Aprendizaje profundo en imágenes de alimentos con etiquetas múltiples y ruidosas

Deep learning from noisy multi-label food images

Roberto Morales*[1] https://orcid.org/0009-0008-1796-5278
Ángela Martínez[2] https://orcid.org/0000-0001-6380-5701
Eduardo Aguilar[1] https://orcid.org/0000-0002-2463-0301

Recibido 5 de abril de 2024, aceptado 06 de mayo de 2024
Received: April 05, 2024 Accepted: May 06, 2024

RESUMEN

El rendimiento de los métodos de aprendizaje profundo no solo depende del diseño del modelo, sino también de la cantidad, variedad y calidad de los datos. La recopilación de abundantes datos de repositorios públicos es factible, pero la revisión y anotación resulta laboriosa. Como alternativa, se han desarrollado bases de datos no supervisadas, donde la asignación automática de etiquetas puede generar ruido debido a posibles desviaciones en los datos recopilados. En este trabajo proponemos un modelo de aprendizaje profundo robusto a etiquetas ruidosas para la tarea de clasificación de imágenes de alimentos a nivel de ingredientes, mediante la extensión del método de etiqueta única AFM. La propuesta, ML-AFM, utiliza Attentive Grouping y MixUp para mitigar el ruido de las etiquetas y capturar relaciones complejas entre características y etiquetas en los datos de entrenamiento. Adicionalmente, se adapta la función de activación y pérdida para que sea apropiada a problemas de clasificación multi-etiqueta. La evaluación experimental se realiza sobre el conjunto de datos público Food-101N, con anotaciones ampliadas a nivel de ingredientes. De los resultados se observa que ML-AFM proporciona un mejor rendimiento que el modelo de la línea base, alcanzando un F1 de 86,99%, un AUPRC de 92,85% y un índice de Jaccard de 77,19%. La mejora del rendimiento demuestra la robustez del modelo propuesto frente al problema planteado, lo que respalda su utilidad en aplicaciones prácticas de reconocimiento de alimentos.

Palabras clave: Aprendizaje con etiquetas ruidosas, aprendizaje multi-etiqueta, reconocimiento de ingredientes, aprendizaje profundo, análisis de comida.

ABSTRACT

The performance of deep learning methods depends not only on the model's design but also on the data's quantity, variety, and quality. Collecting abundant data from public repositories is feasible, but their review and annotation are laborious. As an alternative, unsupervised databases have been developed, where the automatic assignment of labels may generate noise due to possible deviations in the collected data. This paper proposes a robust deep-learning model for noisy labels to classify food images at the ingredient level by extending the single-label AFM method. The proposed ML-AFM uses Attentive Grouping and MixUp to mitigate label noise and capture complex feature-label relationships in the training data. Additionally,

[1] Universidad Católica del Norte. Departamento de Ingeniería y Sistemas de Computación. Antofagasta, Chile. E-mail: roberto.morales02@alumnos.ucn.cl; eduardo.aguilar@ucn.cl
[2] Universidad de Valparaíso. Centro de Investigación del Comportamiento Alimentario, Escuela Nutrición y Dietética. Valparaíso, Chile. E-mail: angela.martinez@uv.cl
* Autor de correspondencia: eduardo.aguilar@ucn.cl

the activation and loss function is adapted to be suitable for multi-label classification problems. The experimental evaluation is performed on the public Food-101N dataset, with extended annotations at the ingredient level. The results show that ML-AFM performs better than the reference model, achieving an F1 of 86.99%, an AUPRC of 92.85%, and a Jaccard index of 77.19%. The improved performance demonstrates the proposed model robustness to the given problem, which supports its usefulness in practical food recognition applications.

Keywords: Noisy label learning, multi-label learning, ingredient recognition, deep learning, food analysis.

INTRODUCCIÓN

La obesidad representa un serio desafío global de salud, siendo un factor clave en enfermedades cardiovasculares, musculoesqueléticas y varios tipos de cáncer [1]-[3]. La Organización Mundial de la Salud enfatiza la importancia de patrones alimentarios saludables, reduciendo alimentos ultraprocesados y aumentando legumbres, frutas y verduras, mientras se promueve la actividad física [3]. Para cambiar los patrones de consumo, estrategias de prevención y control de salud son esenciales, incluyendo la evaluación de calidad de la alimentación y análisis de comportamientos alimentarios. Sin embargo, los métodos tradicionales como recordatorios de 24 horas, registros alimentarios y encuestas carecen de precisión por problemas de memoria y falta de información actualizada en softwares, afectando la toma de decisiones informadas por parte de profesionales, investigadores y la población en la elección de una alimentación adecuada [4]. Nuevos enfoques para medir la dieta pueden reducir errores, mejorar la comprensión nutricional y brindar herramientas efectivas para mejorar hábitos alimentarios [5]. La ciencia e inteligencia artificial han avanzado con tecnologías que son capaces de estimar la ingesta calórica, tamaño de porción y perfil nutricional de alimentos [6]. Mejorar la calidad de información alimentaria ayudaría en el autocuidado de la población.

Hasta ahora, avances en los modelos de inteligencia artificial han cobrado importancia en el reconocimiento de alimentos, mediante diversas métricas que indican la precisión del modelo en las predicciones [7]. Sin embargo, el rendimiento del modelo no solo depende de su implementación, sino también de factores externos como la cantidad y calidad de imágenes en el conjunto de datos [8] y la precisión en el proceso de anotación de sus etiquetas. Por su naturaleza, este proceso puede presentar errores, lo que genera ruido en las anotaciones y afecta el rendimiento del modelo, perjudicando la calidad de la clasificación de alimentos. Por lo tanto, es fundamental la creación de modelos robustos capaces de gestionar el ruido presente en las anotaciones en imágenes de comida.

