INSTITUTO POLITÉCNICO NACIONAL

CENTRO DE INVESTIGACIÓN EN COMPUTACIÓN

Detección de emociones usando inteligencia artificial para la recomendación personalizada de bebidas

TESIS

PARA OBTENER EL GRADO DE

MAESTRÍA EN CIENCIAS DE LA COMPUTACIÓN

PRESENTA:

Lic. Maraya Nataly Briones Pérez

DIRECTORES DE TESIS:

Dr. Marco Antonio Moreno Armendáriz

Dr. Francisco Hiram Calvo Castro

CIUDAD DE MÉXICO DICIEMBRE 2022

SIP-13
REP 2017

INSTITUTO POLITÉCNICO NACIONAL
SECRETARIA DE INVESTIGACIÓN Y POSGRADO

ACTA DE REGISTRO DE TEMA DE TESIS
Y DESIGNACIÓN DE DIRECTOR DE TESIS

Ciudad de México, a 29 de 11 del 2022

El Colegio de Profesores de Posgrado del **Centro de Investigación en Computación** en su Sesión
(Unidad Académica)

Extraordinaria No 16 celebrada el día 25 del mes octubre de 2022, conoció la solicitud presentada por el (la) alumno (a):

Apellido Paterno:	BRIONES	Apellido Materno:	PÉREZ	Nombre (s):	MARAYA NATALY

Número de registro: A 2 1 0 2 3 1

del Programa Académico de Posgrado: **Maestría en Ciencias de la Computación**

Referente al registro de su tema de tesis; acordando lo siguiente:

1.- Se designa al aspirante el tema de tesis titulado:

"Detección de emociones usando inteligencia artificial para la recomendación personalizada de bebidas"

Objetivo general del trabajo de tesis:

Desarrollar un sistema que a partir de la imagen de una persona obtenga un conjunto de características que permitan realizar recomendaciones personalizadas de una bebida de té con tapioca.

2.- Se designa como Directores de Tesis a los profesores:

Director: **Dr. Marco Antonio Moreno Armendáriz** 2° Director: **Dr. Francisco Hiram Calvo Castro**

No aplica: ☐

3.- El Trabajo de investigación base para el desarrollo de la tesis será elaborado por el alumno en:

Centro de Investigación en Computación

que cuenta con los recursos e infraestructura necesarios.

4.- El interesado deberá asistir a los seminarios desarrollados en el área de adscripción del trabajo desde la fecha en que se suscribe la presente, hasta la aprobación de la versión completa de la tesis por parte de la Comisión Revisora correspondiente.

Director(a) de Tesis

Dr. Marco Antonio Moreno Armendáriz

Aspirante

C. Maraya Nataly Briones Pérez

2º Director de Tesis

Dr. Francisco Hiram Calvo Castro

Presidente del Colegio

Dr. Francisco Hiram Calvo Castro

IPN-CIC

SIP-14
REP 2017

INSTITUTO POLITÉCNICO NACIONAL

SECRETARÍA DE INVESTIGACIÓN Y POSGRADO

ACTA DE REVISIÓN DE TESIS

En la Ciudad de México siendo las 12:00 horas del día 06 del mes de diciembre del 2022 se reunieron los miembros de la Comisión Revisora de la Tesis, designada por el Colegio de Profesores de Posgrado de: Centro de Investigación en Computación para examinar la tesis titulada:

"Detección de emociones usando inteligencia artificial para la recomendación personalizada de bebidas".

del (la) alumno (a):

Apellido Paterno:	BRIONES	Apellido Materno:	PÉREZ	Nombre (s):	MARAYA NATALY

Número de registro: A 2 1 0 2 3 1

Aspirante del Programa Académico de Posgrado: **Maestría en Ciencias de la Computación**

Una vez que se realizó un análisis de similitud de texto, utilizando el software antiplagio, se encontró que el trabajo de tesis tiene 02 % de similitud. **Se adjunta reporte de software utilizado**.

Después que esta Comisión revisó exhaustivamente el contenido, estructura, intención y ubicación de los textos de la tesis identificados como coincidentes con otros documentos, concluyó que en el presente trabajo **SI** ☐ **NO** ☒ **SE CONSTITUYE UN POSIBLE PLAGIO.**

JUSTIFICACIÓN DE LA CONCLUSIÓN: *(Por ejemplo, el % de similitud se localiza en metodologías adecuadamente referidas a fuente original)*
El porcentaje es muy bajo.

****Es responsabilidad del alumno como autor de la tesis la verificación antiplagio, y del Director o Directores de tesis el análisis del % de similitud para establecer el riesgo o la existencia de un posible plagio.**

Finalmente y posterior a la lectura, revisión individual, así como el análisis e intercambio de opiniones, los miembros de la Comisión manifestaron **APROBAR** ☒ **SUSPENDER** ☐ **NO APROBAR** ☐ la tesis por **UNANIMIDAD** ☒ o **MAYORÍA** ☐ en virtud de los motivos siguientes:
Cumple con los requisitos de una tesis de maestría.

COMISIÓN REVISORA DE TESIS

Dr. Marco Antonio Moreno Armendáriz
Director de Tesis

Dr. Grigori Sidorov

M. en C. José Eduardo Valdez Rodríguez

Dr. Francisco Hiram Calvo Castro
2° Director de Tesis

Dr. Miguel Jesús Torres Ruiz

Dr. Rolando Menchaca Méndez

Dr. Francisco Hiram Calvo Castro
PRESIDENTE DEL COLEGIO DE PROFESORES

DIRECCIÓN
IPN-CIC

INSTITUTO POLITÉCNICO NACIONAL

SECRETARÍA DE INVESTIGACIÓN Y POSGRADO

CARTA DE AUTORIZACIÓN DE USO DE OBRA PARA DIFUSIÓN

En la Ciudad de México el día 15 del mes de diciembre del año 2022, la que suscribe Maraya Nataly Briones Pérez alumna del programa de Maestría en Ciencias de la Computación con número de registro A210231, adscrita al Centro de Investigación en Computación manifiesta que es autora intelectual del presente trabajo de tesis bajo la dirección del Dr. Marco Antonio Moreno Armendáriz y el Dr. Francisco Hiram Calvo Castro cede los derechos del trabajo intitulado Detección de emociones usando inteligencia artificial para la recomendación personalizada de bebidas, al Instituto Politécnico Nacional, para su difusión con fines académicos y de investigación.

Los usuarios de la información no deben reproducir el contenido textual, gráficas o datos del trabajo sin el permiso expresado del autor y/o directores. Este puede ser obtenido escribiendo a la siguiente dirección de correo marayabp04@gmail.com. Si el permiso se otorga, al usuario deberá dar agradecimiento correspondiente y citar la fuente de este.

Maraya Nataly Briones Pérez

Resumen

Las expresiones faciales son una forma importante de interacción social entre los seres humanos. La construcción de un sistema capaz de reconocer automáticamente las expresiones faciales a partir de imágenes o vídeos ha sido un campo de estudio intenso en los últimos años. Existen dos clases de expresiones faciales: las macroexpresiones y las microexpresiones. La principal diferencia entre estas dos clases radica en su duración e intensidad. Las macroexpresiones suelen durar entre 0.5 y 4 segundos, mientras que las microexpresiones duran entre 0.065 y 0.5 segundos. Además, las macroexpresiones son expresiones voluntarias que abarcan el movimiento de una gran área facial, a diferencia de las microexpresiones que son involuntarias, y causan deformaciones localizadas causadas por la contracción involuntaria de los músculos faciales. Ekman es pionero en la relación entre emociones y expresiones faciales, determinando que existen emociones básicas universales, y proponen el *Facial Action Coding System (FACS)* que permite factorizar la composición de expresiones faciales en unidades de acción. La investigación sobre el reconocimiento de expresiones faciales permite una mayor conciencia y sensibilidad a los comportamientos faciales sutiles, y es un tema importante para la comprensión de las emociones humanas y los fenómenos afectivos. Un dispositivo de inteligencia artificial similar a un humano tiene una mayor aceptación cuando tiene la capacidad de mostrar empatía e interacción en relación con el consumidor humano. Debido a la importancia del entendimiento mutuo entre consumidores y agentes de servicios, la empatía es un constructor relevante en el marketing relacional y en la industria de servicios. Mediante el análisis de expresiones faciales se puede determinar el estado emocional de un individuo, y así, dar una respuesta empática para el usuario.

Abstract

Facial expressions are an important form of social interaction between humans. The construction of a system capable of automatically recognizing facial expressions from images or videos has been a field of intense study in recent years. There are two kinds of facial expressions: macroexpressions and microexpressions. The main difference between these two classes lies in their duration and intensity. Macro expressions usually last between 0.5 and 4 seconds, while micro expressions last between 0.065 and 0.5 seconds. In addition, macroexpressions are voluntary expressions that encompass the movement of a large facial area, unlike microexpressions which are involuntary, and cause localized deformations caused by involuntary contraction of facial muscles. Ekman pioneered the relationship between emotions and facial expressions, determining that there are universal basic emotions, and proposed the Facial Action Coding System (FACS), which allows factoring the composition of facial expressions into action units. Research on facial expression recognition enables greater awareness and sensitivity to subtle facial behaviors, and is an important topic for understanding human emotions and affective phenomena. A human-like artificial intelligence device has greater acceptance when it has the ability to show empathy and interaction in relation to the human consumer. Due to the importance of mutual understanding between consumers and service agents, empathy is a relevant construct in relationship marketing and the service industry. By analyzing facial expressions, it is possible to determine the emotional state of an individual, and thus, provide an empathetic response to the user.

Agradecimientos

A mi mamá y hermana, por su amor y apoyo incondicional,
por la confianza brindada y esas palabras de aliento.

A mis directores de tesis, por su apoyo y guía
durante el desarrollo de este trabajo.

A mis amigos, que siempre creyeron en mí.

Al CIC, por brindarme los conocimientos y herramientas
necesarias para el desarrollo de esta investigación.

A CONACyT, por su apoyo económico.

Índice general

Índice de tablas

Índice de figuras

Capítulo 1

Introducción

El análisis de emociones es un campo de investigación bastante amplio, existiendo diversas aportaciones científicas, desde la definición de la emoción hasta su origen neurológico. Entender como se comporta un ser humano siempre ha sido un tema de interés, y el estudio de las expresiones faciales es una forma de hacerlo.

Con ayuda de las expresiones faciales se puede determinar el estado emocional de una persona. Las investigaciones realizadas por el psicólogo Paul Ekman demostraron que las expresiones faciales son universales, es decir, son evolutivas y no tienen que ver con el entorno cultural. Teniendo en cuenta la relación existente entre las expresiones faciales y emociones, Ekman define seis emociones como básicas: ira, asco, felicidad, tristeza, miedo y sorpresa. De modo que, el reconocimiento automático de las expresiones faciales es la unión de las ciencias cognitivas, la neurología y la inteligencia artificial, permitiendo el desarrollo de sistemas con agentes empáticos o sistemas con ambientes empáticos, mejorando la relación humano-computadora.

La detección automática de expresiones faciales se puede dividir en dos partes: el reconocimiento de emociones o el reconocimiento de unidades de acción faciales. Para ello, se necesita del aprendizaje automático, la cual utiliza un conjunto de datos para entrenar a un clasificador. Considerando que la detección de expresiones faciales requiere un gran número de ejemplos para determinar las ca-

racterísticas necesarias para entrenar un clasificador, y este clasifique de manera correcta, el aprendizaje profundo es una mejor opción de aprendizaje automático.

McCulloch y Pitts (1943) propusieron el primer modelo matemático, dando origen al perceptrón (Rosenblatt, 1958). Mucho tiempo después, Fukushima propone las redes neuronales convolucionales, una variación del perceptrón multicapa, pero aplicado a matrices bidimensionales, y resultaron ser muy efectivas para tareas de visión artificial, como la clasificación de imágenes.

1.1 Justificación

El interés en mejorar la interacción social, ha llevado al desarrollo de diversas soluciones. La empatía se define como la habilidad de percibir y entender los estados emocionales de un individuo, y generalmente dar una respuesta en consecuencia de dichas emociones. Dado que las emociones se pueden detectar a través de las expresiones faciales cuando un individuo se comunica con otro cara a cara, el análisis de estas puede tener un efecto positivo en la interacción, por ejemplo, la experiencia del consumidor. En la actualidad existe un gran interés en la relación existente entre consumidor y producto, lo cual ha llevado al desarrollo de diversos estudios, aplicaciones, métodos, entre otros; capaces de entender, explicar o mejorar esta relación.

La detección de emociones ha sido un tema de gran interés, ya que en la comunicación la palabra solo tiene un 7 % de significado, el tono de voz un 38 % y las expresiones faciales un 55 %. Más aún, la detección automática del estado emocional de un individuo permite mejorar la relación entre agentes o servicios inteligentes con el consumidor humano.

1.2 Hipótesis

Las expresiones faciales son una forma fundamental de interacción social entre las personas. Comprender las expresiones faciales y por consiguiente las emociones permite mejorar la interacción social. Por lo tanto, se puede generar o dar una respuesta más asertiva con base a dicha emoción. Por ejemplo, la recomendación de productos alimenticios. De modo que se puede utilizar un modelo capaz de identificar la emoción de una persona para dar una respuesta más asertiva a sus necesidades, en este caso la recomendación de una bebida.

1.3 Objetivos

1.3.1 Objetivo general

Desarrollar un sistema que a partir de la imagen de una persona obtenga un conjunto de características que permitan realizar recomendaciones personalizadas de una de bebida de *bubble tea*.

1.3.2 Objetivos específicos

- Revisar trabajos relacionados con la clasificación de emociones.
- Seleccionar el/los corpus a utilizar.
- Preprocesar el corpus.
- Desarrollar un modelo capaz de detectar las emociones.
- Validar y evaluar el desempeño de los modelos propuestos.
- Proponer un recomendador de bebidas.

1.4 Contribuciones

En este trabajo se presenta una nueva forma de clasificar las emociones básicas de los individuos a partir del rostro, y posteriormente, dar una recomendación de una bebida de *buble tea*. El modelo propuesto permite la detección de las unidades de acción y luego la clasificación de la emoción básica, permitiendo ocupar la parte del modelo que solo detecta las unidades de acción con otros objetivo, y no para la clasificación de emociones.

1.5 Estructura de la tesis

Una visión general del contenido de cada capítulo se describe a continuación. El capítulo 2 se resumen algunos trabajos relacionados con temas de interés para esta investigación. Por consiguiente, el capítulo en cuestión contiene referencias al estado del arte de la clasificación de emociones, los conjuntos de datos desarrollados para el análisis de las expresiones faciales, investigaciones realizadas para recomendación de alimentos, entre otros. En el capítulo 3 se detalla el marco teórico. En este capítulo se da una breve introducción a la inteligencia artificial y se cubre la teoría detrás de los modelos utilizados para el desarrollo de la solución. El capítulo 4 explica la solución propuesta con cada una de sus etapas. En el capítulo 5 se presentan los experimentos realizados en cada una de las etapas de la solución propuesta y los resultados obtenidos. Finalmente, en el capítulo 6 se dan las conclusiones de esta investigación y se establece el trabajo futuro.

Capítulo 2

Estado del arte

Para comprender el panorama y entender el contexto relacionado con el objetivo que se busca alcanzar en este trabajo, en este capítulo presentamos detalladamente los trabajos relacionados con nuestra propuesta.

2.1 Emociones

El estudio de las emociones humanas de manera científica fue expuesta por primera vez por Darwin (1872), quien planteaba que algunas expresiones presentes en las emociones eran universales, y que además mostraban igualdades con expresiones homólogas a los animales. Calhoun y Solomon (1989) fueron los primeros filósofos interesados en ordenar las emociones, y proponen que el análisis de las emociones se hace a través de la expresión en la conducta. Una noción muy presente en la literatura de emociones, es que algunas suelen ser "especiales", o bien, las suelen llamar básicas, primarias o fundamentales. El psicólogo Plutchik (1980) elaboró un diagrama de emociones en forma de rueda, mostrado en la figura 2.1, de tal manera que las menos similares se encuentren en mutua oposición. Aquí se muestran cuatro ejes: *joy* - *sadness*, *anger* - *fear*, *anticipation* - *surprise*, *trust* - *disgust*.

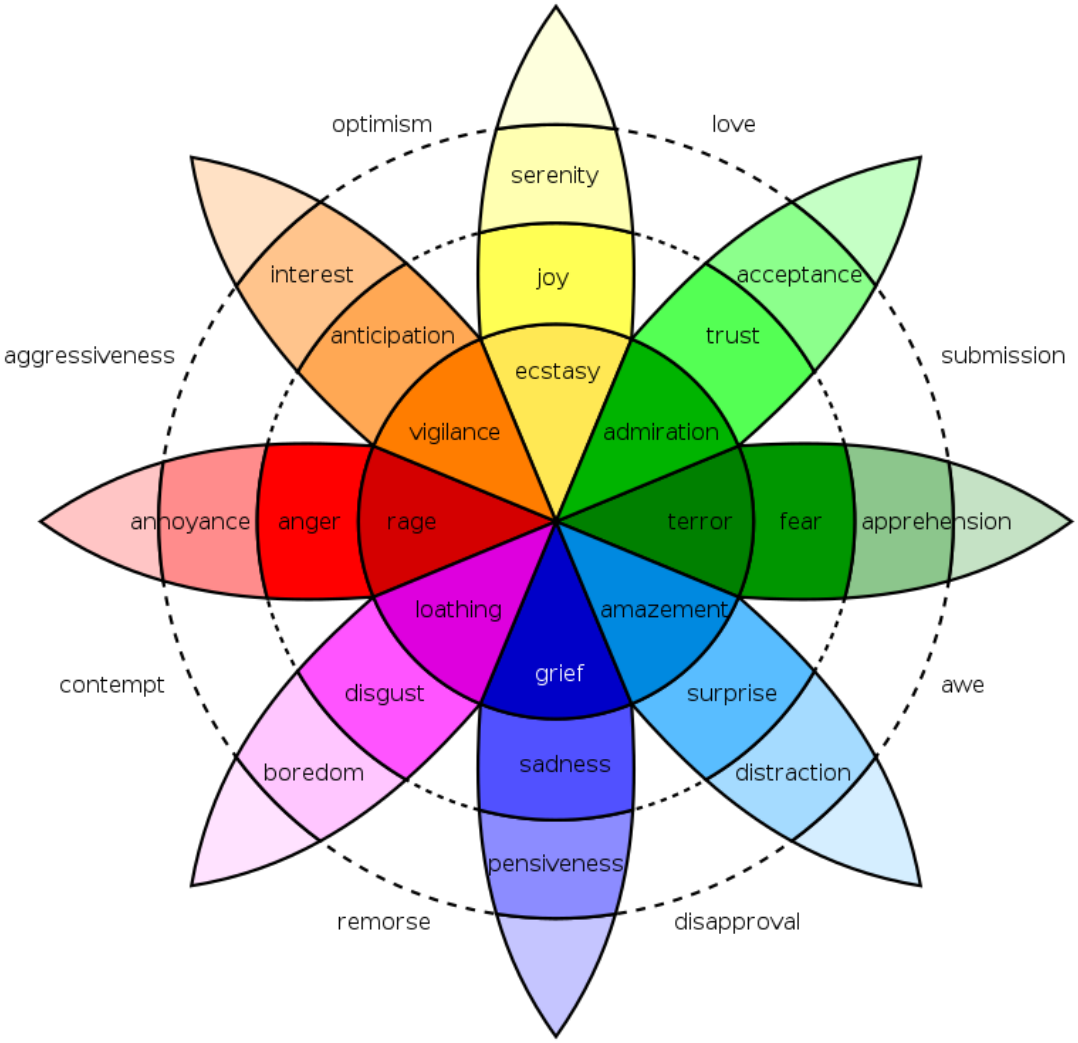

Figura 2.1: Rueda de las emociones propuesta por Robert Plutchik

Dentro de las teorías propuestas, la idea de la existencia de un conjunto de emociones básicas difería del número de emociones básicas, cuáles eran esas emociones básicas, y por qué eran básicas. Por ejemplo, Mowrer (1960) sugirió solo dos emociones básicas: el placer y el dolor. Panksepp (1982) mencionó en su teoría solo cuatro emociones básicas: esperanza, miedo, rabia y pánico; Kemper (1987) dijo que el miedo, la ira, la depresión y la satisfacción; Frijda (1986) identificó 18 emociones básicas, como la arrogancia, humildad, indiferencia, miedo, ira y tristeza, quien posteriormente solo recomendó 2 emociones básicas. Tomkins (1984a) creía en la existencia de nueve emociones básicas. La teoría más aceptada es la de Ekman

(1982), quienes proponen seis emociones básicas: ira, asco, miedo, felicidad, tristeza y sorpresa. Ekman es pionero dentro del estudio de las emociones y su relación con las expresiones faciales. Ekman realizó una investigación de campo en Nueva Guinea, llenado a la conclusión de que "las emociones se expresan a través del cuerpo y no son determinadas culturalmente; más bien son universales. Por ende, el origen de las emociones es biológico, tal como lo planteaba en su hipótesis Charles Darwin" (Ekman, 2003). Ekman llegó a esta conclusión estudiando cada parte del rostro con las que se ejecutaba la emoción. Ekman menciona que las emociones además de transmitir cambios en las expresiones faciales, también se pueden ver cambios en la voz y postura corporal. Por otra parte, indica que las expresiones emocionales duran más o menos dos segundos, aunque algunas suelen ser más breves, de apenas medio segundo, o bien, alargarse hasta 4 segundos. La duración de una expresión viene relacionada con la intensidad de la emoción. Además, descubrió que la cara puede hacer más de diez mil expresiones. Las investigaciones de Ekman demuestran que cada una de las seis emociones posee una expresión facial diferenciada y universal.

