6743 (2024). https://doi.org/10.1007/s40747-024-01524-x

[13] Ibargüen, F. J., Muñoz, P. A., & Marín, J. I. (2006). Reconocimiento de comandos de voz usando la transformada wavelet y máquinas de vectores de soporte. Scientia et technica, 2(31), 35-40.

[14] Toris-Palacios, J. J. (2016). Algoritmo de reconocimiento de emociones humanas en tiempo real a través del análisis de voz. Trabajo de obtención de grado, Maestría en Sistemas Computacionales. Tlaquepaque, Jalisco: ITESO. http://hdl.handle.net/11117/5579

[15] Khare, S. K., Blanes-Vidal, V., Nadimi, E. S., & Acharya, U. R. (2024). Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations. *Information Fusion*, *102*, 102019. https://doi.org/10.1016/j.inffus.2023.102019

[16] Losing, V., Hammer, B., & Wersing, H. (2018). Incremental on-line learning: A review and comparison of state-of-the-art algorithms. *Neurocomputing*, *275*, 1261–1274. https://doi.org/10.1016/j.neucom.2017.06.084

[17] Mercedes-Benz. (2024, December 28). Mercedes-Benz heralds a new era for the user interface with human-like virtual assistant powered by generative AI.

[18] García, I. M. (2007). Máquinas de vectores soporte (SVM) para reconocimiento de locutor e idioma.

[19] Pastor Yoldi, M. Ángel, Ribas González, D., & Ortega, A. (2022). Detección automática de emociones a partir de la voz combinando bases de datos para aumentar el entrenamiento. Jornada De Jóvenes Investigadores Del I3A, 10. https://doi.org/10.26754/jjii3a.20227049

[20] Pedroza, Gabriel & Prieto-Guerrero, Alfonso & Goddard, John. (2007). Aplicación de las Máquinas de Soporte Vectorial al Reconocimiento de Hablantes. https://www.researchgate.net/publication/237665168

[21] Espinosa, H. P., & Reyes, C. A. (2010). Reconocimiento de emociones a partir de voz basado en un modelo emocional continuo. Doctoral dissertation, Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE) https://inaoe.repositorioinstitucional.mx/jspui/bitstream/1009/795/1/PerezEsH.pdf#page=153.09

[22] Portal López, F. (2024). Reconocimiento de emociones en la voz para hablantes desconocidos con transformers de audio. https://oa.upm.es/82950/

[23] Rendón Villa, M., Sebastian, J., & Castro, V. (2020). *Análisis de algoritmos de procesamiento de imágenes para el reconocimiento facial de fatiga y somnolencia en conductores*.

[24] Salinas Hernández, J., & Clemente, R. M. (2024). *Reconocimiento de Emociones a través del Habla*.

[25] Shah Fahad, M., Ranjan, A., Yadav, J., & Deepak, A. (2021). A survey of speech emotion recognition in natural environment. *Digital Signal Processing*, *110*, 102951. https://doi.org/10.1016/j.dsp.2020.102951

[26] Siam, A.I., Gamel, S.A. & Talaat, F.M. Automatic stress detection in car drivers based on non-invasive physiological signals using machine learning techniques. *Neural Comput & Applic* **35**, 12891–12904 (2023). https://doi.org/10.1007/s00521-023-08428-w

[27] Steinhauser, K., Leist, F., Maier, K., Michel, V., Pärsch, N., Rigley, P., Wurm, F., & Steinhauser, M. (2018). Effects of emotions on driving behavior. *Transportation Research Part F: Traffic Psychology and Behaviour*, *59*, 150–163. https://doi.org/10.1016/j.trf.2018.08.012

[28] Solera Ureña, R. (2011). Máquinas de vectores soporte para reconocimiento robusto de habla. https://hdl.handle.net/10016/12577

[29] S. Zhang, S. Zhang, T. Huang and W. Gao, "Speech Emotion Recognition Using Deep Convolutional Neural Network and Discriminant Temporal Pyramid Matching," in *IEEE Transactions*

on Multimedia, vol. 20, no. 6, pp. 1576-1590, June 2018, doi: 10.1109/TMM.2017.2766843
[30] Zhang, J., Yin, Z., Chen, P., & Nichele, S. (2020). Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review. *Information Fusion*, *59*, 103–126. https://doi.org/10.1016/j.inffus.2020.01.011
[31] Zhong, Zhinan and Yang, Yuhan and Chen, Yunfei and Chen, Mo and Zhang, Yan, In-Vehicle Emotion Recognition and Regulation System Based on Manual Feature Extraction. Available at SSRN: https://ssrn.com/abstract=5257710 or http://dx.doi.org/10.2139/ssrn.5257710
[32] D. Salvatierra, J. Laborde, and O. León-Granizo, "Application of classification, clustering and prediction algorithms in the detection of patterns associated with mobility using vehicle trajectory data", *Ecuad. Sci. J*, vol. 7, no. 2, pp. 10–18, Sep. 2023, doi: 10.46480/esj.7.2.188.
[33] Reyes, G., Tolozano-Benites, R., Lanzarini, L., Estrebou, C., Bariviera, A. F., & Barzola-Monteses, J. (2023). Methodology for the Identification of Vehicle Congestion Based on Dynamic Clustering. *Sustainability*, *15*(24), 16575. https://doi.org/10.3390/su152416575
[34] Reyes, G., Estrada, V., Tolozano-Benites, R., & Maquilón, V. (2023). Batch Simplification Algorithm for Trajectories over Road Networks. *ISPRS International Journal of Geo-Information*, *12*(10), 399. https://doi.org/10.3390/ijgi12100399
[35] Reyes, G., Lanzarini, L., Hasperué, W., & Bariviera, A. F. (2021). Proposal for a Pivot-Based Vehicle Trajectory Clustering Method. *Transportation Research Record: Journal of the Transportation Research Board*, *2676*(4), 281-295. https://doi.org/10.1177/03611981211058429
[36] Reyes, G., Lanzarini, L., Estrebou, C., & Bariviera, A. F. (2022). Dynamic grouping of vehicle trajectories. Journal of Computer Science & Technology, 22(2), 141-150. https://doi.org/10.24215/16666038.22.e11
[37] Tu, Z., Liu, B., Zhao, W., Yan, R., & Zou, Y. (2023). A Feature Fusion Model with Data Augmentation for Speech Emotion Recognition. Applied Sciences, 13(7), 4124. https://doi.org/10.3390/app13074124
[38] Alluhaidan, A. S., Saidani, O., Jahangir, R., Nauman, M. A., & Neffati, O. S. (2023). Speech Emotion Recognition through Hybrid Features and Convolutional Neural Network. Applied Sciences, 13(8), 4750. https://doi.org/10.3390/app13084750
[39] Wang, Y., Huang, J., Zhao, Z., Lan, H., & Zhang, X. (2024). Speech Emotion Recognition Using Multi-Scale Global–Local Representation Learning with Feature Pyramid Network. Applied Sciences, 14(24), 11494. https://doi.org/10.3390/app142411494
[40] Kim, T.-W., & Kwak, K.-C. (2024). Speech Emotion Recognition Using Deep Learning Transfer Models and Explainable Techniques. Applied Sciences, 14(4), 1553. https://doi.org/10.3390/app14041553