En el contexto de la clasificación de alimentos, las investigaciones se han centrado principalmente en la clasificación de etiqueta-única, en la que cada imagen se categoriza según el contenido principal. Varios estudios han abordado este tema [7], [9]-[13], tanto en la creación de nuevos métodos como en la elaboración de conjuntos de datos que permiten recopilar una cantidad significativa de imágenes. Este tipo de investigaciones requieren de datos anotados para el entrenamiento de los modelos, anotaciones que principalmente son llevadas a cabo por una persona por lo que siempre existirá la posibilidad de error, lo que puede introducir ruido en las etiquetas. Esta es una de las razones que han motivado el desarrollo de métodos basados en etiqueta-única capaces de manejar el ruido generado en el proceso de anotación [14]-[24].

Por otro lado, para tener una mayor certeza sobre todos los alimentos representados en una imagen, es necesario cambiar el enfoque de clasificación de etiqueta-única a múltiples etiquetas. Recientemente este enfoque se ha estudiado en diferentes conjuntos de datos para evaluar cómo reaccionan los métodos en diferentes escenarios de la vida real [25]-[31]. Sin embargo, no hay evidencias de algoritmos con un desempeño robusto ante el ruido en las anotaciones en el problema de reconocimiento de múltiples comidas. Lo que lleva a la necesidad de crear modelos capaces de clasificar múltiples alimentos y tener en cuenta el manejo del ruido en las anotaciones.

La principal diferencia en el presente trabajo radica en la forma de abordar las anotaciones ruidosas en el

proceso de aprendizaje de clasificación de alimentos. Este trabajo se centra en un esquema de clasificación multi-etiqueta, dotando al modelo propuesto la capacidad de clasificar varios simultáneamente. Esto es especialmente importante en el contexto de la alimentación, ya que muchos platos y comidas contienen múltiples ingredientes y nutrientes. Además, al abordar las anotaciones ruidosas, se está mejorando el proceso de aprendizaje del modelo y, por lo tanto, la precisión de la clasificación de los alimentos. En concreto, el presente trabajo se propone un método de clasificación de múltiples alimentos que tenga en cuenta la presencia de ruido en las anotaciones de los datos de entrenamiento durante el proceso de aprendizaje. Para lograr este objetivo, se extiende un método que ha demostrado un buen rendimiento en la clasificación de etiquetas simples con ruido en sus anotaciones y se adapta para su funcionamiento en la clasificación de etiquetas múltiples. A su vez, teniendo en cuenta que actualmente no existe una base de datos de imágenes de múltiples comidas con anotaciones ruidosas, para la evaluación experimental del método propuesto, se usa la base de datos Food-101N [14] y se amplían sus anotaciones considerando las anotaciones a nivel de ingredientes disponibles en Ingredients101 [27] para cada una de las clases de comida existentes en Food-101N.

Las contribuciones del presente trabajo son: 1) Se extiende la base de datos Food-101N con la adición de Ingredients101, una base de datos que incluye información detallada sobre los ingredientes de los alimentos y de esta manera obtener una base de datos multi-etiqueta con anotaciones ruidosas; 2) Se desarrolla un modelo de aprendizaje profundo para la clasificación multi-etiqueta, entrenado con presencia de ruido en las anotaciones; y 3) Se evidencia por parte del modelo propuesto una mayor robustez al ruido de las anotaciones para la clasificación de alimentos con etiquetas múltiples en comparación con el método de la línea base. La experimentación se validó sobre la base de datos Food-101N extendida.

REVISIÓN DE LA LITERATURA

En esta sección se describen los trabajos relacionados con la investigación desarrollada, centrándose en los avances logrados en el campo de los modelos de aprendizaje profundo aplicados al reconocimiento de alimentos, así como en el reconocimiento de múltiples objetos teniendo en cuenta etiquetas ruidosas.

Reconocimiento de comida de etiqueta única con anotaciones ruidosas

En el campo del aprendizaje profundo, uno de los desafíos más comunes es el ruido presente en las etiquetas de entrenamiento, lo cual afecta la calidad y precisión de los modelos de clasificación de imágenes. Para abordar este problema, se han desarrollado varios enfoques y métodos que han sido evaluados en conjuntos de datos de imágenes, especialmente para la clasificación de imágenes con una sola etiqueta.

CleanNet [14] y Co-learning [15] presentan enfoques innovadores para abordar el problema del ruido en las etiquetas en el entrenamiento de modelos de clasificación de imágenes a gran escala. CleanNet identifica el ruido en las etiquetas verificando manualmente solo una fracción de las clases, y luego transfiere esa información a otras clases, reduciendo así la necesidad de supervisión humana. CleanNet logra una disminución del 41,5% en la tasa de error de detección de ruido de etiquetas en comparación con los métodos actuales de supervisión débil. Además, mejora en un 47% el rendimiento de la clasificación de imágenes, incluso con solo el 3,2% de las imágenes verificadas. Por otro lado, Co-learning combina el aprendizaje supervisado y el auto-supervisado utilizando un codificador de características compartido y aplicando restricciones de similitud para mejorar el rendimiento en presencia de etiquetas ruidosas, en experimentos con CIFAR-10, CIFAR-100, Animal-10N y Food-101N, se lograron mejoras en la precisión con datos de ruido simétrico y asimétrico. Los resultados destacados fueron 92,21%, 91,07%, 66,58% y 65,26% de precisión, respectivamente. Otro método que combina el aprendizaje supervisado y el auto-supervisado es LongReMix [22], el cual está basado en una etapa de aprendizaje no supervisado para clasificar muestras de entrenamiento limpias y ruidosas, seguida de una etapa de aprendizaje semi-supervisado para minimizar la tasa de error de variabilidad (EVR) utilizando un conjunto etiquetado formado por muestras clasificadas como limpias y un conjunto no etiquetado con muestras clasificadas como ruidosas. Se observó un rendimiento significativamente mejor en los experimentos con Food-101N y Clothing1M