2.1.1 Expresiones faciales

Existen dos tipos de expresiones faciales: las macroexpresiones y las microexpresiones, las principales diferencias están en su duración e intensidad. Las macroexpresiones son voluntarias y suelen durar entre 0,5 y 4 segundos. Además, son realizados mediante movimientos faciales subyacentes que abarcan una área facial (Corneanu et al., 2016). En cambio, las microexpresiones son manifestaciones involuntarias, rápidas y locales, cuya duración suele ser de 0.065 y 0.5 segundo (Yan et al., 2013).

Los músculos faciales que producen las expresiones faciales son activados por núcleos nerviosos faciales, que a su vez son controlados por circuitos corticales y subcorticales de neuronas motoras superiores. Uno de los responsables de las expresiones faciales voluntarias es el circuito cortical. En cambio, el circuito subcortical es el responsable de las expresiones faciales involuntarias (Matsumoto & Hwang, 2011).

Las microexpresiones son deformaciones faciales localizadas, causadas por la contracción involuntaria de los músculos faciales (Ekman & Friesen, 1969). Las macroexpresiones implican más músculos en una zona facial más grande y la intensidad del movimiento muscular es también relativamente más fuerte. Por lo tanto, las microexpresiones en comparación con las macroexpresiones tienen una duración muy corta, una variación muy ligera y menos zonas de acción en los rasgos faciales externos (XiaoLan et al., 2010).

Ekman menciona que cuando se intenta eliminar todo signo de emoción, el resultado puede ser una microexpresión en la que la expresión se muestra brevemente. De igual forma, las microexpresiones pueden surgir cuando la inhibición de la expresión se da sin intervención de la conciencia, es decir, la persona no sabe conscientemente cómo se siente. Por lo tanto, las microexpresiones pueden ser expresiones plenas pero muy breves, o expresiones parciales y/o leves, muy breves. Por tal motivo, las microexpresiones son difíciles de percibir, pero con entrenamiento son fáciles de ser distinguidas. Además, Ekman sugiere que hay ciertos músculos faciales que no se pueden controlar conscientemente, a los que se refiere como músculos fiables; estos músculos sirven como indicadores sólidos de la aparición de emociones relacionadas (Ekman & O'Sullivan, 2006).

Como resultado de las investigaciones realizadas por Ekman, se creó un sistema denominado *Facial Action Coding System* (FACS), el cual permite factorizar la composición de las expresiones faciales (Ekman & Rosenberg, 2005). FACS es el esquema de codificación más utilizado para descomponer las expresiones faciales en movimientos musculares individuales, llamados *action units* (UA). Con el FACS, cada expresión facial posible puede describirse como una combinación de UA's. Hay 32 acciones relacionadas con los músculos faciales, y 6 descriptores de acción (ADs) (Wang et al., 2015). La figura 2.2 muestra algunos ejemplos de las expresiones faciales descritas por Ekman.

Uno de los problemas a los que se enfrenta FACS es la formación profesional de los expertos en codificación, dado que se requiere de mucho tiempo de

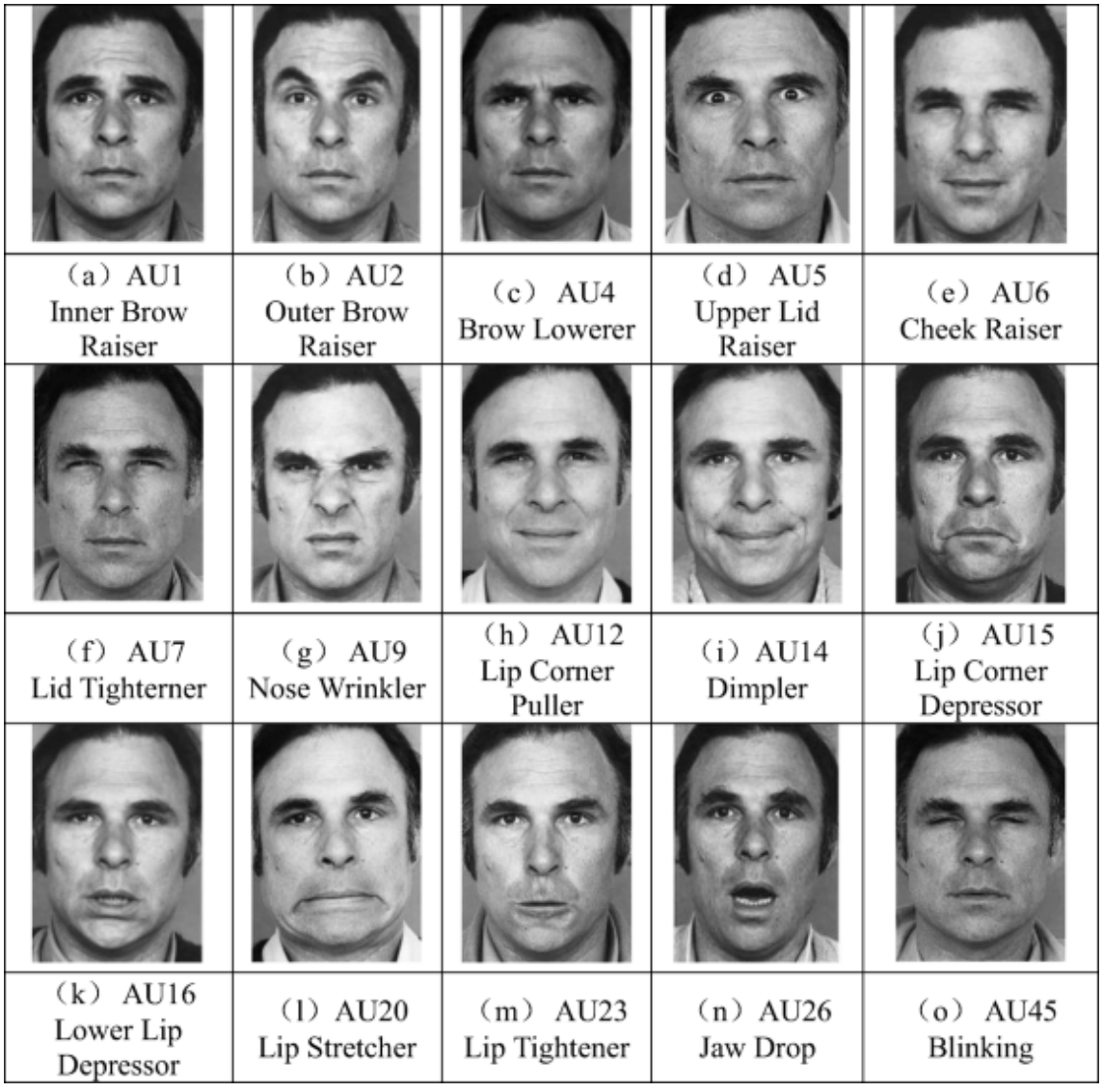

Figura 2.2: Ejemplos de expresiones faciales (Ekman, 2003)

entrenamiento, un aproximado de 100 horas de formación y, en práctica, el proceso de codificación tarda aproximadamente dos horas para codificar un video de 1 minuto (Pantic, 2009).

Debido a que las unidades de acción son descriptivas para determinar configuraciones faciales, se proponen sistemas específicos para explorar la relación entre los movimientos de los músculos faciales (UA's) y las emociones humanas. Por ejemplo, *Emotional Facial Action Coding System* (EMFACS-7) (Renneberg et al., 2005), *Facial Action Coding System affect Interpretation Dictionary* (FACSAID) (Ekman & Friesen, 2002) y *System for Identifying Affect Expressions by Holistic Judgments* (Affex) (Cairns et al., 1996). La tabla 2.1 muestra la relación ideal entre unidades de acción y emociones propuesta por Ekman y Friesen.

Tabla 2.1: Relación entre emociones y unidades de acción

Emoción	Unidades de acción
Ira	$4+5+7+23$
Desprecio	$12+14$
Asco	$9+15+16$
Miedo	$1+2+4+5+7+20+26$
Felicidad	$6+12$
Tristeza	$1+4+15$
Sorpresa	$1+2+5+26$

2.1.2 Reconocimiento de expresiones faciales

El reconocimiento de expresiones faciales puede dividirse de dos formas: reconocimiento de emociones y reconocimiento de unidades de acción faciales. De esta manera, el reconocimiento automático de las expresiones faciales es la intersección de las ciencias cognitivas, la neurología y la inteligencia artificial. Dicho esto, la capacidad de reconocer las expresiones faciales, es de gran utilidad y tiene diversas aplicaciones. Por ejemplo, la construcción de entornos adaptativos (Maat & Pantic, 2007); sistemas con agentes empáticos (DeVault et al., 2014); agentes con habilidades sociales, como los robots AIBO y Robavie (Ishiguro et al., 2001). Kapoor et al. (2007) muestra como la detección de la frustración en estudiantes puede mejorar

la experiencia del aprendizaje remoto. Otro ejemplo es la detección de dolor, que puede permitir controlar la evolución de pacientes en entornos clínicos (Kaltwang et al., 2012). El análisis de los malestares psicológicos puede ayudarse a partir del reconocimiento de la depresión (Joshi et al., 2012). Por último, han desarrollado aplicaciones comerciales como *Kairos* [1] o *Affectiva* [2], que realizan evaluaciones a gran escala en internet, con el objetivo de obtener información de las reacciones que tienen los espectadores a anuncios o material relacionado, con el fin de predecir un comportamiento de compra.

Reconocimiento de emociones

Para el reconocimiento de emociones, la mayoría de aportaciones está enfoca en cinco o seis emociones básicas, dado que como se había mencionado anteriormente, estas son universales. Pero también existen aportaciones que intentan detectar estados afectivos, como el dolor (Littlewort et al., 2007), la fatiga (Ji et al., 2006), incluso detectar estados emocionales como: interés, inseguridad, desacuerdo, concentrado, entre otras (Kapoor et al., 2007). Sajjanhar et al. (2018) presentan el reconocimiento de emociones con modelos de *deep learning*, más específicamente CNN o modelos pre-entrenados (VGG, *Inception-v3*), con dos preprocesamientos en las imágenes (región de interés, patrón binario local), obteniendo *accuracy* entre 51,65 % y 82,26 %. Quinn et al. (2017) exponen varios modelos capaces de reconocer siete emociones (alegría, tristeza, enfado, miedo, sorpresa, asco y neutro) mediante el conjunto FER13, obteniendo un *accuracy* de 45,95 % con una máquina de soporte vectorial y un 66,677 % con una CNN, asimismo un 98,4 % con el conjunto CK+. Por otra parte, Li y Xu (2020) menciona que el rendimiento de los métodos de aprendizaje profundo dependen mucho de la calidad de imágenes con que se trabaje, por tal razón proponen aprendizaje por refuerzo para una preselección de imágenes favorables para la clasificación de emociones, y posteriormente, entrenar el clasificador de emociones. Este selector de imágenes muestra una mejora en el ren-

[1] www.kairos.com
[2] www.affectiva.com

dimiento de la clasificación de emociones, obteniendo un *accuracy* de 72,35 % con el conjunto de datos FER13. Zhang et al. (2018) diseñan una red multitarea capaz de aprender atributos como el sexo, la edad, postura de la cabeza y la expresión facial, obteniendo un *accuracy* del 73,73 % para el reconocimiento de emociones. Borgalli y Surve (2022) obtienen un *accuracy* de 86,78 % con el dataset FER13, entrenando la red por 10 *folds*.

Reconocimiento de unidades de acción

En el reconocimiento de unidades de acción, las investigaciones se enfocan en su intensidad y excitación. Cohn y Schmidt (2004) muestra como diferentes tipos de sonrisas espontáneas, difieren en amplitud y duración. También se ha demostrado que unidades de acción espontáneas de las cejas (UA1, UA2, UA4) poseen diferencias en su intensidad, duración y orden de ocurrencia, en comparación con las unidades de acción de cejas posadas (Valstar et al., 2006). Yang et al. (2019) proponen un método basado en el reconocimiento de unidades de acción que puede ejecutarse en un ordenar de bajo rendimiento y obtener buenos resultados. Este método esta divido en dos partes, la primera parte consiste en marcar 68 puntos de referencia y obtener la imagen histogramas de gradientes orientados para calcular las unidades de acción. La segunda parte consiste en utilizar diferentes métodos de clasificación para realizar el paso de UA al reconocimiento de expresiones faciales. Pu et al. (2015) trabajan con *random forest* para la clasificación de UA, cabe destacar que aquí no hacen la clasificación por una UA individual, si no, por subconjuntos de UA. Además, la detección de UA se ha centrado más en la extracción de características faciales afectivas, la relación entre las UA y detectar el momento en que se presentó la UA. Existen investigaciones donde se dividía la cara en regiones, en seguida se extraían las características de textura o geométricas, para finalmente clasificarlas. Zhao et al. (2015) entrenaron las redes para detectar UAs por regiones, con la ayuda de 49 puntos clave, que se encontraban en la boca, nariz y ojos. Posteriormente, extraían las características para la detección adecuada de UAs. Gudi et al. (2015) por medio de una CNN calcularon la intensidad de la detección de la UA. Por otra

parte, Zhao et al. (2016) subdivide el rostro en 8 regiones iguales, las cuales pasan a redes profundas para obtener las características por regiones.

2.1.3 Corpus

Existen diversos conjuntos de datos para expresiones faciales. Zhang et al. (2018) mencionan que en la investigación de la percepción de emociones o caras, se requieren conjuntos de datos bastante amplios, compuestos por diversas imágenes de expresiones faciales. La tabla 2.2 resume algunos de estos conjuntos de datos. Cabe resaltar que la diferencia entre conjuntos de datos con imágenes posadas o espontáneas, radica en la duración e intensidad de las unidades de acción que se activan para cada expresión facial.

Tabla 2.2: Conjuntos de datos de expresiones faciales más conocidos

Conjunto de datos	Cantidad	Tipo de expresiones
JAFFE (Lyons et al., 1999)	213 imágenes de 10 sujetos	Posadas
MMI (Pantic et al., 2005)	238 secuencias de imágenes de 28 sujetos	Posadas
CK+ (Lucey et al., 2010)	593 secuencias de imágenes de 123 sujetos	Posadas
FER (Goodfellow et al., 2013)	35,587 imágenes	Espontáneas
SFEW (Dhall et al., 2015)	1,635 imágenes	Espontáneas
FER+ (Courville et al., 2013)	23,744	Espontáneas
DISFA+ (Mavadati et al., 2016)	32,875 fotogramas de 9 sujetos	Posadas y espontáneas

2.1.4 Inteligencia emocional

Naranjo (2018) menciona que el término de "inteligencia emocional" es atribuido generalmente a Wayne Payne, en su tesis doctoral. Mayer y Salovey (1989) definen a la inteligencia emocional como la habilidad de percibir, asimilar, regular y comprender las emociones propias y las de terceros, promoviendo un crecimiento emocional e intelectual. Considerar la forma en que ha evolucionado el cerebro, se puede comprender el poder que ejercen las emociones sobre la parte intelectual de la mente. El tronco encefálico es la región más primitiva del cerebro, esta regula las funciones básicas vitales, como el metabolismo o la respiración. El neocórtex (cerebro pensante) emergió de este cerebro primitivo. La amígdala y el hipocampo

fueron piezas importantes del cerebro primitivo. La amígdala es la especializada de cuestiones emocionales, además, es considera una estructura límbica ligada a procesos de aprendizaje y memoria. Por otra parte, el hipocampo es fundamental para el registro de hechos puros. La amígdala reacciona emocionalmente de forma impulsiva y ansiosa, pero existe una parte del cerebro que se encarga de regular los impulsos emocionales. Dicho regulador cerebral se encarga de desconectar los impulsos de la amígdala y parece encontrarse en el extremo de una vía nerviosa del neocórtex (lóbulo prefrontal). Las existentes conexiones entre la amígdala y el neocórtex componen un centro de gestión entre pensamientos y emociones. Por tal motivo, las emociones forman un papel muy relevante para el ejercicio de la razón, guiando las decisiones y trabajando con la parte racional de la mente. La figura 2.3 muestra la anatomía cerebral.

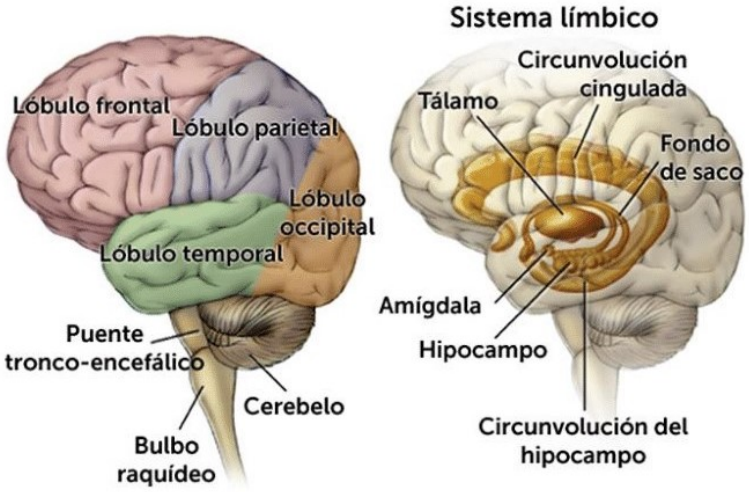

Figura 2.3: Anatomía cerebral

2.2 Recomendador

Un recomendador es un servicio o aplicación capaz de filtrar información personalizada, intentando adaptarse a los gustos del consumidor para realizar una sugerencia apropiada. Gutjar et al. (2015) menciona que la estimación del gusto del consumidor por productos alimentarios no se suele predecir con éxito en el mercado. Se

cree que las emociones evocadas por los alimentos pueden proporcionar una medida sensible para describir los productos que se eligen, además de las pruebas sensoriales. En su estudio miden las respuestas emocionales evocadas por los alimentos de los participantes por medio de las herramientas *PrEmo* y *EsSense*. También investigaron la relación entre las emociones provocadas por los alimentos, el gusto y el comportamiento para elegir dichos alimentos. Como resultados muestran que el gusto se asocia solo en parte con las respuestas emocionales, y que las elecciones de dicho producto estaban más relacionadas con las emociones positivas. Intentar medir la aceptabilidad por sí sola no es suficiente para el desarrollo y la prueba de productos en el mercado (King et al., 2010). Existen varios estudios que intentan relacionar las características sensoriales y las respuestas emocionales, comparando diversas categorías de productos (Cardello et al., 2012), o en la misma categoría, pero con diferentes productos, por ejemplo el chocolate negro (Thomson et al., 2010), los suavizantes (Porcherot et al., 2013) y calabazas de grosella (Ng et al., 2013). Samant et al. (2017) proponen un estudio donde intentan predecir el gusto y la preferencia del consumidor a partir de sus respuestas emocionales. Sus resultados mostraron que las respuestas emocionales autodeclaradas y el análisis de la expresión facial, más la intensidad percibida en las muestras, fueron las que mejor predijeron el gusto general de los participantes. Además, concluyeron que la combinación entre emociones evocadas y la percepción sensorial podría ayudar a comprender de mejor manera la preferencia del consumidor a soluciones con sabor básico.