en comparación con las líneas de base. Esto mejoró la capacidad de generalización de las redes neuronales profundas (DNN). En ambos casos, se optimizó el modelo base (ResNet50), logrando un rendimiento del 82,52% y 72,50% en Food101N y Clothing1M, respectivamente. De igual manera [18] propone un marco de auto-aprendizaje iterativo para el entrenamiento en conjuntos de datos ruidosos del mundo real. Se utiliza la corrección de etiquetas y el entrenamiento conjunto para lograr un marco de entrenamiento efectivo sin necesidad de redes o supervisión adicionales. También el método NoiseRank [16] gestiona el ruido en las etiquetas mediante un algoritmo de clasificación de ruido de etiquetas no supervisado basado en Campos Aleatorios de Markov. El modelo de dependencia se utiliza para estimar y clasificar las instancias según la probabilidad posterior de estar etiquetadas incorrectamente en el conjunto de datos, este enfoque fue creado con la capacidad de ser flexible a algoritmos de clasificación de ruido de etiquetas no supervisadas. NoiseRank supera a los métodos de clasificación actuales en Food101-N, logrando un destacado 85,78% de precisión en el top 1, en comparación con el 85,11% sin entrenamiento no supervisado. Además, en Clothing-1M, NoiseRank mejora la precisión de clasificación de 68,94% a 73,82% (una reducción de errores del 16%) incluso en condiciones de alto ruido y sin supervisión. En el método MSLG [23] nuevamente se propone un marco agnóstico del tipo de backbone que se emplee, pero en este caso se propone un enfoque de meta-aprendizaje que busca una distribución óptima de etiquetas suaves sujeta a un meta-objetivo, que es minimizar la pérdida del pequeño meta conjuntó de datos. Posteriormente, la red se entrena con estas etiquetas suaves predichas. Estas dos etapas se repiten consecutivamente durante el proceso de entrenamiento. MSLG supera con claridad a métodos actuales, siendo modelo-independiente y logrando 75% de precisión incluso en condiciones extremas con 80% de ruido. Requiere pocos metadatos; 1.000 metadatos logran el mejor rendimiento en CIFAR10 con 50.000 datos ruidosos. En Clothing1M, supera con 76,02% de precisión, 2,3% por encima del estado del arte. En el mismo sentido del meta aprendizaje se propone WarPi [21] concebido con el fin de aprender a rectificar de manera adaptativa el proceso de entrenamiento en el escenario de meta-aprendizaje, este método agrega el enfoque probabilístico mediante la formulación del proceso de aprendizaje como un modelo probabilístico jerárquico y considerando el vector de rectificación como una variable latente, para lograr una estimación efectiva de la posterior predictiva. Los resultados superan consistentemente a otros enfoques de meta-aprendizaje como L2RW, MWNet y MLC. En CIFAR-100 con un 40% de ruido asimétrico, supera a MWNet en un 3,73%. Además, establece un nuevo estado de la técnica en conjuntos de datos desafiantes del mundo real, cómo Clothing1M y Food-101N, con mejoras significativas en comparación con métodos de meta-aprendizaje.

Por otro lado, el método Probabilistic Noise Predicción (PNP) [32] se centra en aprender de etiquetas ruidosas utilizando un enfoque probabilístico. En lugar de umbrales de selección difíciles de ajustar, se utilizan dos redes para predecir la categoría y el tipo de ruido. Se emplea una tarea de regresión para mejorar la predicción del tipo de ruido y se utiliza la regularización de consistencia para mejorar la capacidad de discriminación. Los resultados experimentales en conjuntos de datos sintéticos y del mundo real destacan la superioridad del método propuesto. Tanto PNP-Hard como PNP-Soft lograron precisiones de prueba del 87,31% y 87,50% en Food101N, superando enfoques líderes y demostrando la efectividad de PNP en aplicaciones a gran escala del mundo real. También se propuso un algoritmo progresivo de corrección de etiquetas [33] que corrige las etiquetas de forma iterativa y refina el modelo sobre la base de esta suposición de ruido. Se propone un método de recalibración de datos que está garantizado teóricamente para converger para ser consistente con el clasificador de Bayes. Enfoques como el presentado en [34] se basan en elementos comunes del aprendizaje profundo para abordar el problema de las etiquetas ruidosas. Utilizan una matriz de confusión para representar la distribución del ruido y eliminan la capa de ruido antes de hacer predicciones en un conjunto de prueba limpio. Otros métodos existentes dividen los datos de entrenamiento en subconjuntos limpios y ruidosos [17], pero pueden confundir las muestras etiquetadas incorrectamente. Este enfoque separa los datos ruidosos y limpios y entrena una red con etiquetas ruidosas y suaves, logrando mejoras significativas en el rendimiento. Otro enfoque, Tripartite [35], divide los datos de entrenamiento en tres subconjuntos: duro, ruidoso y limpio, utilizando criterios basados en las predicciones inconsistentes y la discrepancia entre

predicciones y etiquetas. Su objetivo es minimizar el impacto de las etiquetas ruidosas y maximizar el valor de los datos ruidosos mediante estrategias auto-supervisada en datos etiquetados ruidosos. Por otra parte, MORPH [36] consta de dos fases: una para estimar el punto de transición óptimo y otra para completar el entrenamiento utilizando un conjunto seguro de alta calidad. MORPH ha demostrado mejoras significativas en robustez y eficiencia en comparación con los métodos actuales en diversos conjuntos de datos ruidosos.

Otros enfoques novedosos son MetaCleaner [24], Jo-SRC [20] y Attentive Feature MixUp (AFM) [37]. MetaCleaner [24] consta de dos submódulos: 1) Ponderación Ruidosa y 2) Generación Limpia. La primera estima la importancia de las imágenes en un subconjunto mediante la comparación de representaciones semánticas. El segundo genera una representación limpia considerando los pesos de las diferentes representaciones de las imágenes. Esto mejora la robustez y la generalización de las redes neuronales profundas a etiquetas ruidosas. En los experimentos con Food-101N y Clothing1M, se alcanzó un rendimiento notablemente superior a las líneas de base, mejorando la capacidad de generalización de las DNN. En ambos casos, el método mejoró el modelo base (ResNet50), logrando un rendimiento del 82,52% y 72,50% en Food101N y Clothing1M, respectivamente. En cuanto a Jo-SRC [20], se aborda el problema del ruido en las etiquetas al identificar muestras limpias utilizando la divergencia de Jensen-Shannon y distinguiendo muestras ruidosas según su consistencia *in-distribution* (ID) y *out-of-distribution* (OOD). Propone una pérdida conjunta que combina términos de clasificación y regularización de consistencia, mejorando el rendimiento y la robustez del modelo frente a etiquetas ruidosas. En CIFAR100NC, Jo-SRC supera a los métodos modernos con alta precisión de prueba, logrando 79,20% en datos simétricos y 78,60% en asimétricos. En CIFAR80N-O, que simula situaciones reales, destaca en todos los niveles de simetría, incluso en un desafiante 80%. En Clothing1M, mejora ResNet-18 en 1,48% y ResNet50 de 74,76% a 75,93%. En Food101N, Jo-SRC lidera con 86,05%, superando DeepSelf en 1,55%. Finalmente, Attentive Feature MixUp (AFM) [37] aborda el problema de las etiquetas ruidosas en modelos de aprendizaje profundo al asignar pesos de atención a las muestras, permitiendo que los modelos presten más atención a las muestras limpias y menos a las ruidosas. AFM mejora la calidad de las muestras mediante la interpolación de muestras agrupadas con ruido suprimido. No requiere conjuntos de datos limpios adicionales y optimiza conjuntamente los pesos de interpolación con los clasificadores. AFM logra resultados destacados en conjuntos de datos ruidosos del mundo real, como Food-101N y Clothing1M.

En la Tabla 1 se muestran los métodos utilizados en las investigaciones que no solo hacen público el código fuente, sino que también trabajan con la base de datos Food-101N. Los métodos se encuentran ordenados ascendentemente en base al rendimiento obtenido.

Reconocimiento de múltiples objetos con anotaciones ruidosas

Recientemente, aunque no en el dominio de la comida, se han propuesto métodos para abordar el reconocimiento de objetos con anotaciones ruidosas en la clasificación de imágenes considerando una

Tabla 1. Resultados de los métodos estudiados, públicamente disponibles, sobre la base de datos Food-101N en términos de accuracy.

Reference	Method	Accuracy
ICPR 2021 [24]	MSLG	79,06%
CVPR 2018 [15]	CleanNet Whard	83,47%
CVPR 2018 [15]	CleanNet Wsoft	83,95%
CVPR 2022 [33]	PNP	85,28%
Pattern Recognition 2022 [22]	WarPI	85,95%
CVPR 2021 [21]	Jo-SRC	86,66%
ECCV 2020 [38]	AFM	87,27%

predicción a nivel de múltiples etiquetas. Estos enfoques permiten asignar múltiples etiquetas a una sola imagen, modelando las relaciones complejas entre las etiquetas y mejorando la precisión de la clasificación en comparación con técnicas que no consideran un diseño consciente del ruido en las etiquetas del conjunto de datos. CCMN [25] se basa en estimadores imparciales eficientes para abordar los problemas de CCMN (clasificación multi-etiqueta con etiquetas ruidosas condicionales a clase) se demostró su consistencia en términos de pérdida de Hamming y pérdida de clasificación. Además, se propone en conjunto con este, un método novedoso llamado uPML para resolver problemas de PML (Probabilistic Multi-Label), que se puede considerar como un caso especial dentro del marco de CCMN. Por otra parte, el artículo [28] presenta dos enfoques de aprendizaje semi-supervisado para la clasificación de múltiples etiquetas. El primero utiliza un conjunto de datos de poses corporales sobre un enfoque de etiquetas binarias de prendas, mejorando el rendimiento mediante la corrección manual de etiquetas y el entrenamiento con etiquetas ruidosas. Por otra parte, se presenta una red de profesor-estudiantes [29] que utiliza una transformación no lineal de características para aprender de manera eficiente a partir de etiquetas ruidosas masivas. Los resultados experimentales muestran que el enfoque propuesto logra superar el estado del arte en la clasificación multi-etiqueta a partir de etiquetas ruidosas. Por otra parte, en [30], se demostró que el ruido sustractivo tiene un impacto negativo significativo en el rendimiento de los modelos en comparación con el ruido aditivo. Se recomendó el uso de métodos de mejora de etiquetas para casos de bajo ruido y métodos de regularización para casos de alto ruido. Además, se resaltó la importancia de considerar diferentes tipos de ruido y priorizar el uso de ruido aditivo al crear conjuntos de datos multi-etiqueta.