Existen cinco sabores principales: umami, dulce, picante, ácido y amargo. Estudios han revelado que las soluciones con sabor dulce evocan emociones posotivas mientras que las soluciones con sabor salado evocan emociones negativas (Rousmans et al., 2000). O'Doherty et al. (2001) muestran que el consumir soluciones con sabor salado o dulce provoca activaciones neuronales en la amígdala, una parte del cerebro asociada al procesamiento emocional.

En Moreno-Armendáriz et al. (2021) se generó un vector carácteristico para cada bebida *bubble tea*, mediante el uso de *Coffee Taster's Flavor Wheel* que solo considera el sabor. La vectorización se llevó a cabo generando inicialmente el vector

característico de cada ingrediente. Consecuentemente, los vectores de los sabores básicos de cada bebida, se aplicó función *softmax* al vector resultante de la suma de todos los ingredientes para cada bebida.

Capítulo 3

Marco teórico

3.1 Emociones

La palabra emoción proviene del latín *emotio*, que significa "movimiento", "impulso".

La RAE (2021) define a la emoción como alteraciones del ánimo intensas y pasajeras, agradables o penosas, que están acompañadas de cierta conmoción somática.

Por otra parte, las emociones se entienden como el conjunto de respuestas orgánicas, fisiológicas, psicológicas o conductuales, que experimentan los individuos ante estímulos externos que permiten adaptarse a situaciones en relación con personas, objetos, lugares, etc. Entonces las emociones son una parte inherente de la vida humana y surgen de manera intensional o inconsciente. Además, son importantes para usar la razón. Entre sentir y pensar, las emociones guían nuestras decisiones, trabajando junto a la mente racional y habilitando o inhabilitando los pensamientos.

Según el psicólogo, Ekman (2003) las emociones son un proceso, un tipo especial de evaluación automática influenciada por nuestra evolución y nuestro pasado personal. Las palabras son solo un medio por el cual procesamos nuestras

emociones, las utilizamos cuando nos emocionamos, pero no podemos reducir las emociones a palabras.

3.1.1 Emociones básicas

Existen varios teóricos que concuerdan que existe un conjunto de emociones básicas, la tabla 3.1 muestra algunas de estas teorías.

Tabla 3.1: Algunos ejemplos de emociones básicas

Referencia	Emociones básicas
Arnold (1960)	Ira, aversión, coraje, abatimiento, deseo, desesperación, miedo, odio, esperanza, amor, tristeza
Plutchik (1980)	Aceptación, ira, anticipación, disgusto, alegría, miedo, tristeza, sorpresa
Ekman (1982)	Ira, asco, miedo, alegría, tristeza, sorpresa
Tomkins (1984b)	Ira, interés, desprecio, asco, angustia, miedo, alegría, vergüenza, sorpresa
Frijda (1986)	Deseo, felicidad, interés, sorpresa, asombro, tristeza

Como se puede observar, existe variaciones entre la clasificación de emociones básicas, esto puede deberse a cómo se etiqueta cada emoción, y a que algunas etiquetas suelen significar lo mismo. Además, la mayoría de clasificaciones propuestas suelen coincidir con la ira, la felicidad, el miedo y la tristeza.

La teoría más aceptada es la del psicólogo Paul Ekman, pionero en el estudio de las emociones y la relación con las expresiones faciales. Ekman realizó una investigación de campo en Nueva Guinea, determinando que "las emociones se expresan a través del cuerpo, y dichas expresiones no son determinadas culturalmente; más bien son universales. Por consiguiente, tienen origen biológico, tal como lo planteaba Charles Darwin en su hipótesis" (Ekman, 2003). Porlo tanto, Ekman clasificó las emociones básicas estudiando cada parte del rostro con las que se ejecutaba cada emoción. Por otra parte, menciona que las emociones producen cambios en la voz y la postura corporal, no solo en las expresiones faciales. Asimismo, nos indica que las expresiones emocionales duran aproximadamente dos segundos, aunque algunas suelen ser más breves, de apenas medio segundo, o bien, alargarse hasta 4 segundos. Esta duración viene relacionada con la intensidad de la emoción. Por lo tanto, las investigaciones de Ekman demostraron que el rostro es capaz de hacer más de diez mil expresiones faciales, y que cada una de las seis emociones básica propuestas por

él, poseen una expresión facial diferenciada y universal.

3.2 Inteligencia Artificial

El término inteligencia artificial (IA) fue mencionado por primera vez en 1956, por expertos como John McCarthy, Newell, Simon y Marvin Minsky durante una conferencia en Dartmouth.

Existe diversas definiciones sobre la inteligencia artificial, y no todas son aceptadas por los expertos. De forma simple, la inteligencia artificial hace referencia a máquinas o sistemas capaces de "imitar" la inteligencia humana para realizar tareas, intentando mejorar iterativamente a partir de la información recopilada.

3.2.1 Aprendizaje automático

El aprendizaje automático es una rama de la inteligencia artificial que tiene como objetivo capacitar a las computadoras a aprender, mediante el desarrollo de algoritmos aptos para reconocer patrones, y generalizar comportamientos haciendo uso de los datos proporcionados. Existen tres paradigmas de aprendizaje automático: aprendizaje supervisado, aprendizaje no supervisado y aprendizaje por refuerzo.

Aprendizaje supervisado

El aprendizaje supervisado requiere de un conjunto Y de entes denotados por y, que se relacionan con los patrones x (vectores, t-uplas o arreglos) del conjunto de patrones X. Las tareas de clasificación, regresión y recuperación necesitan del ente y para determinar el tipo de tarea a realizar en reconocimiento de patrones. Es decir, si el ente y es una etiqueta de clase, la tarea a realizar es de clasificación; si el ente y es un número real, la tarea será regresión; y si el ente y es un patrón, el tipo de tarea será de recuperación.

Aprendizaje no supervisado

Para el caso de aprendizaje no supervisado no existe el conjunto de entes Y. La tarea de clustering se asocia a este paradigma de aprendizaje automático. Dicha tarea se encarga de formar agrupamientos con los patrones $x \epsilon X$. Aquí los algoritmos son capaces aprender patrones sobre los datos proporcionados, aunque no estén etiquetados, calculando la similitud que tienen.

Aprendizaje por refuerzo

El paradigma de aprendizaje por refuerzo esta enfocado en proporcionar a los agentes inteligentes las herramientas necesarias para que sean capaces de aprender y escoger las acciones dentro del entorno en que se encuentren, con el fin de maximizar alguna noción de "recompensa" o premio.

3.2.2 Aprendizaje profundo

El aprendizaje profundo es un subconjunto del aprendizaje automático e inteligencia artificial. Los sistemas de aprendizaje profundo suelen mejorar el rendimiento, puesto que intentan simular el comportamiento del cerebro humano, lo cual permite que aprendan de una mayor cantidad de datos. Entonces los sistemas obtiene más experiencia cuando aprenden. La diferencia entre el aprendizaje profundo y el aprendizaje automático clásico tiene que ver con el tipo de datos con los que trabaja y los métodos con los que aprende. Por lo general, los algoritmos de aprendizaje automático clásico requieren de una etapa de preprocesamiento de los datos, en cambio, en el aprendizaje profundo no es necesario este preprocesamiento, ya que los algoritmos de aprendizaje profundo suelen automatizar la extracción de características necesarias para aprender. Además, procesos como propagación hacia atrás y descenso del gradiente, ayudan a ajustar los parámetros de forma automática, buscando mejorar la exactitud.

3.3 Redes neuronales tradicionales

Las redes neuronales artificiales están inspiradas en las redes neuronales del cerebro, diseñadas para modelar la forma en que el cerebro realiza una tarea.

La neurona biológica esta constituía por un cuerpo celular y un núcleo o soma, además de algunos elementos específicos como el axón, que es una ramificación de salida de la neurona donde se propagan impulsos electro-químicos. Por otra parte, la neurona cuenta con un gran número de ramificaciones de entrada, llamadas dendritas, cuyo objetivo es propagar la señal al interior de la neurona. La sinapsis recoge la información procedente de las neuronas vecinas a las que la neurona en cuestión está conectada, dicha información es procesada en el núcleo para generar una respuesta que será propagada por el axón a las dendritas de otras células a través de lo que se denomina sinapsis. La figura 3.1 muestra un diagrama simplificado de dos neuronas biológicas.

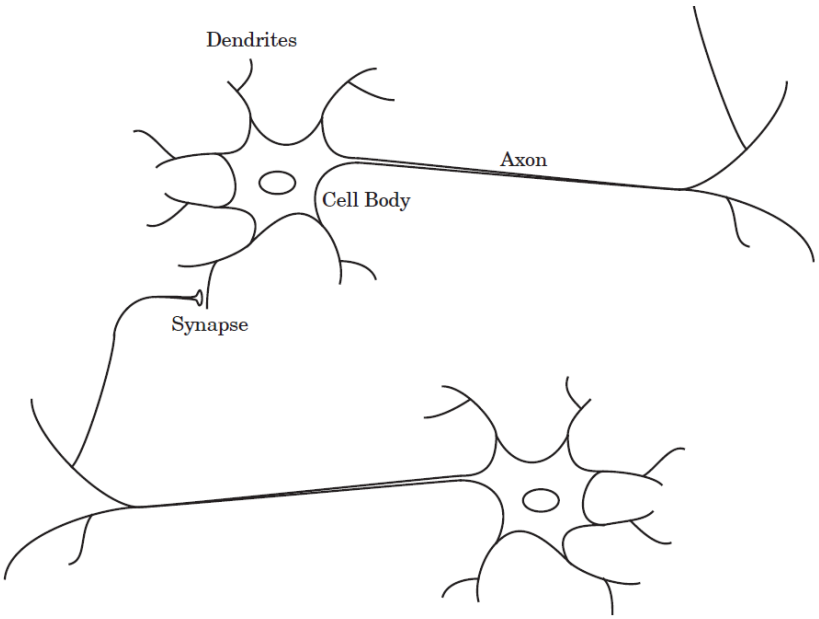

Figura 3.1: Neurona biológica

McCulloch y Pitts (1943) fueron quienes realizaron el primer modelo matemático de redes neuronales artificiales. El modelo propuesto está basado en la idea de que las neuronas operan por impulsos binarios, introduciendo así, una función de umbral de paso.

La neurona artificial, poseé un estado interno llamado nivel de activación, el cual recibe señales que permiten cambiar de estado. Dicho estado interno se denomina función de transición de estado o función de activación. Este nivel de activación depende de las entradas recibidas y de los valores sinápticos. Para hacer el cálculo del estado de activación, se debe de calcular en primer lugar la suma de todas las entradas ponderadas por ciertos valores. La figura 3.2 muestra esta idea. Obsérvese que un grupo de entradas $x_1, x_2, ..., x_n$ son introducidas a la neurona y multiplicadas por un peso asociado $w_1, w_2, ..., w_n$ para posteriormente sumarlos $\sum$. Estos pesos corresponden a la fuerza de conexión sináptica. De forma vectorial puede ser definido como:

$$\sum = X^T W = x_1 w_1 + x_2 w_2 + ... + x_n w_n \tag{3.3.1}$$

Posteriormente, las señales son procesadas por la función de activación F, produciendo la señal de salida de la neurona S. Existen diversos modelos, los cuales dependen de a función de activación F (Isasi & Galván, 2004); por ejemplo:

- Lineal: $S = KE$ con K constante.
- Umbral: $S = 1$ si $E \geq \theta$, $S = 0$ si $E \leq \theta$; siendo θ el umbral constante.
- Cualquier función: $S = F(I)$; siendo F una función cualquiera.

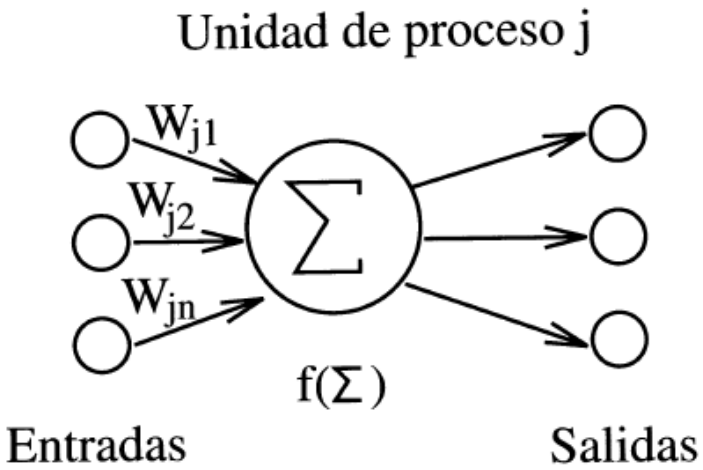

Figura 3.2: Esquema de una neurona artificial típica

3.3.1 Perceptrón

El modelo del perceptrón fue una aportación de Rosenblatt (1958), quien generalizó el modelo de células de McCulloch-Pitts añadiéndole aprendizaje. Su modelo consta de dos niveles donde se ajustan los pesos de las conexiones existentes entre los niveles de entrada y salida, dicho ajuste esta basado en el error entre la salida deseada y la salida obtenida. La figura 3.3 muestra un ejemplo, donde x_1, x_2 son las entradas, y la salida, w_1, w_2 los pesos. También se tiene un parámetro llamado umbral θ, el cual se utiliza como factor de comparación para producir la salida. La salida se puede representar en una sola ecuación:

$$y = F(\sum_{i=1}^{n} w_i x_i + \theta) \tag{3.3.2}$$

donde F ya no depende de ningún parámetro:

$$F(s) = \begin{cases} 1 & \text{si } s > 0 \\ -1 & \text{en caso contrario} \end{cases} \tag{3.3.3}$$

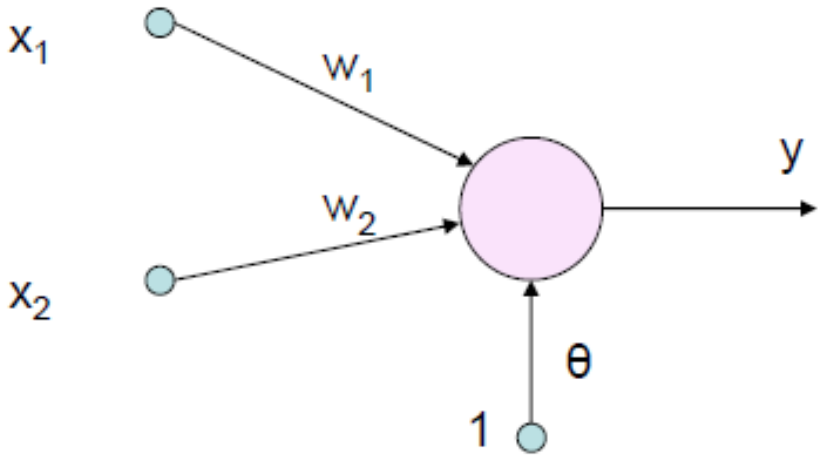

Figura 3.3: Arquitectura del perceptrón con dos entradas y una salida

En caso de que se clasifique de forma incorrecta, es necesario ajustar los pesos. El

ajuste de estos pesos esta dado por la regla de aprendizaje del perceptrón que se representa como sigue:

$$w^{new} = w^{old} + ep \tag{3.3.4}$$

donde w^{new} representa el nuevo peso, w^{old} el peso anterior, e el error de clasificación, obtenido de la diferencia entre el valor real de la etiqueta y el valor predicho, y p es la entrada a clasificada.

3.3.2 Perceptrón multicapa (MLP)

El perceptrón multicapa es la generalización del perceptrón simple. Los preceptrones multicapa tienen como habilidad aprender a partir de un conjunto de datos (entrenamiento), cuyo objetivo es modificar los pesos de la red con el fin de conseguir las salidas deseada. La arquitectura del perceptrón multicapa tiene como característica principal la agrupación de sus neuronas en capas de diferentes niveles. Cada capa está formada por un grupo de neuronas, los cuales se distinguen en tres tipos diferentes de capas: la capa de entrada, las capas ocultas y la capa de salida. La figura 3.4 representa dicha arquitectura.

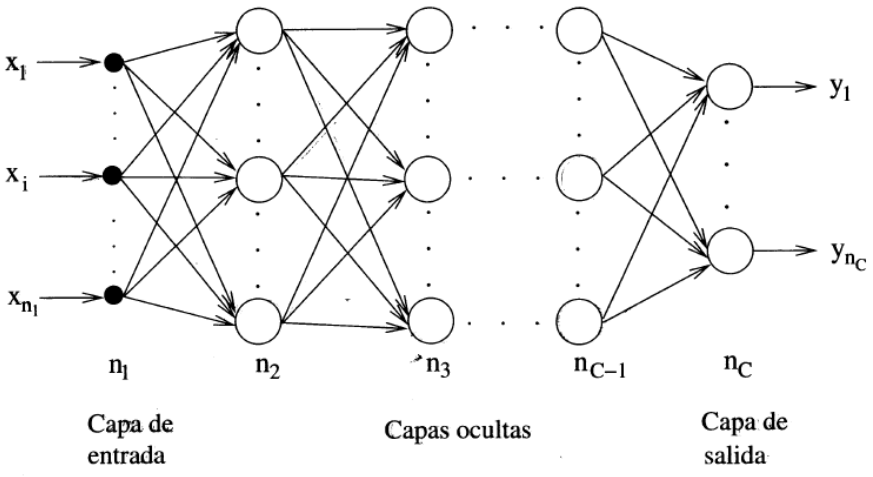

Figura 3.4: Arquitectura del perceptrón multicapa

Las neuronas de la capa de entrada son las encargadas de recibir los patrones externos y propagar hacia adelante dichas señales a través de las neuronas de la siguiente capa. La capa de salida proporciona la respuesta final para cada uno de los patrones de entrada. Como se puede ver en la figura 3.4, todas las conexiones del MLP están dirigidas hacia adelante, de ahí el nombre de redes alimentadas hacia adelante (*feedforward*). Cada conexión tiene asociado un peso y un umbral (comúnmente igual a 1). Generalmente, las neuronas de una capa se encuentran conectadas con todas las neuronas de la siguiente capa, es decir, las neuronas de la capa de entrada se encuentran conectadas con todas las neuronas de la primera capa oculta, las cuales, a su vez se encuentran conectadas a la siguiente capa, etc. Por lo cual existe una conectividad total, o bien, que el MLP se encuentra totalmente conectado.