A partir de la evidencia de la literatura reciente, en la que sólo se ha abordado la clasificación de alimentos con etiquetas ruidosas a nivel de etiqueta-única, en este trabajo de investigación se propone adaptar el modelo que proporciona el mejor rendimiento en la clasificación de alimentos a nivel de etiqueta-única para abordar la clasificación multi-etiqueta mediante la adaptación de componentes claves, como la función de activación y la función de pérdida específica de la clasificación multi-etiqueta. Nuestro propósito es disponer de un método robusto al ruido que pueda estar contenido en las anotaciones de los datos de entrenamiento con la intención de disponer de un modelo de clasificación multi-etiqueta preciso que sea capaz de aprender sobre datos con etiquetas ruidosas. Con ello se pretende aliviar el elevado coste que supone el proceso de revisión y corrección de las etiquetas disponibles.

METODOLOGÍA

En esta sección se describe el método AFM teniendo en cuenta las principales características de este y la adaptación del método para que pueda ser usado sobre datos con anotaciones de multi-etiqueta.

Attentive feature mixup

AFM se utiliza para mejorar el aprendizaje en presencia de imágenes ruidosas. Mediante el uso de redes convolucionales (CNN), AFM divide las imágenes en grupos pequeños y aprende pesos de atención para cada muestra, estos pesos indican la importancia relativa de cada muestra dentro del grupo y se utilizan para determinar cuánta atención se debe prestar a cada muestra. Luego, se generan nuevas muestras y etiquetas suaves mediante la interpolación y mezcla de estas muestras (Figura 1). El AFM utiliza un enfoque de minimización del riesgo vecinal para mitigar el sobreajuste y optimiza conjuntamente los pesos de interpolación y el clasificador. AFM hereda parcialmente la minimización del riesgo vecinal de confusión para aliviar el sobreajuste con interpolaciones masivas. Además, con la optimización conjunta de los pesos de interpolación de mezcla y el clasificador, AFM mejora la mezcla al muestrear menos vectores de características-objetivo en torno a datos mal etiquetados. El enfoque AFM consta de cuatro pasos fundamentales. En primer lugar, se seleccionan aleatoriamente K muestras repetidamente para construir grupos. Luego, se utilizan capas completamente conectadas (FC) para mapear las muestras de cada grupo en nuevas características incorporadas, estas nuevas características se obtienen al combinar y procesar interacciones entre muestras dentro de un grupo. Estas nuevas representaciones se obtienen utilizando capas FC para proyectar linealmente las características originales de las muestras. Estas proyecciones se utilizan posteriormente en el cálculo de los pesos de atención dentro del grupo. Este proceso de interacción y generación de nuevas representaciones

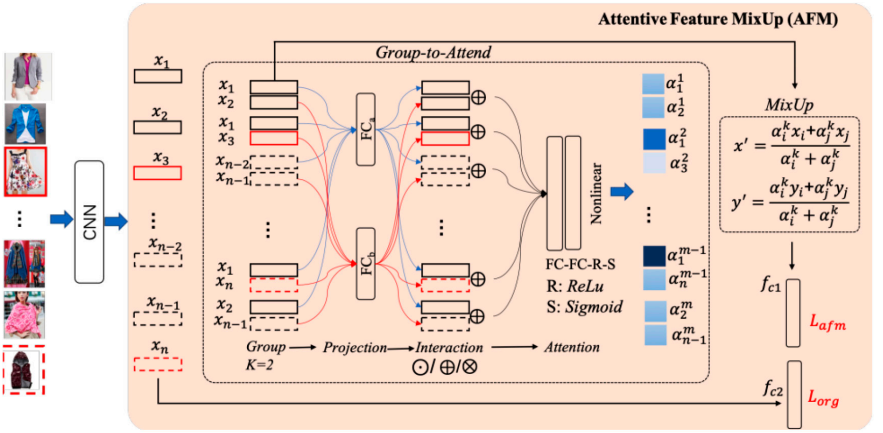

Figura 1. Flujo de proceso del método AFM propuesto para la clasificación de etiqueta-única en muestras ruidosas [37].

permite obtener pesos de atención significativos que distinguen entre muestras limpias y muestras con etiquetas incorrectas. A continuación, se fomenta la interacción entre las muestras para el aprendizaje de pesos grupales. Por último, se utiliza una red de autoatención para estimar los pesos de atención grupales. La interacción entre las funciones y las capas de proyección desempeñan un papel crucial en el aprendizaje de pesos de atención significativos.

El módulo AG (Attentive Grouping) reduce el impacto de las muestras con etiquetas ruidosas y disminuye la proporción de grupos completamente ruidosos. Además, permite que los grupos parcialmente ruidosos proporcionen supervisión útil a través de la red de atención bien entrenada. Sin embargo, un valor grande de K puede llevar a características excesivamente suavizadas, lo cual es perjudicial para el aprendizaje de características discriminativas. El módulo AG, considerando un $K=2$, está representado por la siguiente ecuación (1):

$$\frac{N_{noisy}}{N_{total}} > \frac{N_{noisy}\left(N_{noisy}-1\right)}{N_{total}\left(N_{total}-1\right)} \approx \frac{N_{noisy}^2}{N_{total}^2} \tag{1}$$

Donde N_{noisy} y N_{total} representan el número de imágenes ruidosas y el total de imágenes de un conjunto de datos ruidosos, respectivamente; y $\frac{N_{noisy}}{N_{total}}$ corresponde a la proporción de imágenes ruidosas.

El módulo de mezcla (MixUp) [38] realiza la interpolación de vectores virtuales de características y objetivos durante el entrenamiento. El módulo de mezcla se puede formalizar mediante la ecuación (2) de la siguiente manera:

$$\begin{aligned} x' &= \frac{1}{\sum a}\left(a_i\, x_i + a_j\, x_j\right), \\ y' &= \frac{1}{\sum a}\left(a_i\, y_i + a_j\, y_j\right), \end{aligned} \tag{2}$$

Donde x' e y' representan las características interpoladas y la etiqueta suave, los términos a_i y a_j representan los coeficientes de mezcla que controlan el grado de interpolación entre las características extraídas (x_i y x_j) y las etiquetas correspondientes (y_i y y_j) de las imágenes i-ésima y j-ésima respectivamente. Estos coeficientes son aprendidos durante el entrenamiento del modelo.

La pérdida total en el entrenamiento considera a f_{c1} y f_{c2} como las notaciones de los clasificadores para los datos interpolados y las muestras originales respectivamente. La fórmula de la pérdida de entrenamiento en un mini lote se puede expresar mediante la ecuación (3) de la siguiente manera:

$$\begin{aligned} L_{total} &= \lambda L_{afm} + \left(1+\lambda\right) L_{org} = \\ &\frac{\lambda}{m}\sum_{i=1}^{m} L\left(f_{c1}\left(x_i'\right), y_i'\right) + \\ &\frac{\left(1+\lambda\right)}{n}\sum_{i=1}^{n} L\left(f_{c2}\left(x_i\right), y_i\right) \end{aligned} \tag{3}$$

Donde n es el tamaño del lote, m es el número de interpolaciones y λ es el peso de compensación.

Se utilizó la pérdida de Entropía Cruzada tanto para L_{afm} como para L_{org}, lo que permite que AFM se visualice como un regularizador de redes profundas. Además, los parámetros f_{c1} y f_{c2} pueden compartirse, este fenómeno se debe a que tanto las características originales como las interpolaciones tienen las mismas dimensiones. Durante el proceso de MixUp, se generan nuevas muestras interpolando las características de las muestras originales. Estas características interpoladas mantienen la misma dimensionalidad que las características originales. Esto implica que no es necesario tener clasificadores separados para las muestras originales y las interpolaciones, ya que ambas comparten la misma representación dimensional. Por lo tanto, se puede ahorrar memoria y recursos computacionales al compartir los parámetros entre los clasificadores.

Multi-label attentive feature mixup (ML-AFM)
Teniendo en cuenta que AFM se diseñó para el reconocimiento de imágenes con anotaciones ruidosas de etiqueta-única, en este artículo se analiza su compatibilidad y se adecua para que sea posible su uso en problemas de clasificación de multi-etiqueta. En AFM se destacan dos módulos principales para abordar el entrenamiento de un modelo de aprendizaje profundo considerando datos con anotaciones ruidosas, estos son AG y MixUp. Ambos módulos pueden ser usados en su formulación original para el problema de clasificación multi-etiqueta. En el caso del AG, cuyo objetivo es reducir la proporción de grupos puros de imágenes ruidosas y que los grupos parciales de imágenes ruidosas (en los que algunas imágenes se corrigen correctamente) pueden proporcionar un seguimiento útil por parte de una red de atención bien entrenada, se mantiene sin modificaciones debido a que se aplica a nivel de imagen y, por tanto, se puede utilizar independientemente del número de etiquetas ruidosas que posea la imagen. Por otra parte, en el caso de MixUp, para la mezcla de anotaciones (y') se usa en formato one-hot encoding por lo que también es compatible. Sin embargo, a diferencia de la etiqueta-única, en multi-etiqueta las anotaciones mezcladas resultantes pueden contener más de 2 etiquetas suavizadas. Por otra parte, en el caso de MixUp, para la mezcla de anotaciones (y') se usa en formato one-hot encoding por lo que también es compatible. El módulo AG resulta beneficioso para el aprendizaje multi-etiqueta, ya que puede ayudar a mitigar el impacto del ruido en las etiquetas y mejorar la calidad de las predicciones de múltiples etiquetas. Al reducir la proporción de grupos puros de imágenes ruidosas, el enfoque podría aumentar la capacidad del modelo para aprender características discriminativas y mejorar la precisión del aprendizaje multi-etiqueta. Por otra parte, en el caso de MixUp, la importancia en la clasificación de imágenes multi-etiqueta radica en su capacidad para generar imágenes sintéticas que combinan características y etiquetas de diferentes imágenes. Al crear estas imágenes "mixtas" que contienen múltiples etiquetas, se mejora la capacidad del modelo de aprendizaje automático para capturar relaciones más complejas entre características y etiquetas en los datos de entrenamiento. Además, MixUp también puede ayudar a suavizar las transiciones entre diferentes clases y evitar la sobre confianza en predicciones incorrectas. Al interpolar características y etiquetas, se reduce la probabilidad de que el modelo realice predicciones extremas o confíe demasiado en características específicas, lo que puede llevar a errores de clasificación en la clasificación multi-etiqueta.