El ajuste de los pesos en el MLP se puede realizar generalizando la regla de aprendizaje del perceptrón como sigue. Para actualizar el i-ésimo renglón de la matriz de pesos usamos:

$$\mathbf{w}_i^{new} = \mathbf{w}_i^{old} + e_i\mathbf{p} \tag{3.3.5}$$

Para actualizar el i-ésimo elemento del vector de bias usamos:

$$b_i^{new} = b_i^{old} + e_i \tag{3.3.6}$$

Por lo tanto, la regla de aprendizaje para el MLP puede escribirse de forma matricial como:

$$\mathbf{W}^{new} = \mathbf{W}^{old} + \mathbf{e}\mathbf{p}^T \tag{3.3.7}$$

$$\mathbf{b}^{new} = \mathbf{b}^{old} + \mathbf{e} \tag{3.3.8}$$

3.4 Redes neuronales profundas

Una red neuronal profunda contiene varias capas ocultas con millones de neuronas artificiales conectadas entre sí. Estas redes por lo general necesitan mayor tiempo para su entrenamiento, además de millones de ejemplos de datos.

3.4.1 Redes neuronales convolucionales

Las redes neuronales convolucionales (CNN) fueron creadas por Fukushima en 1980. Las neuronas de las CNN corresponden a los campos receptivos equivalentemente a las neuronas de la corteza visual primaria (V1) del cerebro biológico [1]. Esta red se puede ver como una variación de MLP, solo que su aplicación es realizada en matrices bidimensionales. Las CNN suelen ser muy efectivas a las tareas de visión artificial, como la clasificación y/o segmentación de imágenes, entre otras aplicaciones. Las CNN están compuestas por varias capas ocultas especializadas, las cuales poseen una jerarquía, es decir, las primeras capas detectan propiedades como formas básicas como líneas o curvas, y posteriormente se van especializando de forma que en las capas más profundas se pueden identificar formas más complejas como siluetas o rostros.

Inicialmente, la red toma como entrara los píxeles de una imagen, dicha imagen se representa como una matriz de píxeles con valores entre 0 y 255. Ahora bien, si se tiene una imagen de $48x48$ píxeles de alto y ancho, esta equivaldría a $2,304$ neuronas en el caso de que la imagen de entrada este en un solo color (escala de grises). En caso de tener imágenes a color, es necesario tres canales (rojo, verde, azul), por lo que se estarían usando $48x48x3 = 6,912$ neuronas de entrada. Previo a alimentar la red, es necesario normalizar los datos, es decir, que los valores estén entre 0 y 1, para ello se divide los valores de cada pixel por 255.

[1] https://www.juanbarrios.com/redes-neurales-convolucionales/

Filtro: conjunto de (*kernels*)

Los filtros en las CNN permiten extraer algunas características importantes de la imagen, o bien, patrones en esta. Algunas características relevantes a detectar son los bordes, el enfoque, el desenfoque, etc.

Convolución

Las convoluciones consisten en tomar un grupo de píxeles de la imagen de entrada y multiplicarlos por un kernel. La convolución está definida por la ecuación siguiente:

$$s(t) = \sum_{a} I(t+a) \cdot K(a) \tag{3.4.1}$$

donde I representa los datos de entrada y K el kernel. Cada posición del kernel estará representado por un índice a, y t indica el desplazamiento en los datos de entrada I. Un ejemplo de la operación de convolución en una dimensión se muestra en la figura 3.5.

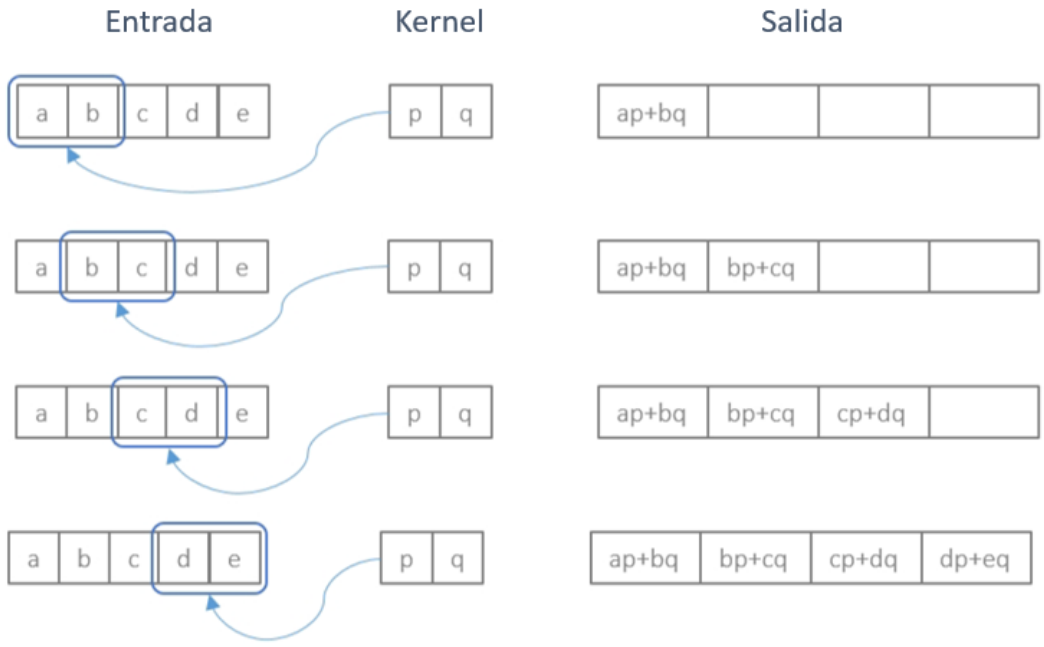

Figura 3.5: Ejemplo de convolución en 1-dimensión

La operación de convolución se puede extender a más dimensiones, por lo que la ecuación en dos dimensiones vendría siendo:

$$s(t) = \sum_a \sum_b I(m+a, n+b) \cdot K(a,b) \tag{3.4.2}$$

donde a y b son los índices de los filtros, m y n representan el desplazamiento en los datos de entrada, como se muestra en la figura 3.6.

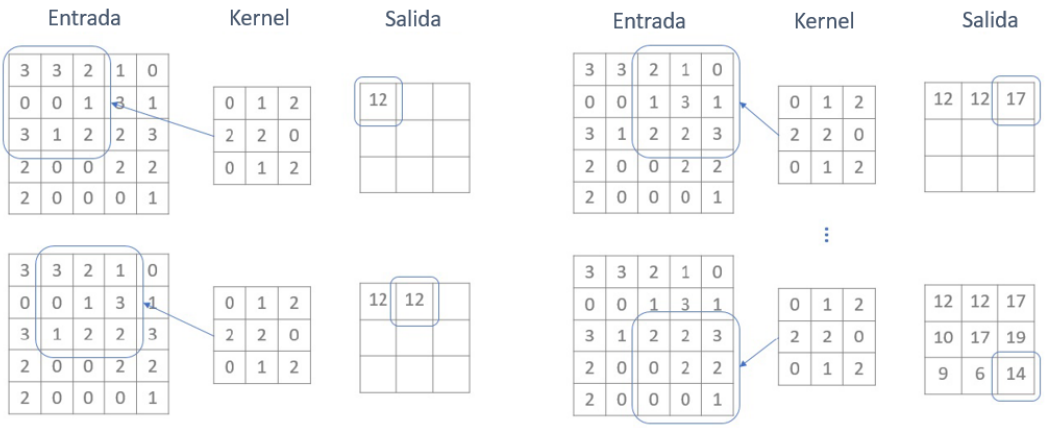

Figura 3.6: Ejemplo de convolución en 2-dimensiones

Para el caso de imágenes a color, dado que se representan con tres canales (rojo, verde, azul) que al combinarse se ven todas la mezclas de colores, el kernel debe tener de igual forma tres canales (3 capas). La salida obtenida en la convolución suele denominarse *feature map*.

Función de activación

Existen diversas funciones de activación, la cuales pueden ser lineales o no lineales. La función de activación se elige para satisfacer alguna especificación del problema que la neurona intenta resolver.

ReLu

La *ReLu* por sus siglas en inglés *Rectified Linear Unit*, devuelve 0 en caso de recibir una entrada negativa, y dará salida al mismo valor de entrada en caso de un valor positivo. Se puede describir por la ecuación siguiente:

$$f(x) = max(0, x) \tag{3.4.3}$$

Sigmoid

La función sigmoidea esta representada por la siguiente ecuación. Esta función toma valores reales y emite valores entre 0 y 1.

$$f(x) = \frac{(1)}{(1 + e^{-x})} \tag{3.4.4}$$

Tanh

La función de activación tangente hiperbólica es similar a la sigmoidea. Esta función toma cualquier valor real y emite valores entre -1 y 1. Se puede describir por la ecuación siguiente.

$$f(x) = \frac{(e^{x} - e^{-x})}{(e^{x} + e^{-x})} \tag{3.4.5}$$

Padding

El *padding* es una operación que se ocupa en las CNN, el cual agrega píxeles de valor cero alrededor de la imagen original, con el fin de evitar reducir la dimensión por el kernel en la convolución, además permite conservar información relevante que se llegará a encontrar en las esquinas de la imagen. La figura 3.7 muestra un ejemplo de un kernel $3x3$ aplicado a una entrada con *padding*.

Figura 3.7: Convolución con kernel $3x3$, *padded*

Función de agrupación: *Maxpooling*

La función de agrupación sustituye la salida de una capa por un resumen estadístico de las salidas cercanas, deslizando una ventana a través de la entrada. Un ejemplo es la función *maxpooling*, la cual produce una salida máxima dentro de un vecindario rectangular, dicha función se puede observar en la figura 3.8. El objetivo de la función de agrupación es reducir el tamaño de la próxima capa, pero conservando las características más importantes que el filtro detectó.

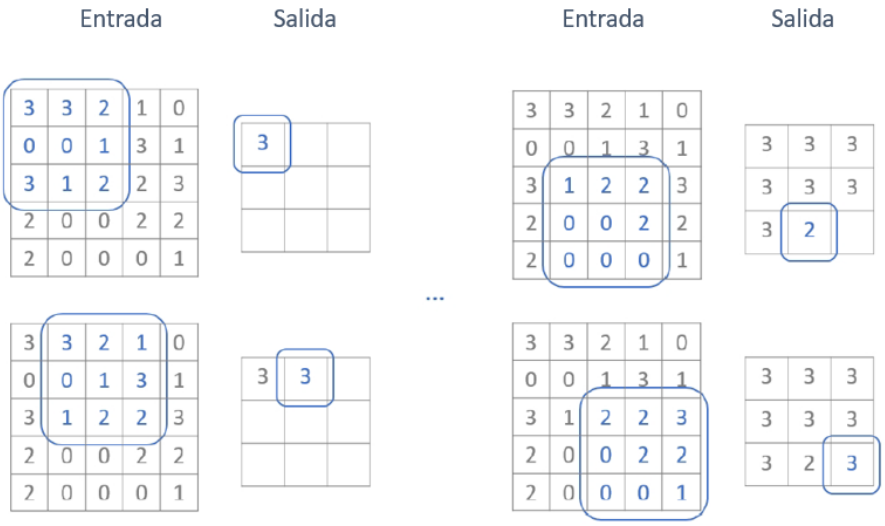

Figura 3.8: *Maxpooling* con ventana $3x3$

Backpropagation

Backpropagation es un método de cálculo que se aplica a un ciclo de propagación. Es decir, al ingresar un patrón a la red neuronal, este se propaga desde la primera capa a las capas siguientes de red, obteniendo una salida. La salida se compara con la salida esperada, calculando el error para cada salida obtenida. Mediante el procesamiento de ajuste hacia adelante y hacia atrás, se va mejorando el valor de los pesos de las conexiones entre las capas de las neuronas, y en cada ciclo se van ajustando los pesos hasta obtener los óptimos.

La figura 3.9 es un esquema general de todas las operaciones y características previamente mencionadas de las redes neuronales convolucionales [2]. Por lo general, las capas convolucionales son una alternativa para mejorar el rendimiento de una red completamente conectada, por lo que generalmente encontramos capas densas que reciben como entrada el mapa de características obtenido en las capas previas para realizar una tarea de clasificación.

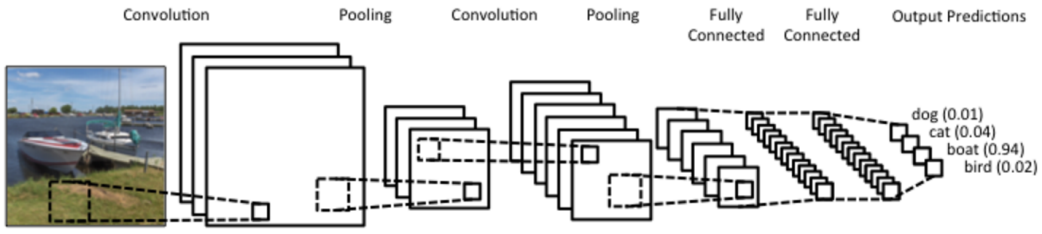

Figura 3.9: CNN

[2]https://towardsdatascience.com/build-your-own-convolution-neural-network-in-5-mins-4217c2cf964f

Capítulo 4

Descripción general de nuestra propuesta

En la sección 3.1 se presentaron algunas teorías de las emociones según la psicología, entre los diferentes enfoques el más utilizado y aceptado por los psicólogos es el modelo propuesto por Ekman y Friesen, quienes consideran 6 emociones básicas: ira, asco, miedo, alegría, tristeza y sorpresa. Por otra parte, sugieren la relación existente entre las unidades de acción y las emociones, mostrado en la tabla 2.1. Esta clasificación ha sido retomado por investigadores del área de la computación, quienes han tomado como base las investigaciones psicológicas para crear modelos inteligentes capaces de determinar las emociones en los individuos, estos trabajos presentados en el capítulo 2 han empleado diversos modelos de *machine learnig* o de *deep learnig* para lograr sus objetivos. Por ello, en este trabajo se plantea utilizar la clasificación desarrollada por Ekman para determinar las emociones, o bien, determinar las unidades de acción para posteriormente determinar la emoción. Para rezar el aprendizaje del modelo en la apariencia de una persona, las imágenes usadas se limitan a imágenes de tipo *selfie*, en donde solo aparece el rostro de una sola persona, intentando que el modelo aprenda a determinar las características necesarias para detectar la emoción.

Para determinar la emoción presente en una persona se trabajó con la detección de 14 unidades de acción, dado que algunas emociones comparten ciertas unidades de acción. Posteriormente, se identificó la emoción a partir de la relación existente entre unidades de acción y emoción, mostrada previamente en la tabla 2.1. Además, se determinó la emoción directamente a partir del rostro, sin la necesidad de pasar por las unidades de acción. De este modo, el proceso que se llevó a cabo para determinar la emoción en de este trabajo consta de cuatro pasos, el cual se muestra en la figura 4.1. Como primer paso se tiene la recolección de datos, en donde se trabajó con tres conjuntos de datos con la finalidad de observar el comportamiento de los modelos propuestos. El segundo paso es el preprocesamiento de los conjuntos de datos. Como tercer paso tenemos la implementación de las arquitecturas propuestas, y finalmente, la selección de la mejor arquitectura. Cada uno de los pasos se detalla a continuación.

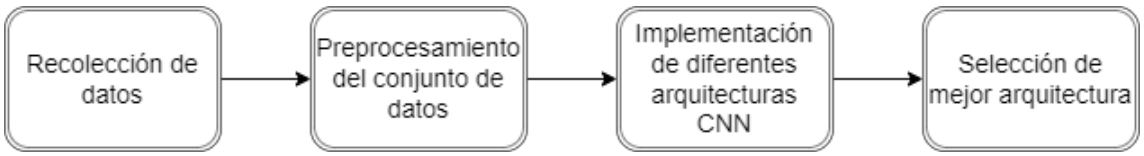

Figura 4.1: Diagrama de flujo para la detección de emociones.

4.1 Detección de emociones

4.1.1 Selección del corpus

Existen diversos corpus para la identificación de emociones, pero existen pocos para la identificación de unidades de acción. En el presente trabajo se utilizaron tres corpus que se detallan a continuación.

Cohn-Kanade ampliado (CK+)

El conjunto de datos Cohn-Kanade ampliado (CK+) (Lucey et al., 2010) contiene 593 secuencias de vídeo de un total de 123 sujetos diferentes, con edades comprendidas entre los 18 y los 50 años. Cada vídeo muestra un cambio facial de la expresión neutra a una expresión máxima, la cual es grabada a 30 fotogramas por segundo (FPS) con una resolución de 640x490 o 640x480 píxeles. De estos vídeos, 327 están etiquetados con una de las siete clases de emociones: ira, desprecio, asco, miedo, felicidad, tristeza y sorpresa, que corresponden al último fotograma (fotograma pico). Además, para cada secuencia existe un solo archivo FACS, donde se encuentra etiqueta una específica unidad de acción con su intensidad. La base de datos CK+ es considerada como la base de datos de clasificación de expresiones faciales más utilizada, la cual fue controlada en laboratorio, y se emplea en la mayoría de los métodos y de clasificación de expresiones faciales. Las figuras 4.2 y 4.3 muestran la distribución de etiquetas para este conjunto de datos.

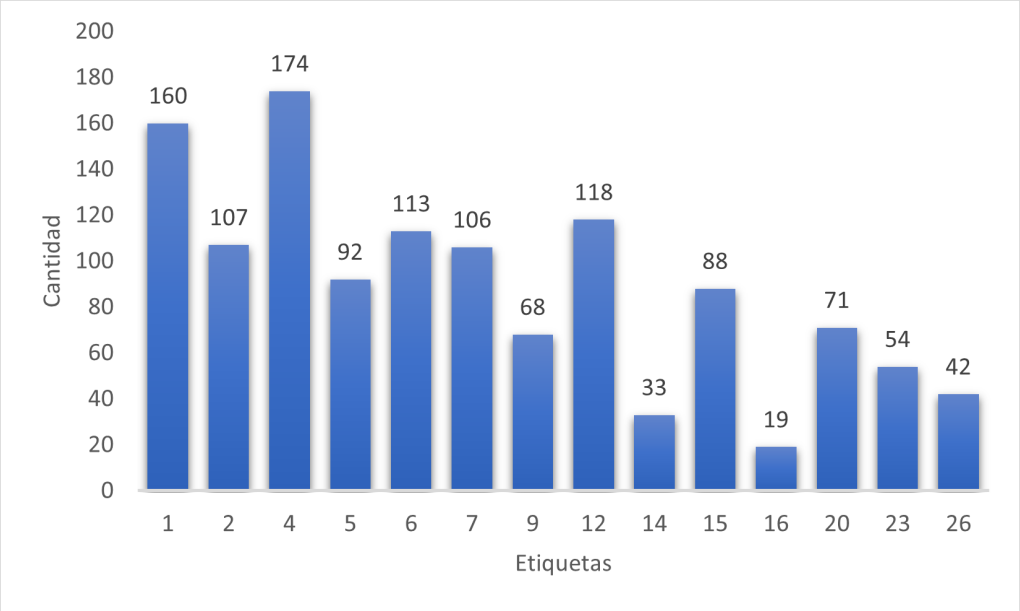

Figura 4.2: Distribución de etiquetas de unidades de acción para CK+.