Los principales ajustes en método corresponden a la modificación de la función de activación usada en la capa de salida (*logits*) y a la función de pérdida. En un problema de clasificación de etiqueta-única el propósito del modelo es producir una predicción que categorice el contenido principal de la imagen de entrada. Por este motivo, comúnmente se adopta una función softmax en la capa de salida para asegurar obtener la categoría más probable para la imagen objetiva. A diferencia de un problema de etiqueta única, en un enfoque de múltiples etiquetas se obtienen predicciones para todos los objetos presentes en la imagen. Por este motivo, se cambia la activación a una sigmoide, tal y como es usado en el resto de los trabajos de multi-etiqueta revisados. La función de activación sigmoide se utiliza para transformar las salidas del modelo en valores entre 0 y 1, que pueden interpretarse como probabilidades para cada una de las clases de manera independiente. Con respecto a la función de pérdida, se sustituye la pérdida Categorical Cross entropy por Binary Cross Entropy. La función de pérdida binaria cruzada se utiliza para calcular la pérdida entre las predicciones y las etiquetas reales. Toma en cuenta la probabilidad de cada clase individualmente y calcula la pérdida para cada clase por separado. Esta función de pérdida acepta como entrada las predicciones (valores logit) generadas por el modelo y las etiquetas reales de los ejemplos.

Dada la n-ésima imagen de entrada x_n, su anotación y_n y la salida del modelo sigmoide $f(x_n)$, la pérdida L_{BCE} se define formalmente por la ecuación (4):

$$L_{BCE} = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{C}\sum_{c=1}^{C} y_{n,c}\log\sigma\left(f\left(x_{n,c}\right)\right) + \left(1-y_{n,c}\right)\log\left(1-\sigma\left(f\left(x_{n,c}\right)\right)\right) \quad (4)$$

Teniendo en cuenta el cambio de la función de pérdida el cálculo de la pérdida total para el nuevo modelo se tiene la ecuación (5).

$$\begin{aligned} L_{total} &= \lambda L_{afm} + \left(1+\lambda\right) \\ L_{org} &= \frac{\lambda}{m}\sum_{i=1}^{m} L_{BCE}\left(f_{c1}\left(x_i^{'}\right), y_i^{'}\right) + \\ &\frac{\left(1+\lambda\right)}{n}\sum_{i=1}^{n} L_{BCE}\left(f_{c2}\left(x_i\right), y_i\right) \end{aligned} \quad (5)$$

Donde L_{afm} es la pérdida asociada a las interpolaciones generadas por el modelo, L_{org} es la pérdida asociada a las muestras originales, λ es un peso de compensación que equilibra las dos pérdidas, m es el número de interpolaciones en el mini-batch, n es el tamaño del mini-batch, L_{BCE} es la función de pérdida Binary Cross Entropy, f_{c1} y f_{c2} son los clasificadores para las interpolaciones y las muestras originales, respectivamente. $x_i^{'}$ y $y_i^{'}$ son una interpolación y su etiqueta correspondiente, x_i y y_i son una muestra original y su etiqueta correspondiente.

Esta ecuación muestra cómo se calcula la pérdida total del modelo mediante la combinación de las pérdidas de las interpolaciones y las muestras originales utilizando la función de pérdida Binary Cross Entropy. El peso λ controla la importancia relativa de cada término en la pérdida total.

Conjuntos de datos

En esta sección se describen los conjuntos de datos usados para la confección del conjunto de datos propuesto (Food-101N e Ingredients101) y el conjunto de datos resultante (ML-Food-101N).

Food-101N

Food-101N [14] se construyó para abordar el ruido en las etiquetas, generado mediante la adquisición de imágenes con una mínima supervisión humana. Este conjunto de datos consta de alrededor de 310.009 imágenes de recetas de comida clasificadas en 101 clases o categorías. Aunque comparte las mismas clases que el conjunto de datos Food-101 [39], Food-101N contiene muchas más imágenes y presenta un mayor nivel de ruido en las anotaciones. En cuanto a las anotaciones, Food-101N contiene dos tipos de etiquetas para las imágenes: 1. Etiquetas de clase, que describen la clase a la que pertenece una imagen; 2. Etiquetas de verificación, estas etiquetas indican si la etiqueta de clase asignada a una imagen es correcta Se estima que existen aproximadamente un 20% de anotaciones ruidosas. Se han asignado manualmente etiquetas de verificación a 52.868 imágenes para el entrenamiento (aproximadamente 523 imágenes por clase) y a 4.741 imágenes para la validación (aproximadamente 47 imágenes por clase), estos datos reflejan que 10.664 imágenes en el conjunto de datos Food-101N tienen etiquetas de clase que se consideran ruidosas debido a la tasa estimada de ruido del 20%. El total de imágenes de Food-101N se consideran para el entrenamiento y para la prueba las imágenes del conjunto de prueba pertenecientes a Food-101. La Figura 2 muestra una representación visual de una clase específica dentro del conjunto de datos.

Ingredients101

Ingredients101 [27] es un conjunto de datos para el reconocimiento de ingredientes. Consiste en una lista de los ingredientes más comunes para cada uno de los 101 tipos de alimentos contenidos en el conjunto de datos Food-101, lo que da un total de 446 ingredientes únicos (9 por receta en promedio) y 227 ingredientes únicos para su versión simplificada. El conjunto de datos se dividió en conjuntos de entrenamiento, validación y prueba, asegurándose de que los 101 tipos de alimentos estuvieran balanceados. La base de datos presenta una lista reducida de ingredientes que simplifican la lista original en función de su categoría general, por ejemplo, "pasta de tomate" y "rodajas de tomates" se categoriza bajo la categoría principal "tomates". La Figura 3 muestra imágenes del conjunto de datos Food-101 con las etiquetas de ingredientes correspondientes.