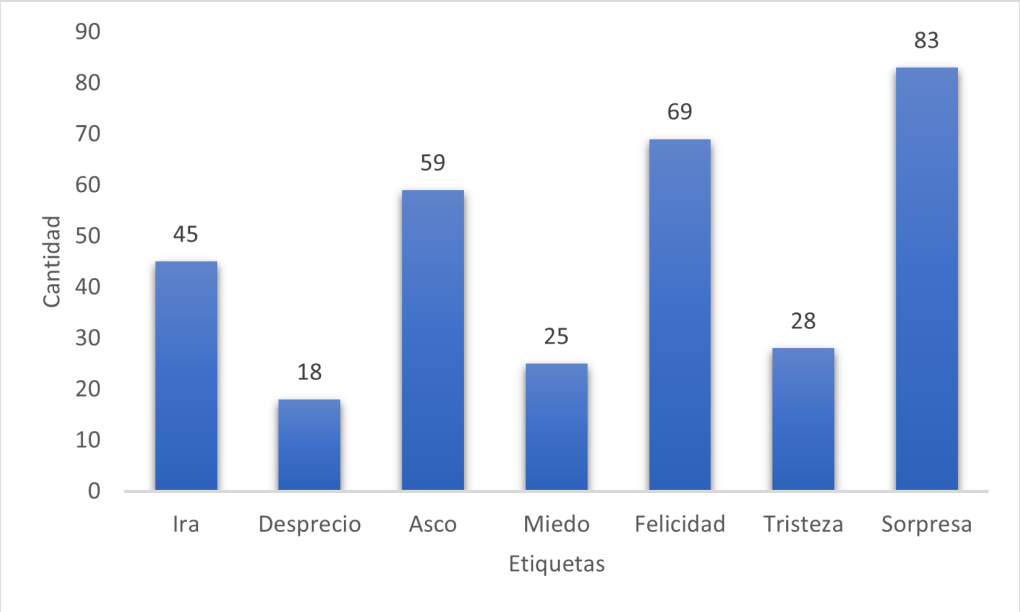

Figura 4.3: Distribución de etiquetas de emociones para CK+.

DISFA ampliado (DISFA+)

La base de datos ampliada de intensidad de las acciones faciales espontáneas de Denver (DISFA+ por sus siglas en inglés) (Mavadati et al., 2016), contiene una gran conjunto de datos de expresiones faciales posadas y espontáneas para nueve individuos. Además, proporciona las anotaciones basadas en $32,875$ fotogramas etiquetados manualmente con cinco niveles de intensidad de doce acciones faciales (FACS), con una resolución de 1280×720 píxeles en 20 FSP. La figura 4.4 muestra la distribución de etiquetas para este conjunto de datos.

Facial Expression Recognition 2013 (FER13)

La base de datos FER13 (Courville et al., 2013) consiste en imágenes de rostros a escala a grises de $48x48$. Los rostros se han registrado automáticamente para que la cara esté más o menos centrada y ocupe aproximadamente el mismo espacio en cada imagen. Contiene $23,744$ imágenes de rostros mostrando la expresión facial de una de las siete categorías: ira, miedo, felicidad, tristeza, sorpresa. La figura 4.5 muestra la distribución de etiquetas para este conjunto de datos.

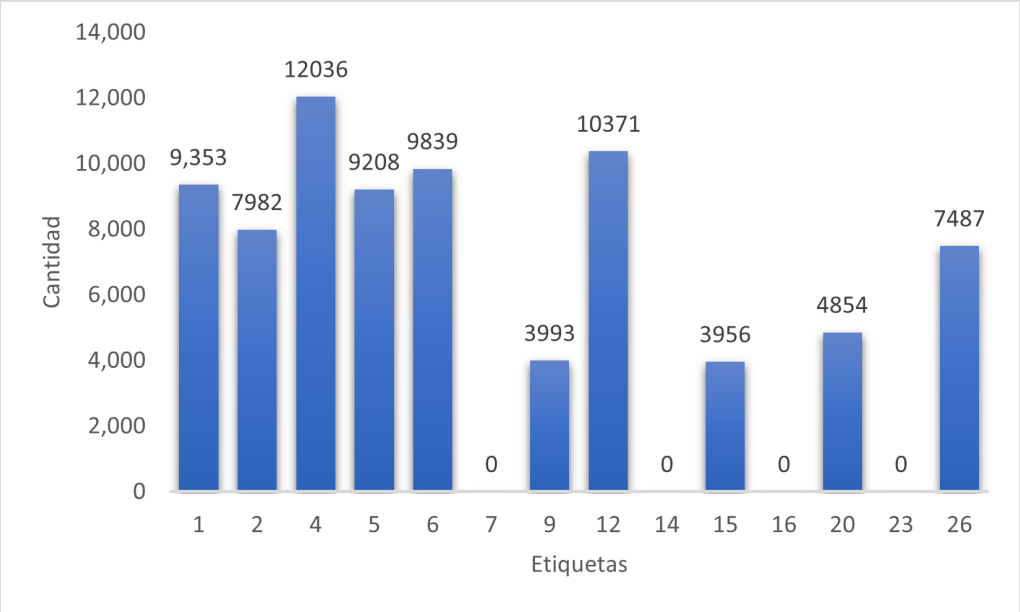

Figura 4.4: Distribución de etiquetas de unidades de acción para DISFA+.

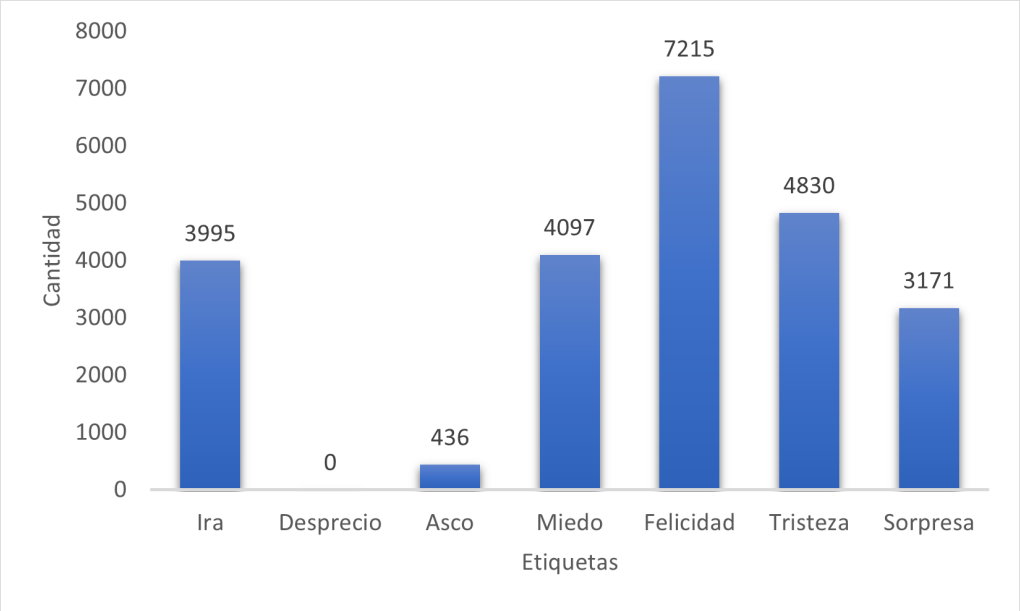

Figura 4.5: Distribución de etiquetas de emociones para FER13.

Las figuras 4.6, 4.7 y 4.8 muestran algunos ejemplos de las imágenes para las emociones en el conjunto de datos CK+, las emociones en FER13 y las unidades de acción en DISFA+ respectivamente.

Figura 4.6: Ejemplos de emociones presentes en CK+.

Figura 4.7: Ejemplos de emociones presentes en FER13.

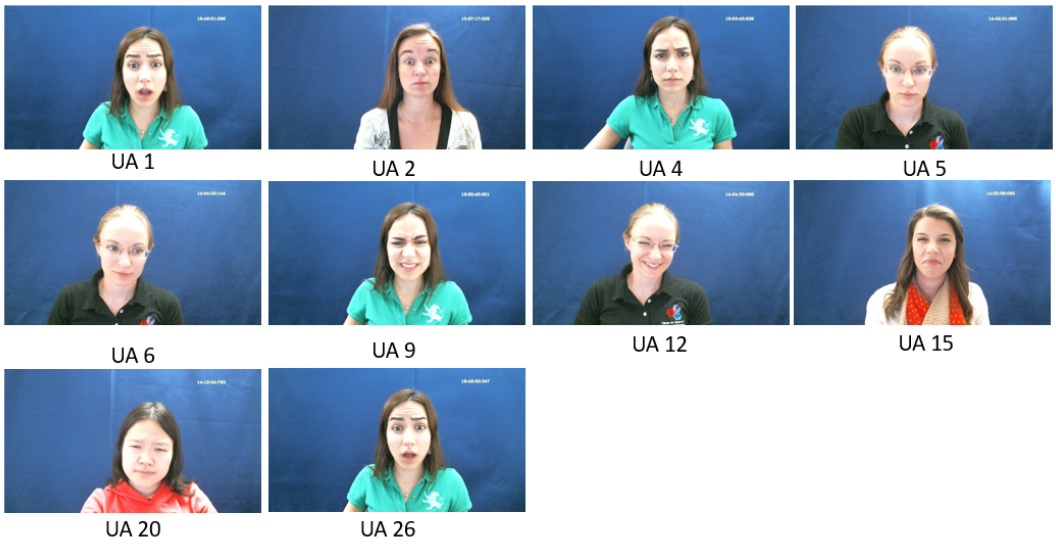

Figura 4.8: Ejemplos de unidades de acción presentes en DISFA+.

4.1.2 Preprocesamiento

Una vez seleccionados los corpus, las imágenes se manipularon para que el formato final fuera útil para el entrenamiento de los modelos planteados. Para el caso de los conjuntos de datos CK+ y DISFA+ el preprocesamiento consistió en:

1. Conversión de las imágenes a escala a grises: Se ocupó el método *cvtColor()* [1] de *OpenCV* basado en la siguiente fórmula:

$$RGB[A]toGray : Y \leftarrow 0{,}299 \cdot R + 0{,}587 \cdot G + 0{,}114 \cdot B \tag{4.1.1}$$

2. Detección del rostro: La detección de objetos mediante clasificadores en cascada basados en características *Haar* es un método eficaz propuesto por Viola y Jones (2001). Se trata de un enfoque basado en el aprendizaje automático en el que se entrena una función en cascada a partir de un lote de imágenes positivas y negativas. Se utiliza para detectar objetos en otras imágenes, en este caso rostros. Aquí se ocupó la detección de *OpenCV CascadeClassifier* puesto que contiene muchos clasificadores pre-entrenados para la detección del rostro. Para más información de como opera, visitar [2].

3. Redimensión de las imágenes: Una vez detectado el rostro, las imágenes fueron redimensionadas con *resize()* [3] a $224x224x3$ o $224x224x1$, de igual forma con ayuda de la librería de *OpenCV*, con el método de *INTER_CUBIC*. Este utiliza la interpolación bicúbica para redimensionar la imagen. Al redimensionar e interpolar los nuevos píxeles, este método actúa sobre los 4×4 píxeles vecinos de la imagen. A continuación, toma la media de los pesos de los 16 píxeles para crear el nuevo píxel interpolado.

4. Eliminación de ruido: La eliminación de ruido se realizó a partir del rostro detectado y a escala en grises. Se hizo la eliminación del ruido mediante la

[1] https://docs.opencv.org/3.4/de/d25/imgproc_color_conversions.html#color_convert_rgb_gray
[2] https://docs.opencv.org/3.4/d2/d99/tutorial_js_face_detection.html
[3] https://docs.opencv.org/3.4/da/d54/group__imgproc__transform.html#ga47a974309e9102f5f08231edc7e7529d

implementación basada en el rango del filtrado de la mediana [4].

5. Ecualización adaptativa del histograma (AHE): Para la ecualización, primero se detectó el rostro y posteriormente las imágenes fueron escaladas a grises. Luego se empleó *createCLAHE()* [5] de *OpenCV*. La ecualización adaptativa de histograma permite mejorar el contraste de las imágenes, su diferencia con la ecualización de histograma ordinaria es que en el método adaptativo se calculan varios histogramas, cada uno corresponde a una sección distinta de la imagen, y son usados para redistribuir los valores de luminosidad de la imagen. La AHE es adecuada para mejorar el contraste local y mejorar las definiciones de los bordes en cada región de la imagen.

6. Detección de bordes (Canny): Se utilizó el algoritmo *Canny* [6] que se encuentra en la librería *OpenCV*. Este algoritmo de visión por computadora para detectar los bordes, fue desarrollado por Canny (1986). Una vez detectado el rostro se procedió a la detección de bordes.

Para el caso del conjunto de datos FER13, el preprocesamiento consistió solo de los últimos tres puntos previamente mencionados. La figura 4.9 muestra ejemplos de las imágenes obtenidas posteriores al preprocesamiento sugerido.

[4] https://scikit-image.org/docs/stable/api/skimage.filters.rank.html#skimage.filters.rank.median

[5] https://docs.opencv.org/3.4/d6/db6/classcv_1_1CLAHE.html

[6] https://docs.opencv.org/3.4/dd/d1a/group__imgproc__feature.html#ga04723e007ed888ddf11d9ba04e2232de

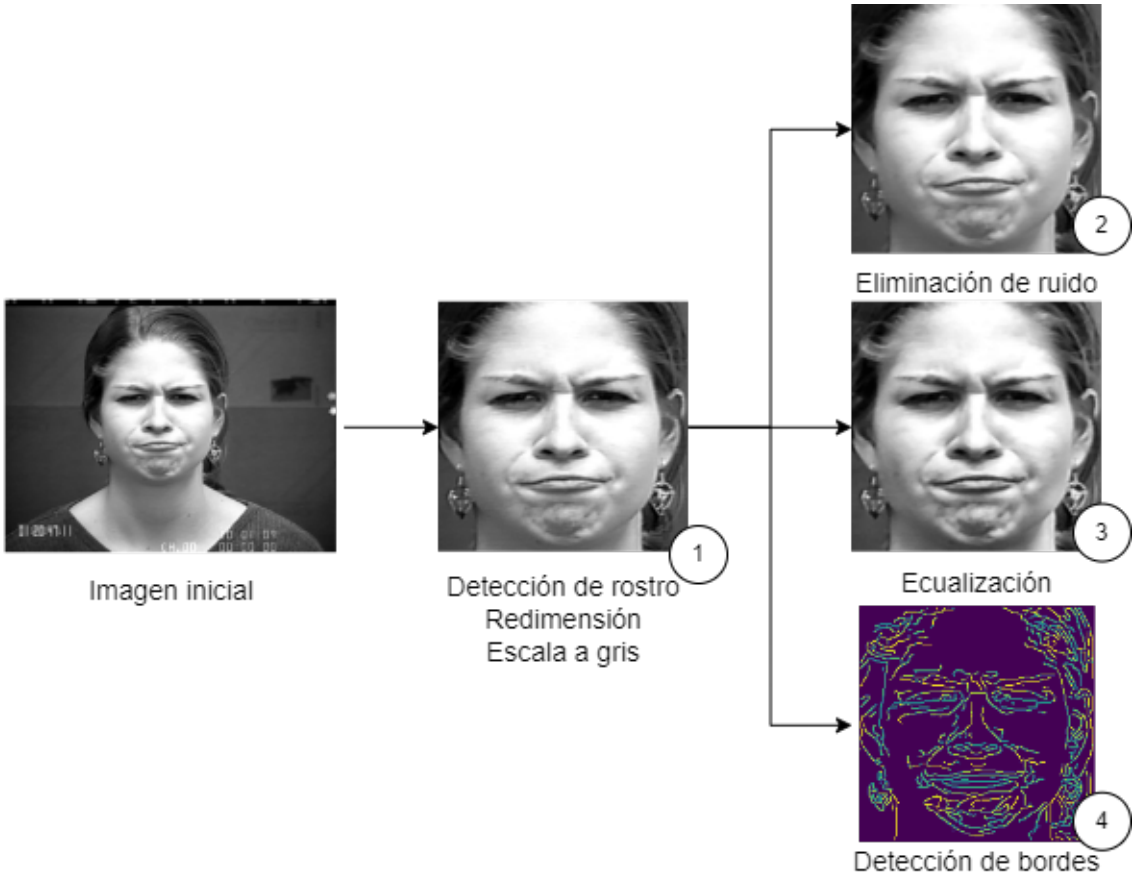

Figura 4.9: Preprocesamiento del conjunto de datos.

4.1.3 Detección de emociones de forma indirecta

Se probaron diversos modelos de redes neuronales convolucionales (CNN) para la detección de UA's, o bien, para la detección de emociones. Para el caso de la detección de emociones de forma indirecta, primero se realizó el entrenamiento de una CNN por UA (14 unidades de acción) para su detección. El entrenamiento de la CCN por UA del modelo propuesto se realizó con el conjunto de datos DISFA+, y se evaluó con el conjunto CK+ (UA1, UA2, UA4, UA5, UA6, UA9, UA12, UA15, UA20, UA26), excepto de cuatro unidades de acción que no se encontraban en DISFA+ (UA7, UA14, UA16,UA23). Para estos casos se utilizó el 90 % de datos disponibles de CK+ para el entrenamiento y 10 % para la evaluación. La arquitectura de la CNN que se entrenó por UA se muestra en la figura 4.10. Ocupando los siguientes parámetros:

- Tamaño de Batch de 64.
- Error cuadrático medio (MSE) como función de pérdida.
- Adam como optimizador con *learning rate* de 0.0001.
- Entrenamiento durante 40 épocas

Una vez entrenadas las CNN por UA, se procedió al ensamble de las CNN's. Como salida del ensamble se obtiene un vector de longitud 14 representando si está o no la UA (1 si esta presente, 0 en caso contrario). El vector de salida obtenido en el ensamble pasará a un perceptrón multicapa previamente entrenado para la clasificación de la emoción. Este proceso se muestra en la figura 4.11.

4.1.4 Detección de emociones de forma directa

Para el caso de detección de emoción de forma directa, se probaron 2 arquitecturas, en ambos casos el entrenamiento se realizó con el conjunto de datos FER13 y se evaluó con el conjunto CK+. La figura 4.12 y 4.13 muestran las arquitecturas propuesta para la clasificación de emociones. Ocupando los siguientes parámetros:

- Tamaño de Batch de 64.
- *categorical_crossentropy* como función de pérdida.
- Adam como optimizador con *learning rate* de 0.0001.
- Entrenamiento durante 50 épocas

4.1.5 Selección del mejor modelo

Las arquitecturas previamente mencionadas fueron entrenadas con las imágenes sin procesar, las imágenes a escala de grises con eliminación de ruido, ecualizadas o con detección de bordes. Posteriormente, se calculó la matriz de confusión y las métricas F1 Score, *presicion*, *recall* y *accuracy* para poder escoger el mejor modelo.

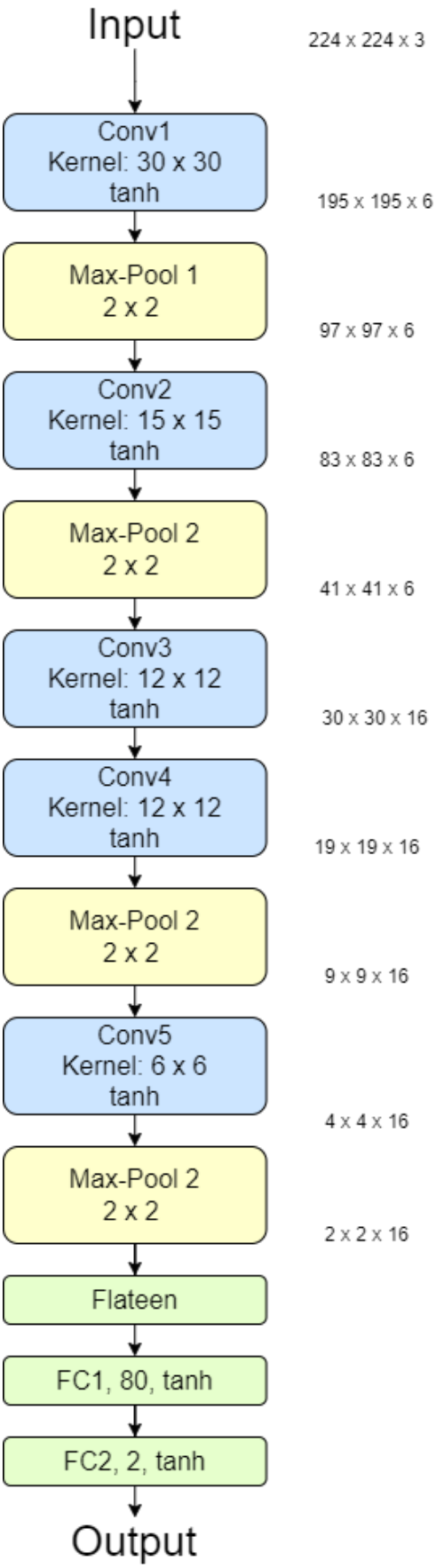

Figura 4.10: Arquitectura propuesta para la detección por UA.

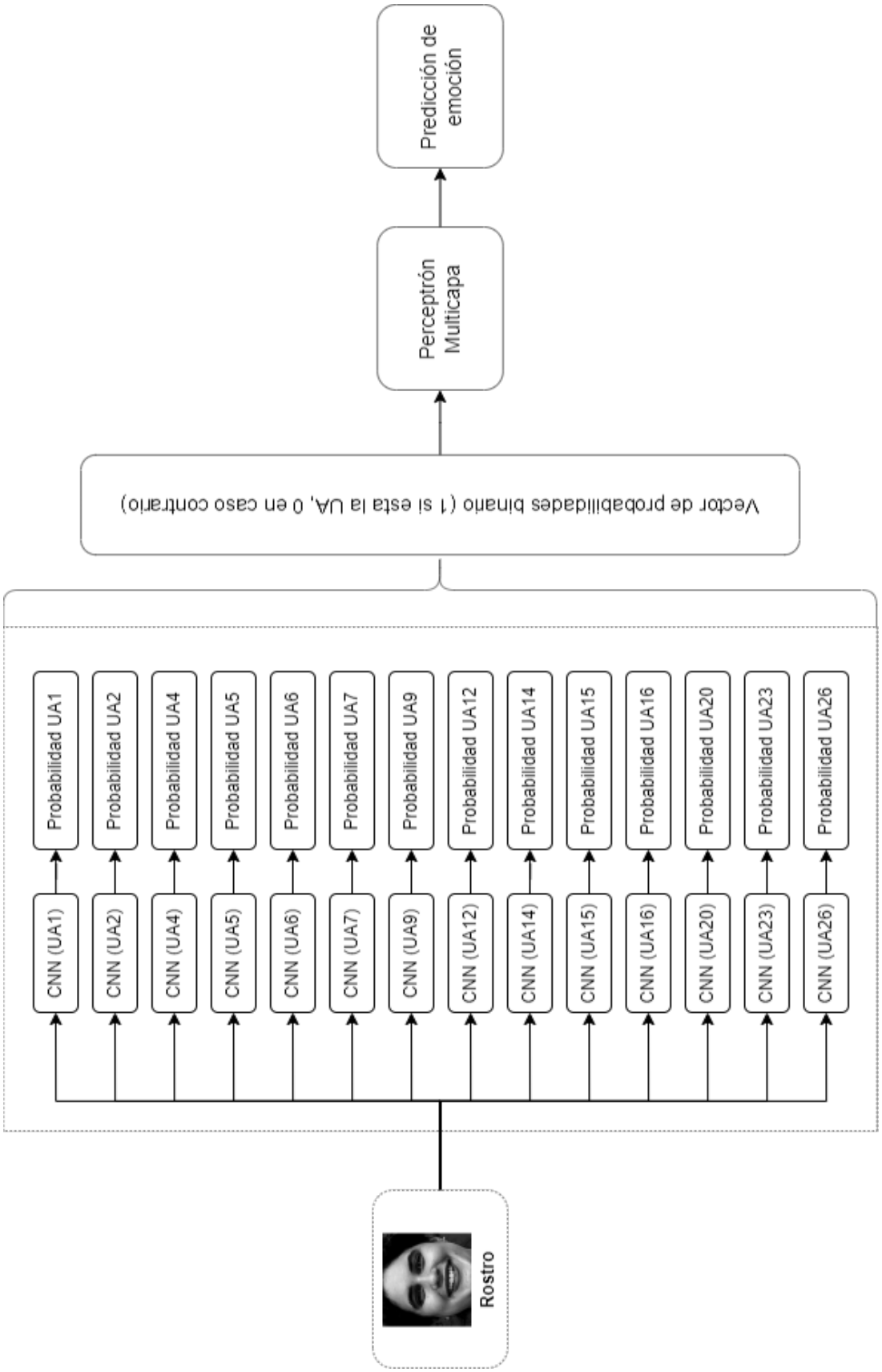

Figura 4.11: Ensamble para la detección de emociones.

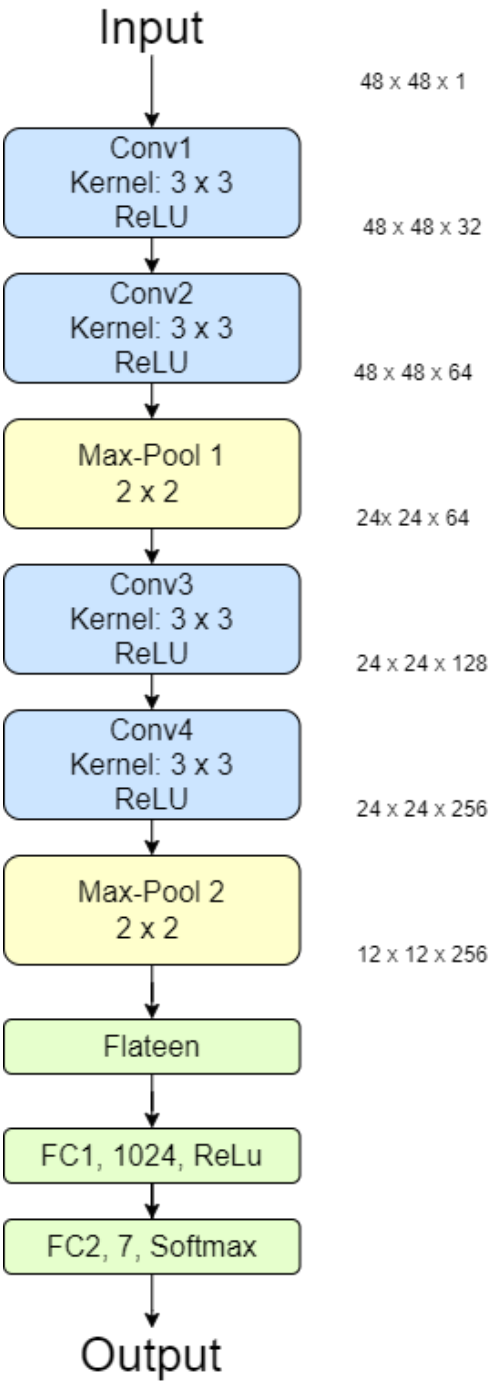

Figura 4.12: CNN1: Arquitectura utilizada para la detección de emociones.

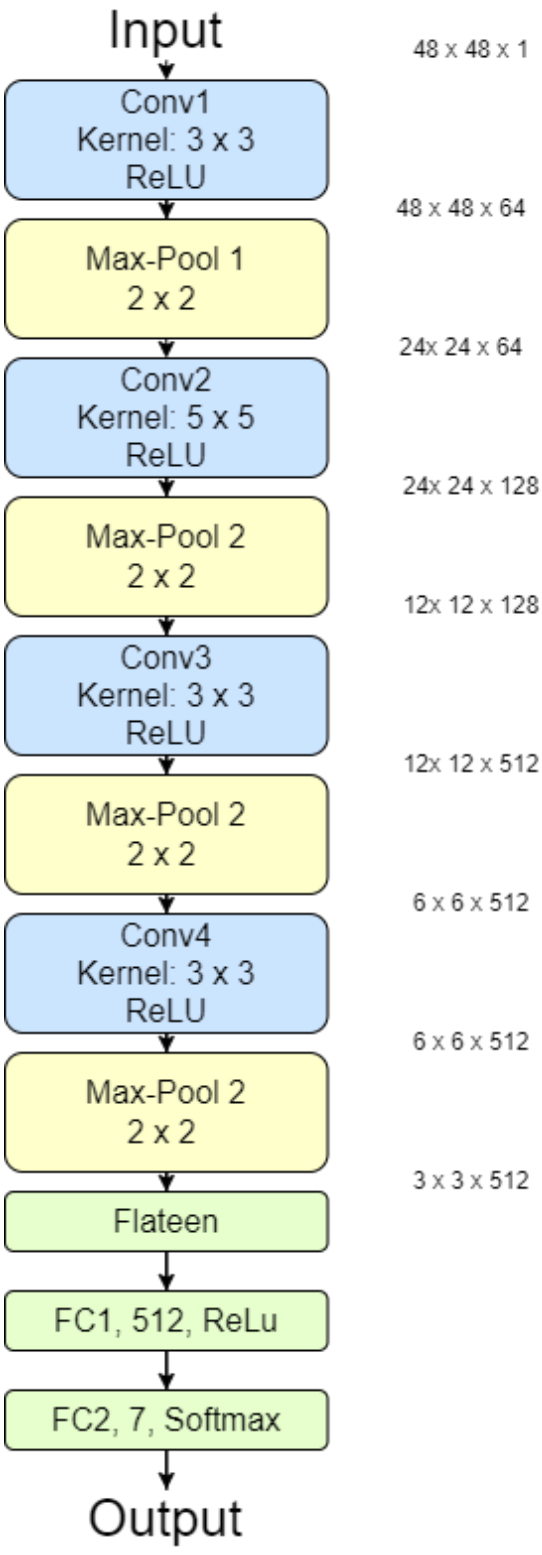

Figura 4.13: CNN2: Arquitectura utilizada para la detección de emociones.

4.2 Recomendador

Una vez elegido el modelo para determinar la emoción en una persona, se procede a realizar una recomendación de alguna bebida de *buuble tea*, previamente clasificada en alguno de los cinco sabores básicos, basándonos en 2.2. Cada sabor fue clasificado en uno de los cinco sabores básicos. Por lo que al hacer una recomendación se hará eligiendo alguna bebida que se encuentre en dicho grupo. Finalmente, la figura 4.14 muestra el diagrama de flujo final de la solución propuesta.

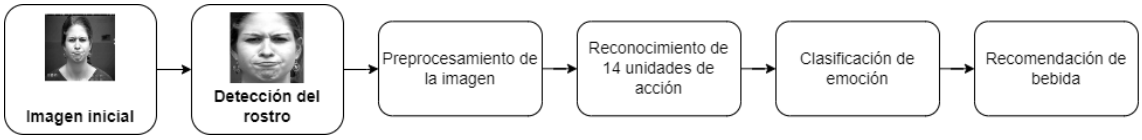

Figura 4.14: CNN2: Arquitectura utilizada para la detección de emociones.

Capítulo 5

Experimentos y resultados

En este capítulo se explican los experimentos con los modelos propuestos en el 4, con la finalidad de comparar las soluciones propuestas entre sí, y contra el estado del arte. Para así poder elegir un modelo capaz de detectar la emoción en una persona de la mejor manera, y poder recomendar alguna bebida de *buble tea*.

5.1 Experimento 1: Detección de emociones de forma indirecta

Como se había explicado previamente, en la clasificación de emociones de forma indirecta primero se entrenaron 14 modelos de CNN para detectar si se encontraba, o no, la unidad de acción. Cabe recalcar que para el entrenamiento de los modelos, 10 CNN fueron entrenadas con subconjuntos de datos de DISFA+. Estos subconjuntos solo poseían imágenes etiquetadas con la UA a detectar, más la misma cantidad de imágenes neutras, es decir, que no estaban etiquetas con ninguna UA. Para la evaluación del entrenamiento de la CNN se utilizó un subconjunto de CK+, el cual, de igual manera que los subconjuntos de DISFA+, solo tenía imágenes con la UA a detectar más imágenes neutras. Para el caso de las otras 4 UAs, las CNN se

entrenaron y evaluaron con subconjuntos de CK+. Dicho entrenamiento se realizó con los preprocesamientos mencionados en el 4: imágenes sin procesar, las imágenes a escala de grises con eliminación de ruido, ecualizadas o con detección de bordes. La tabla 5.1 muestra el promedio de los resultados obtenidos en las 14 CNN por UA con los diferentes tipos de preprocesamiento.

Tabla 5.1: Promedio de los resultados obtenidos de las 14 unidades de acción.

Preprocesamiento	*Accuracy*	*Precision*	*Recall*	*F1 Score*
Detección de rostro	0,80	0,81	0,75	0,75
Eliminación de ruido	0,79	0,85	0,70	0,74
Ecualización adaptativa	0,78	0,88	0,63	0,68
Detección de bordes (Canny)	0,70	0,69	0,60	0,64

Después de que fueran entrenadas las CNN se procedió a realizar el ensamble, el cual funciona en forma de cascada, como se explicó en la sección 4.1.3. El ensamble se evaluó con el conjunto CK+, únicamente con las imágenes que se encontraban etiquetadas con alguna emoción. La figura 5.1 muestra las matrices de confusión de estos resultados.

Por otra parte, se hizo la clasificación solo de 4 emociones: ira, felicidad, tristeza y sorpresa, ya que estas emociones son las que tienen las unidades de acción más marcadas, por lo que la figura 5.2 muestra las matrices de confusión de estos resultados.

5.2 Experimento 2: Detección de emociones de forma directa

La detección de emociones de forma directa, se hizo mediante el entrenamiento de las redes propuestas en el 4. La tabla 5.2 muestra el promedio de los resultados obtenidos para la clasificación de la primera CNN1 propuesta. Además, la 5.3 muestra las matrices de confusión obtenidas con los diferentes entrenamientos.

De igual forma que en la detección indirecta de emociones, se hizo la cla-

sificación solo de 4 emociones: ira, felicidad, tristeza y sorpresa, por lo que la figura 5.4 muestra las matrices de confusión de estos resultados y la tabla 5.3 el promedio de los resultados.

Tabla 5.2: Resultados obtenidos al clasificar siete emociones con CNN1.

Preprocesamiento	Accuracy	Precision	Recall	F1 Score
Detección de rostro	0,64	0,51	0,51	0,42
Eliminación de ruido	0,58	0,41	0,47	0,40
Ecualización adaptativa	0,54	0,43	0,41	0,33
Detección de bordes (Canny)	0,14	0,02	0,14	0,03

Tabla 5.3: Resultados obtenidos al clasificar cuatro emociones con una CNN1.

Preprocesamiento	Accuracy	Precision	Recall	F1 Score
Detección de rostro	0,82	0,77	0,80	0,77
Eliminación de ruido	0,70	0,77	0,75	0,68
Ecualización adaptativa	0,60	0,73	0,66	0,61
Detección de bordes (Canny)	0,32	0,08	0,25	0,12

La tabla 5.4 muestra el promedio de los resultados obtenidos para la clasificación de la primera CNN2 propuesta. La tabla 5.5 el promedio de los resultados obtenidos con la CNN2 pero para cuatro emociones. Las matrices de confusión se muestran en las figuras 5.5 para la clasificación de siete emociones, y las figuras 5.6 para la clasificación de 4 emociones.

Tabla 5.4: Resultados obtenidos al clasificar siete emociones con CNN2.

Preprocesamiento	Accuracy	Precision	Recall	F1 Score
Detección de rostro	0,63	0,50	0,50	0,45
Eliminación de ruido	0,57	0,45	0,44	0,41
Ecualización adaptativa	0,09	0,22	0,14	0,16
Detección de bordes (Canny)	0,14	0,02	0,14	0,03

Tabla 5.5: Resultados obtenidos al clasificar cuatro emociones con una CNN2.

Preprocesamiento	Accuracy	Precision	Recall	F1 Score
Detección de rostro	0,77	0,77	0,79	0,74
Eliminación de ruido	0,71	0,74	0,75	0,68
Ecualización adaptativa	0,63	0,68	0,68	0,59
Detección de bordes (Canny)	0,32	0,08	0,25	0,12

5.3 Discusión

Se puede notar que para el ensamble, a pesar de que en la detección de UA le fue mejor a las CNN entrenadas con imágenes sin preprocesamiento, para la detección de emociones, la mejor clasificación se logró con detección de bordes, logrando detectar la mayoría de las emociones a excepción de desprecio. En la detección de solo 4 emociones se puede notar que, de igual forma, la detección de bordes parece ser la mejor opción. Con la arquitectura CNN1 la mejor clasificación se logró con las imágenes sin procesar al clasificar 7 emociones, pero no logro mejores resultados que el ensamble. Para la clasificación de 4 emociones con la CNN1, las imágenes sin procesar lograron los mejores resultados, y logró superar al ensamble. Finalmente, para la clasificación de 7 emociones con la CNN2, los mejores resultados se obtuvieron entrenando con las imágenes sin procesar, pero solo se logró clasificar 4 emociones bien. Al clasificar 4 emociones con la CNN2 se puede notar que al entrenar con las imágenes sin procesar se lograron mejores resultados, además se comportó de igual forma que la CNN1.

Como se puede observar con las matrices de confusión obtenidas y dado que queremos trabajar con siete emociones básicas. Se eligió el ensamble que clasifica siete emociones para la solución propuesta. La tabla 5.6 muestra la comparación contra algunas investigaciones. Se debe considerar que los artículos mencionados en la tabla trabajan con otros conjuntos de datos, además que el entrenamiento y evaluación la realizaron con el mismo conjunto, a diferencia de nuestra propuesta, ya que se entrenó con un conjunto, y se evaluó con otro. Esto proporciona variabilidad al momento de entrenar y evaluar.

Tabla 5.6: Comparación con el estado del arte

Artículo	Clases	Corpus	Accuracy
(Liu et al., 2016)	6 clases	FER13	65,03 %
(Mavani et al., 2017)	6 clases	CFEE	74,79 %
(Shan et al., 2017)	6 clases	JAFFE	74,74 %
(Sajjanhar et al., 2018)	6 clases	JAFFE	73,53 %
(Ozdemir et al., 2019)	6 clases	JAFFE	63,63 %
Propuesta Ensamble	7 clases	CK+	60,0 %
Propuesta CNN1	7 clases	CK+	64,0 %
Propuesta CNN2	7 clases	CK+	63,0 %

5.4 Recomendador

Para el funcionamiento del recomendador se categorizó cada una de las bebidas de *bubble tea*. Se tomaron los existentes en Moreno-Armendáriz et al. (2021) que ya estaban vectorizadas, y posteriormente se clasificaron en uno de los cinco sabores básicos: dulce, umami, amargo, ácido y picante. Una vez detectada la emoción, el recomendaror seleccionará alguna bebida del catálogo existente, basado en [1]. Es decir, en caso de detectar la emoción feliz, se recomendará una bebida dulce; si la emoción es triste, se recomendará una bebida picante o amarga; una bebida ácida en caso de una emoción de ira o desprecio; y para la sorpresa una bebida umami. La tabla 5.7 muestra algunos ejemplos de bebidas a recomendar por cada sabor.

[1] https://www.yoga.com.mx/los-sabores-y-las-emociones/

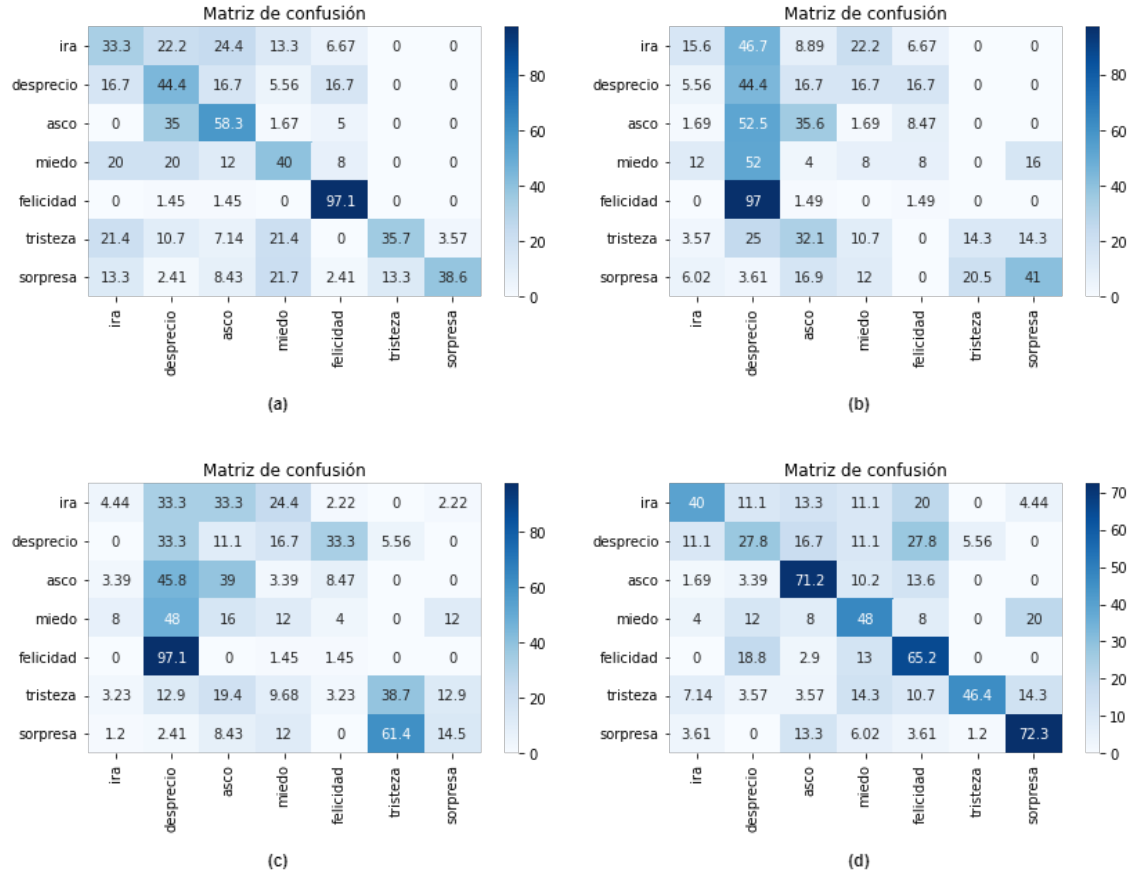

Figura 5.1: Matrices de confusión en la detección de 7 emociones básicas con el ensamble y los diferentes preprocesamientos. (a) Sin preprocesamiento, (b) Sin ruido, (c) Ecualizado, (d)Detección de borde

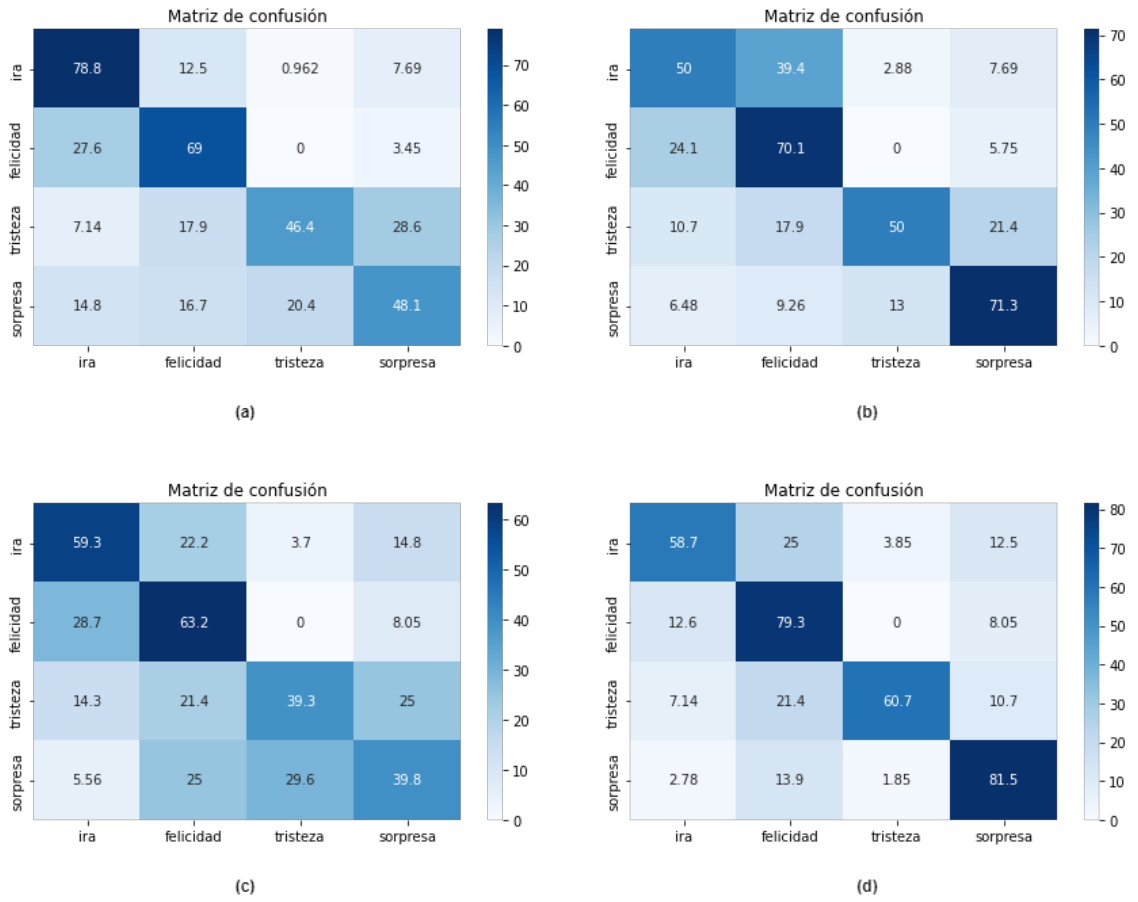

Figura 5.2: Matrices de confusión en la detección de 4 emociones básicas con el ensamble y los diferentes preprocesamientos.(a) Sin preprocesamiento, (b) Sin ruido, (c) Ecualizado, (d)Detección de borde

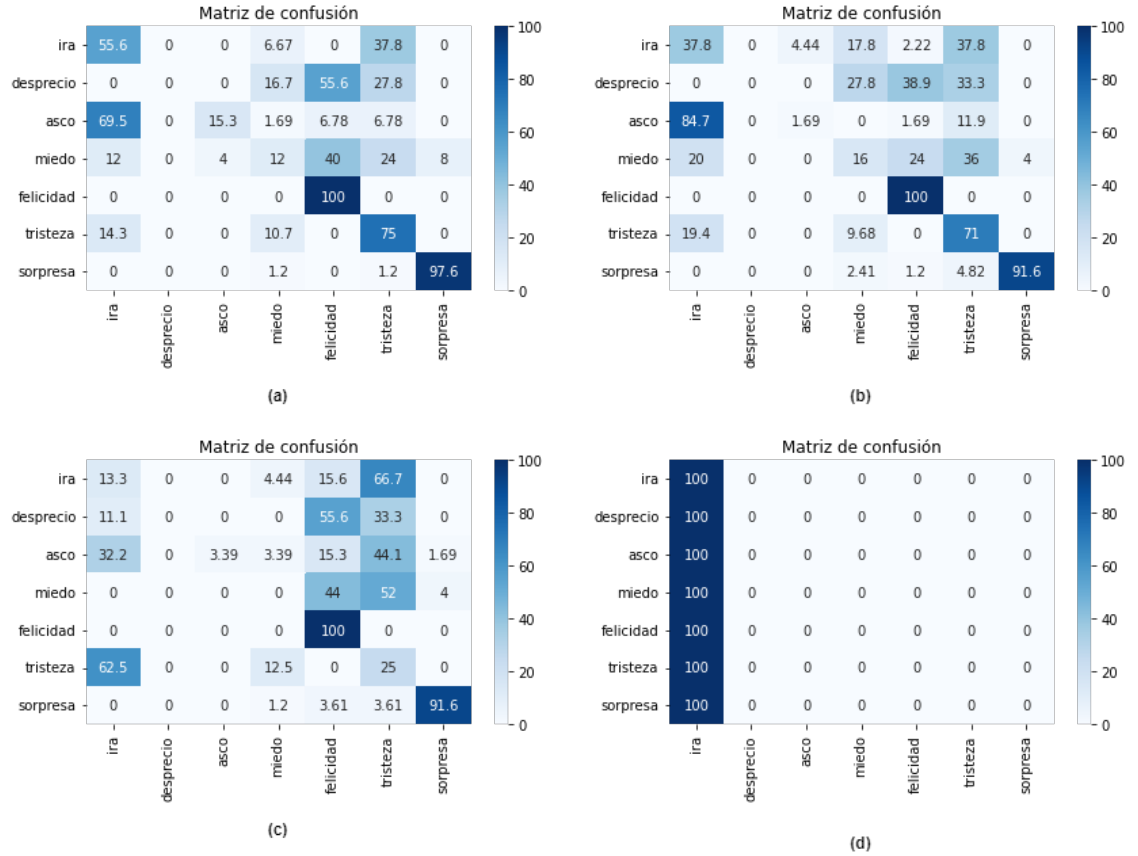

Figura 5.3: Matrices de confusión en la detección de 7 emociones básicas con CNN1 y los diferentes preprocesamientos. (a) Sin preprocesamiento, (b) Sin ruido, (c) Ecualizado, (d)Detección de borde

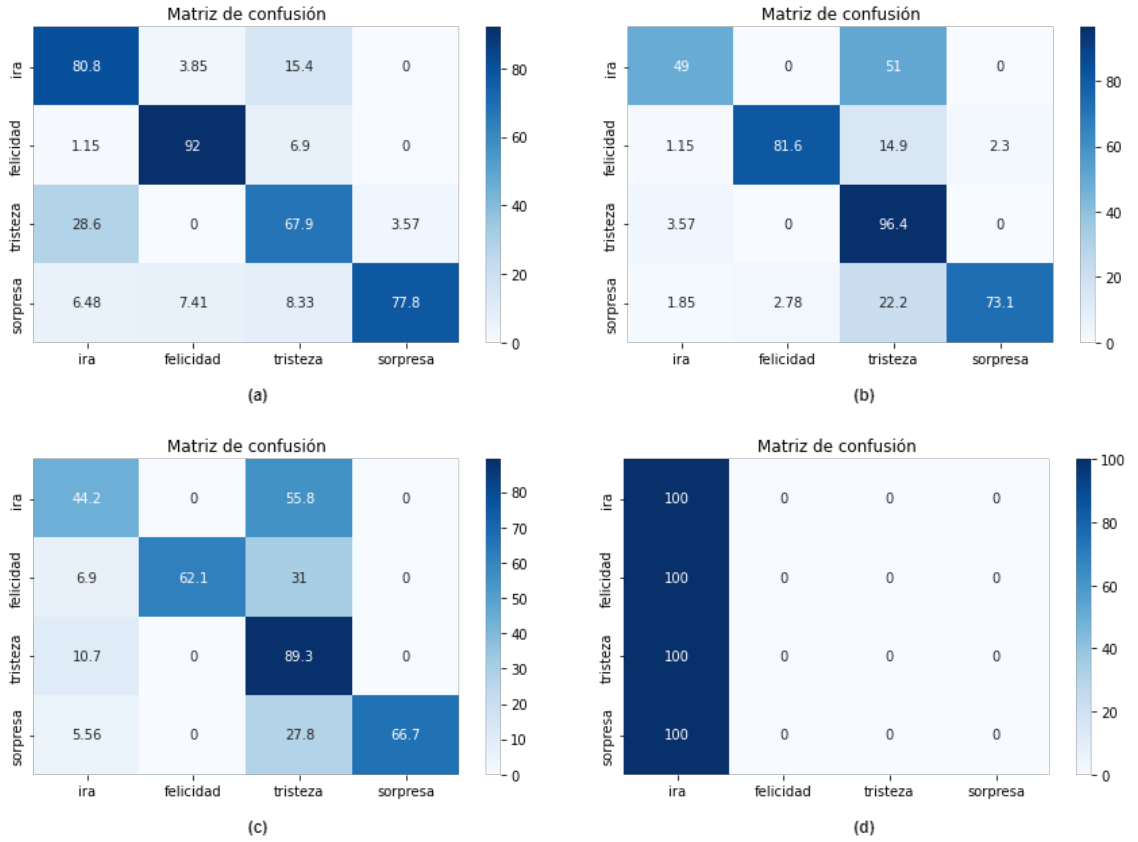

Figura 5.4: Matrices de confusión en la detección de 4 emociones básicas con CNN1 y los diferentes preprocesamientos. (a) Sin preprocesamiento, (b) Sin ruido, (c) Ecualizado, (d)Detección de borde

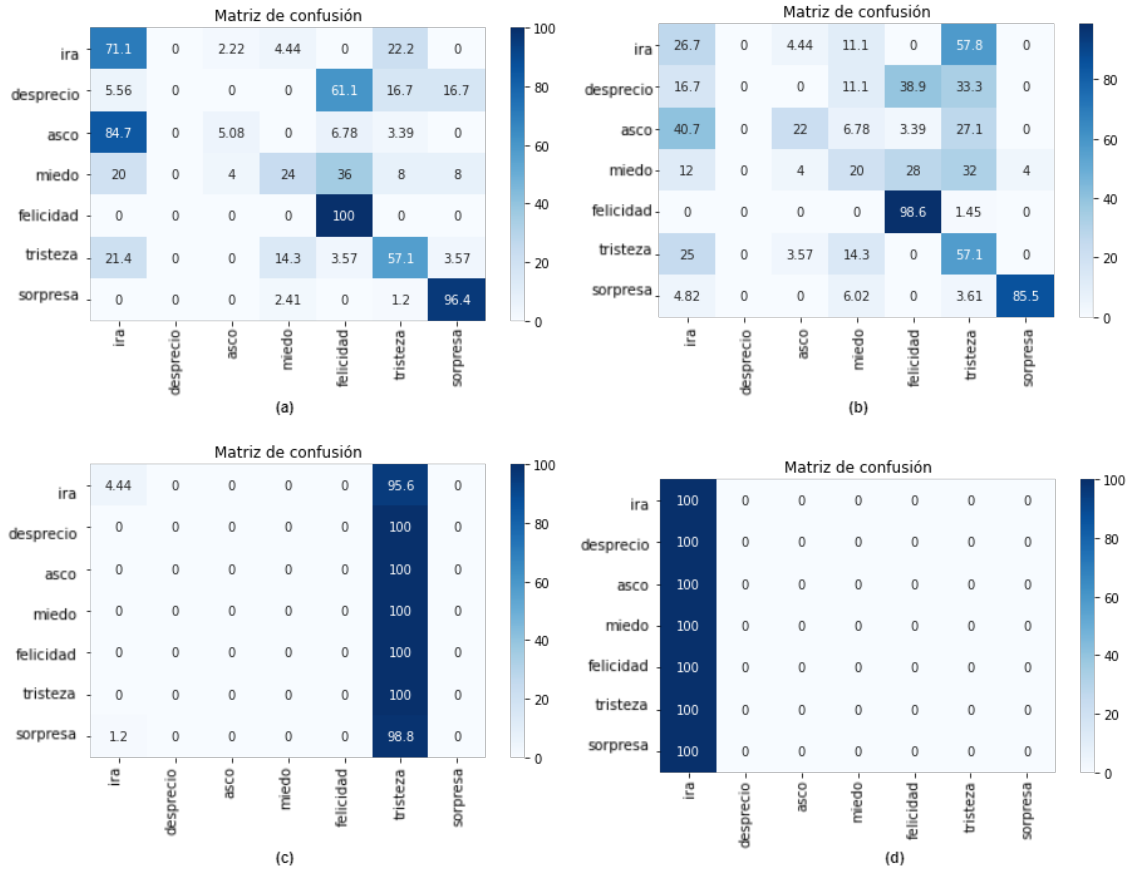

Figura 5.5: Matrices de confusión en la detección de 7 emociones básicas con CNN2 y los diferentes preprocesamientos.(a) Sin preprocesamiento, (b) Sin ruido, (c) Ecualizado, (d)Detección de borde

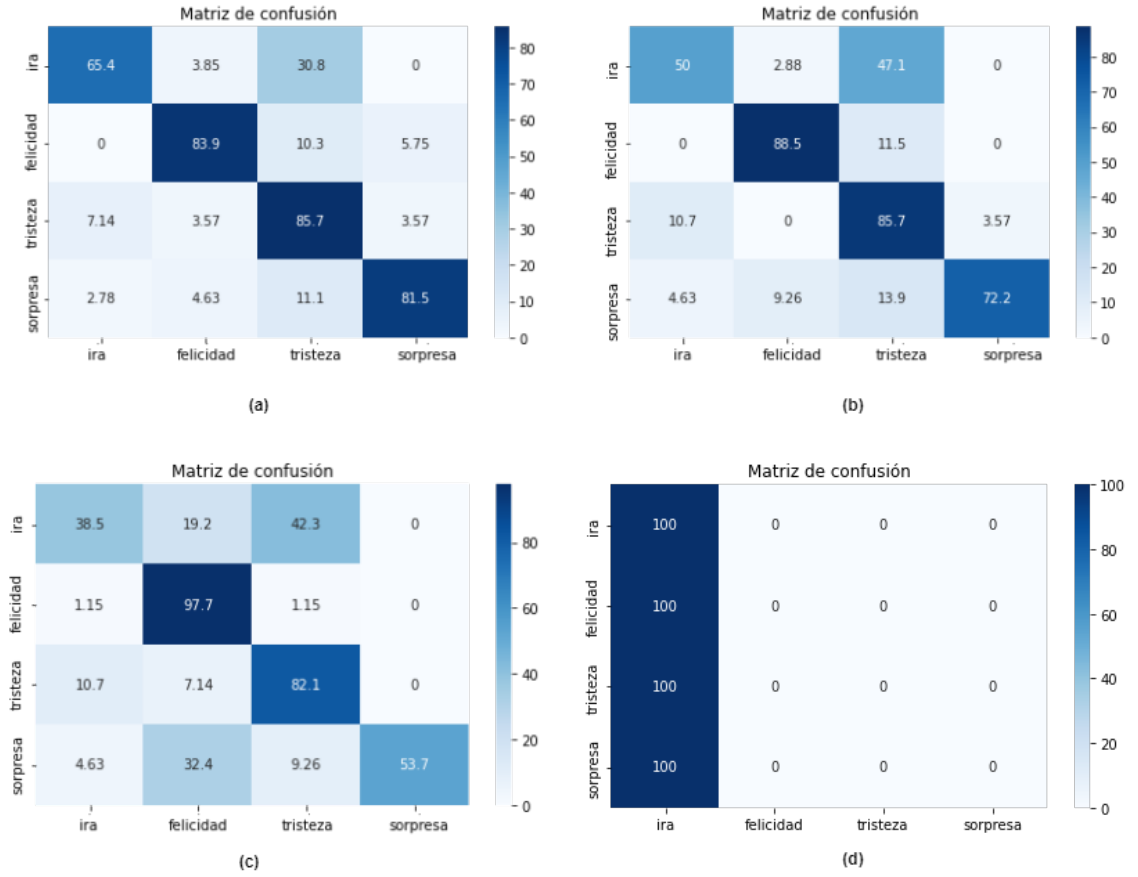

Figura 5.6: Matrices de confusión en la detección de 4 emociones básicas con CNN2 y los diferentes preprocesamientos.(a) Sin preprocesamiento, (b) Sin ruido, (c) Ecualizado, (d)Detección de borde

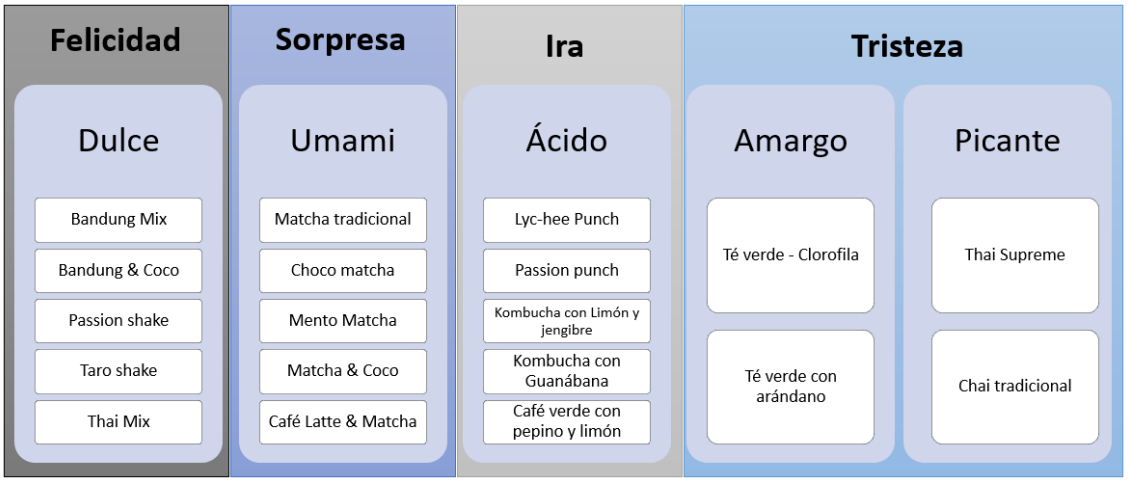

Figura 5.7: Bebidas de *bubble tea*

Capítulo 6

Conclusiones y trabajo futuro

6.1 Conclusiones

En este trabajo se propone una forma nueva forma para la detección de emociones, al detectar inicialmente las unidades de acción de las expresiones faciales, y posteriormente, clasificar la emoción. Generalmente las opciones propuestas para la detección de emociones, están enfocadas a determinar directamente la emoción, o bien, solo detectar las unidades de acción presentes en las expresiones faciales. Uno de los mayores retos en la detección de emociones es el corpus que se elija, ya que la calidad y las condiciones en que se recaba la información, influye demasiado con los resultados esperados. En nuestra propuesta el entrenamiento se realizó con diversos corpus, puesto que así se le agregaba variabilidad al entrenamiento y se intenta validar que las características que se están buscando, en este caso las unidades de acción, son correctas. El rendimiento obtenido para la detección de la mayoría de unidades de acción fue buena, oscilando entre el 70 % al 80 %, a excepción de algunas, de las cuales no se disponía de bastantes ejemplos para el entrenamiento de la CNN. Otra cosa que se pudo observar, fue que algunas imágenes que se tenían de DISFA+, a pesar de estar etiquetadas con alguna unidad de acción, esta no se presentaba de manera intensa, lo cual agregaba ruido al momento de entrenar. Por

otra parte, la clasificación de unidades de acción a emociones, no siempre existía la forma "ideal", es decir, que solo se activarán las unidades de acción que los expertos mencionan, más bien, siempre existía ruido. También en la mayoría de los casos los mejores resultados se obtuvieron sin la necesidad de pre procesar las imágenes, más que con la detección del rostro. Además, que el desempeño mejoró cuando solo se clasificaban cuatro emociones que se consideran como opuestas, en comparación con la clasificación de las siete emociones. Según los expertos, el desprecio suele confundirse bastante con el asco o la ira, este comportamiento se vio reflejado en las matrices de confusión obtenidas. Además, en la clasificación de emociones de forma directa los mejores resultados se obtuvieron al solo tomar en cuenta cuatro emociones. Uno de los mayores retos que presenta el conjunto FER13 es el tipo de imágenes con las que se trabaja, ya que no solo implican el rostro, si no, que suelen presentarse las manos en rostro.

Finalmente, para el caso del recomendador, la mayoría de investigaciones están enfocadas en que emociones son evocadas por ciertos alimentos, si bien, se comienzan hacer investigaciones para poder predecir que productos prefieren las personas a partir de la emoción que están sintiendo, no se encontró una predicción bien fundamentada para hacer esta relación. Es por ello que se optó por un modelo más sencillo para la recomendación de bebidas.

6.2 Trabajo futuro

Aún existe una gran oportunidad de mejora en la detección de emociones con los modelos propuestos. Esto se puede realizar mediante la división por regiones del rostro, ya sea en tres partes: boca, nariz y ojos, o dividiendo por regiones específicas donde se agrupen las unidades de acción. Otra opción podría ser mediante aprendizaje reforzado, el cual se implementaría para la selección de las imágenes más adecuadas para el entrenamiento de la red. También, se podría trabajar con más conjuntos de datos, para agregar mayor variedad y trabajar con un mayor número

de imágenes para el entrenamiento. Además, de buscar otros tipos de preprocesamiento que ayuden a focalizarse en las características que se estan buscando, o bien, un ensamble con diferentes preprocesamientos, adecuando el que mejor se comporte para cada emoción o unidad de acción.

Por otra parte, para el caso del recomendador, nos podemos enfocar a dar recomendaciones para cambiar el estado de ánimo o emoción de un individuo a partir de la emoción detectada. Existen estudios que mencionan que ciertos alimentos pueden llegar a mejorar el estado de ánimo de una persona. En este apartado, aún hay bastante información que revisar para focalizar de mejor manera la recomendación de productos a partir de las emociones.

Referencias

Arnold, M. B. (1960). Emotion and personality.

Borgalli, M. R. A., & Surve, S. (2022). Deep learning for facial emotion recognition using custom CNN architecture. *Journal of Physics: Conference Series*, *2236*(1), 012004.

Cairns, R., Elder, G., & Costello, E. (1996). Developmental Science. https://doi.org/10.1017/CBO9780511571114

Calhoun, C., & Solomon, R. (1989). ¿Qué es una emoción? Lecturas clásicas de psicología filosófica.

Canny, J. (1986). A Computational Approach to Edge Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-8*(6), 679-698. https://doi.org/10.1109/TPAMI.1986.4767851

Cardello, A. V., Meiselman, H. L., Schutz, H. G., Craig, C., Given, Z., Lesher, L. L., & Eicher, S. (2012). Measuring emotional responses to foods and food names using questionnaires. *Food Quality and Preference*, *24*(2), 243-250.

Cohn, J. F., & Schmidt, K. L. (2004). The timing of facial motion in posed and spontaneous smiles. *International Journal of Wavelets, Multiresolution and Information Processing*, *2*(02), 121-132.

Corneanu, C. A., Oliu, M., Cohn, J. F., & Escalera, S. (2016). Survey on RGB, 3D, Thermal, and Multimodal Approaches for Facial Expression Recognition: History, Trends, and Affect-related Applications. https://doi.org/https://doi.org/10.1109/tpami.2016.2515606

Courville, P., Goodfellow, A., Mirza, I., & Bengio, Y. (2013). FER-2013 face database. *Universit de Montreal: Montréal, QC, Canada*.

Darwin, C. (1872). *The Expression of the Emotions in Man and Animals* (1.ª ed.). John Murray.

DeVault, D., Artstein, R., Benn, G., Dey, T., Fast, E., Gainer, A., Georgila, K., Gratch, J., Hartholt, A., Lor-Lhommet, M., Lucas, G., Marsella, S., Morbini, F., Nazarian, A., Scherer, S., Stratou, G., Suri, A., Traum, D., Wood, R., & Morency, L.-P. (2014). SimSensei Kiosk: A Virtual Human Interviewer for Healthcare Decision Support. *13th International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2014*, *2*, 1061-1068.

Dhall, A., Ramana Murthy, O., Goecke, R., Joshi, J., & Gedeon, T. (2015). Video and image based emotion recognition challenges in the wild: Emotiw 2015. *Proceedings of the 2015 ACM on international conference on multimodal interaction*, 423-426.

Ekman, P. (1982). What emotion categories or dimensions can observers judge from facial behavior?, 39-55.

Ekman, P. (2003). *Emotions Revealed. Recognizing faces and feelings to improve communication and emotional life*. Henry Holt; Company.

Ekman, P., & Friesen, W. (2002). Facial action coding system: the manual ; on CD ROM.

Ekman, P., & Friesen, W. V. (1969). The Repertoire of Nonverbal Behavior: Categories, Origins, Usage, and Coding. *Semiotica*, *1*(1), 49-8. https://doi.org/doi:10.1515/semi.1969.1.1.49

Ekman, P., & O'Sullivan, M. (2006). From flawed self-assessment to blatant whoppers: The utility of voluntary and involuntary behavior in detecting deception. *Behavioral sciences the law*, *24*, 673-86. https://doi.org/10.1002/bsl.729

Ekman, P., & Rosenberg, E. L. (2005). What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS). https://doi.org/10.1093/acprof:oso/9780195179644.001.0001

Frijda, N. (1986). The emotions.

Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.-H., et al. (2013). Challenges in representation learning: A report on three machine learning contests. *International conference on neural information processing*, 117-124.

Gudi, A., Tasli, H. E., Den Uyl, T. M., & Maroulis, A. (2015). Deep learning based facs action unit occurrence and intensity estimation. *2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, *6*, 1-5.

Gutjar, S., de Graaf, C., Kooijman, V., de Wijk, R. A., Nys, A., Ter Horst, G. J., & Jager, G. (2015). The role of emotions in food choice and liking. *Food Research International*, *76*, 216-223.

Isasi, P., & Galván, I. (2004). *Redes neuronales artificiales. Un enfoque práctico*.

Ishiguro, H., Ono, T., Imai, M., Maeda, T., Kanda, T., & Nakatsu, R. (2001). Robovie: An interactive humanoid robot. *Industrial Robot: An International Journal*, *28*. https://doi.org/10.1108/01439910110410051

Ji, Q., Lan, P., & Looney, C. (2006). A probabilistic framework for modeling and real-time monitoring human fatigue. *IEEE Transactions on systems, man, and cybernetics-Part A: Systems and humans*, *36*(5), 862-875.

Joshi, J., Dhall, A., Goecke, R., Breakspear, M., & Parker, G. (2012). Neural-net classification for spatio-temporal descriptor based depression analysis. *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, 2634-2638.

Kaltwang, S., Rudovic, O., & Pantic, M. (2012). Continuous pain intensity estimation from facial expressions. *International Symposium on Visual Computing*, 368-377.

Kapoor, A., Burleson, W., & Picard, R. W. (2007). Automatic prediction of frustration. *International journal of human-computer studies*, *65*(8), 724-736.

Kemper, T. (1987). How many emotions are there? Wedding the social and the autonomic components. *American Journal of Sociology*, 263-298.

King, S. C., Meiselman, H. L., & Carr, B. T. (2010). Measuring emotions associated with foods in consumer testing [Eighth Pangborn Sensory Science Symposium]. *Food Quality and Preference*, *21*(8), 1114-1116. https://doi.org/https://doi.org/10.1016/j.foodqual.2010.08.004

Li, H., & Xu, H. (2020). Deep reinforcement learning for robust emotional classification in facial expression recognition. *Knowledge-Based Systems*, *204*, 106172.

Littlewort, G. C., Bartlett, M. S., & Lee, K. (2007). Faces of pain: automated measurement of spontaneousallfacial expressions of genuine and posed pain. *Proceedings of the 9th international conference on Multimodal interfaces*, 15-21.

Liu, K., Zhang, M., & Pan, Z. (2016). Facial expression recognition with CNN ensemble. *2016 international conference on cyberworlds (CW)*, 163-166.

Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010). The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, 94-101. https://doi.org/10.1109/CVPRW.2010.5543262

Lyons, M. J., Budynek, J., & Akamatsu, S. (1999). Automatic classification of single facial images. *IEEE transactions on pattern analysis and machine intelligence*, *21*(12), 1357-1362.

Maat, L., & Pantic, M. (2007). Gaze-X: Adaptive, Affective, Multimodal Interface for Single-User Office Scenarios. En A. Huang Thomas S.and Nijholt, M. Pantic & A. Pentland (Eds.), *Artifical Intelligence for Human Computing* (pp. 251-271). Springer Berlin Heidelberg.

Matsumoto, D., & Hwang, H. S. (2011). Evidence for training the ability to read microexpressions of emotion. *Motivation and Emotion*. https://doi.org/https://doi.org/10.1007/s11031-011-9212-2

Mavadati, M., Sanger, P., & Mahoor, M. H. (2016). Extended DISFA Dataset: Investigating Posed and Spontaneous Facial Expressions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Mavani, V., Raman, S., & Miyapuram, K. P. (2017). Facial expression recognition using visual saliency and deep learning. *Proceedings of the IEEE international conference on computer vision workshops*, 2783-2788.

Mayer, D., & Salovey, P. (1989). Emotional Intelligence. *Imagination, Cognition, and Personality*, *9*, 185-211.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, *5*(4), 115-133. https://doi.org/10.1007/BF02478259

Moreno-Armendáriz, M. A., Calvo, H., Duchanoy, C. A., Lara-Cázares, A., Ramos-Diaz, E., & Morales-Flores, V. L. (2021). Deep-learning-based adaptive advertising with augmented reality. *Sensors*, *22*(1), 63.

Mowrer, O. (1960). Learning theory and behavior.

Naranjo, R. (2018). Inteligencia emocional. *Academia*. https://www.academia.edu/19034778/Inteligencia_emocional_wiki

Ng, M., Chaya, C., & Hort, J. (2013). Beyond liking: Comparing the measurement of emotional response using EsSense Profile and consumer defined check-

all-that-apply methodologies. *Food Quality and Preference*, *28*(1), 193-205. https://doi.org/https://doi.org/10.1016/j.foodqual.2012.08.012

O'Doherty, J., Rolls, E. T., Francis, S., Bowtell, R., & McGlone, F. (2001). Representation of pleasant and aversive taste in the human brain. *Journal of neurophysiology*, *85*(3), 1315-1321.

Ozdemir, M. A., Elagoz, B., Alaybeyoglu, A., Sadighzadeh, R., & Akan, A. (2019). Real time emotion recognition from facial expressions using CNN architecture. *2019 medical technologies congress (tiptekno)*, 1-4.

Panksepp, J. (1982). Toward a general psychobiological theory of emotions. *The Behavioral and Brain Sciences*, 407-467.

Pantic, M. (2009). Machine analysis of facial behaviour: naturalistic and dynamic.

Pantic, M., Valstar, M., Rademaker, R., & Maat, L. (2005). Web-based database for facial expression analysis. *2005 IEEE international conference on multimedia and Expo*, 5-pp.

Plutchik, R. (1980). A general psychoevolutionary theory of emotion. In R. Plutchik H. Kellerman (Eds.), 3-31.

Porcherot, C., Delplanque, S., Gaudreau, N., & Cayeux, I. (2013). Seeing, smelling, feeling! Is there an influence of color on subjective affective responses to perfumed fabric softeners? [Ninth Pangborn Sensory Science Symposium]. *Food Quality and Preference*, *27*(2), 161-169. https://doi.org/https://doi.org/10.1016/j.foodqual.2012.06.011

Pu, X., Fan, K., Chen, X., Ji, L., & Zhou, Z. (2015). Facial expression recognition from image sequences using twofold random forest classifier. *Neurocomputing*, *168*, 1173-1180.

Quinn, M.-A., Sivesind, G., & Reis, G. (2017). Real-time emotion recognition from facial expressions. *Standford University*.

RAE. (2021). Diccionario de la lengua española RAE - ASALE. *«Diccionario de la lengua española» - Edición del Tricentenario*. Consultado el 27 de noviembre de 2022, desde https://dle.rae.es/emoci%5C%C3%5C%B3n

Renneberg, B., Heyn, K., Gebhard, R., & Bachmann, S. (2005). Facial expression of emotions in borderline personality disorder and depression. *Journal of behavior therapy and experimental psychiatry*, *36*(3), 183-196.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, *65 6*, 386-408.

Rousmans, S., Robin, O., Dittmar, A., & Vernet-Maury, E. (2000). Autonomic nervous system responses associated with primary tastes. *Chemical senses*, *25*(6), 709-718.

Sajjanhar, A., Wu, Z., & Wen, Q. (2018). Deep learning models for facial expression recognition. *2018 digital image computing: Techniques and applications (dicta)*, 1-6.

Samant, S. S., Chapko, M. J., & Seo, H.-S. (2017). Predicting consumer liking and preference based on emotional responses and sensory perception: A study with basic taste solutions. *Food research international*, *100*, 325-334.

Shan, K., Guo, J., You, W., Lu, D., & Bie, R. (2017). Automatic facial expression recognition based on a deep convolutional-neural-network structure. *2017*

IEEE 15th International Conference on Software Engineering Research, Management and Applications (SERA), 123-128.

Thomson, D. M., Crocker, C., & Marketo, C. G. (2010). Linking sensory characteristics to emotions: An example using dark chocolate. *Food Quality and Preference*, *21*(8), 1117-1125.

Tomkins, S. S. (1984a). Affect theory. *In K. R. Scherer P. Ekman (Eds.), Approaches to emotion*, 163-195.

Tomkins, S. S. (1984b). Affect theory., 163-195.

Valstar, M. F., Pantic, M., Ambadar, Z., & Cohn, J. F. (2006). Spontaneous vs. posed facial behavior: automatic analysis of brow actions. *Proceedings of the 8th international conference on Multimodal interfaces*, 162-170.

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, *1*, I-I. https://doi.org/10.1109/CVPR.2001.990517

Wang, S.-J., Yan, W.-J., Li, X., Zhao, G., Zhou, C.-G., Fu, X., Yang, M., & Tao, J. (2015). Micro-Expression Recognition Using Color Spaces. *IEEE Transactions on Image Processing*, *24*(12), 6034-6047. https://doi.org/10.1109/TIP.2015.2496314

XiaoLan, [XunBing, S., & Qi], W. (2010). Micro-expression and Its Applications. *Advances in Psychological Science*, *18*(09), Artículo 1359, 1359-1368. http://journal.psych.ac.cn/adps/EN/abstract/article_2115.shtml

Yan, W.-J., Wu, Q., Liang, J., Chen, Y.-H., & Fu, X. (2013). How Fast are the Leaked Facial Expressions: The Duration of Micro-Expressions. https://doi.org/https://doi.org/10.1007/s10919-013-0159-8

Yang, J., Zhang, F., Chen, B., & Khan, S. U. (2019). Facial expression recognition based on facial action unit. *2019 Tenth International Green and Sustainable Computing Conference (IGSC)*, 1-6.

Zhang, Z., Luo, P., Loy, C. C., & Tang, X. (2018). From facial expression recognition to interpersonal relation prediction. *International Journal of Computer Vision*, *126*(5), 550-569.

Zhao, K., Chu, W.-S., De la Torre, F., Cohn, J. F., & Zhang, H. (2015). Joint patch and multi-label learning for facial action unit detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2207-2216.

Zhao, K., Chu, W.-S., & Zhang, H. (2016). Deep region and multi-label learning for facial action unit detection. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3391-3399.