ML-Food-101N

Ante la no existencia de un conjunto de datos con anotaciones ruidosas a nivel de multi-etiqueta dentro del dominio de los alimentos, y considerando que tanto el conjunto de datos Food-101N como

Figura 2. Ejemplo de imágenes recuperadas desde la web usando como término de consulta ‘waffles’. Las imágenes con anotaciones ruidosas son resaltadas en rojo [14].

Figura 3. Ejemplo de comidas en Ingredients101 junto con sus anotaciones. En verde se resaltan las anotaciones limpias y en rojo las ruidosas.

Ingredients101 se crearon basándose en Food-101, lo que significa que ambos conjuntos contienen exactamente las mismas clases. Se propone el conjunto de datos ML-Food-101N, el cual considera los ingredientes proporcionados en Ingredients101 para cada una de las imágenes pertenecientes a Food-101N. Como resultado, para las imágenes recuperadas correctamente, los ingredientes asociados serán precisos, mientras que, para aquellas con ruido en las etiquetas, algunos de los ingredientes también serán incorrectos. Esto último puede ser observado en la Figura 4, específicamente en ella se presentan algunas imágenes ruidosas de la clase “*chicken_quesadilla*” extraídas de ML-Food-101N, junto con los respectivos ingredientes proporcionados por el conjunto de datos Ingredients101. Aunque estas imágenes no representan fielmente la clase en sí, es posible observar que algunos de los ingredientes que la conforman están presentes en algunas de ellas. Por lo tanto, no todos los ingredientes en las imágenes ruidosas son incorrectos.

Experimentación

En esta sección se describe la configuración experimental y se presentan los resultados de la experimentación para el modelo de la línea base y el propuesto.

Configuración experimental

Para la experimentación se consideran dos modelos, ResNet50 y ML-AFM. En el caso de ML-AFM, el backbone usado es ResNet50. En ambos modelos se establece una activación sigmoide y un total de 227 neuronas en la capa de salida, donde el total de neuronas corresponden a la cantidad de ingredientes únicos disponibles en la base de datos objetiva, ML-Food-101N. Para asegurar que los resultados sean comparables, en ambos modelos se utilizan los mismos hiper-parámetros de entrenamiento y técnicas de aumento de datos. Los modelos se entrenan durante 50 épocas usando *Stochastic Gradient Descent* (SGD) con un momentum de 0.9 como optimizador y un *batch-size* de 128. En

Figura 4. Ejemplo de comidas en ML-Food-101N junto con sus anotaciones. En verde se resaltan las anotaciones limpias y en rojo las ruidosas.

cuanto a la tasa de aprendizaje, inicialmente se fija en 1e-2 y luego se reduce en una décima parte en la época 25 y en la época 45. También se utiliza un weight decay de 1e-4 para evitar que los modelos se sobren ajusten. En cuanto a los datos de entrada, estos son redimensionados considerando 256 píxeles para el lado mínimo, la intensidad de los píxeles es escalada en un rango de 0-1 y posteriormente se normalizan restando la media y dividiendo la desviación estándar. Durante el entrenamiento los datos son aumentados por medio de cortes aleatorios de 224x224 píxeles y volteos horizontales aleatorios. Durante la prueba, se realiza un corte en el centro de las imágenes con una dimensión de 224x224 píxeles. Finalmente, en cuanto los parámetros de ML-AFM, los valores predeterminados para λ y K son 0,5 y 2 respectivamente.

Para la evaluación cuantitativa de los resultados, se seleccionan las métricas tradicionales usadas en problemas de clasificación de multi-etiqueta, estas son: Precision, Recall, F1 índice de Jaccard y AUPRC. Todos los experimentos se realizaron usando el framework de deep learning PyTorch v1.10, los cuales fueron ejecutados en un servidor que posee dos tarjetas gráficas NVIDIA de gama media (RTX 3060, 12GB de VRAM), 16GB RAM y un procesador AMD Ryzen 5600X.

RESULTADOS

Se presentan los resultados experimentales de forma cuantitativa y cualitativa para evaluar el rendimiento del modelo propuesto con relación a los datos relacionados con el conjunto de datos de comida que hemos ampliado. De esta manera, se obtiene un conocimiento detallado sobre su desempeño en el ámbito del aprendizaje con etiquetas ruidosas.

La Tabla 2 presenta los resultados obtenidos para cada una de las métricas evaluadas en nuestro estudio sobre todas las imágenes del conjunto de prueba. Aunque existen diversos trabajos sobre las bases de datos Food-101 y Food-101N (versión con ruido), en este proyecto se extienden las anotaciones de esta última base de datos para convertirla en una base de datos de múltiples etiquetas. Mediante la combinación de Food101-N con la base de datos Ingredients101, se logró obtener un nuevo conjunto de datos multi-etiqueta para la evaluación de ambos: el modelo de línea base (ResNet50) y nuestro modelo propuesto (ML-AFM). Los resultados demuestran

Tabla 2. Resultado de la sobre la base de datos Food-101N con enfoque multi-etiqueta en comparación con el baseline.

Método	Precision	Recall	F1	AUPRC	Jaccard
ResNet50	94,74%	74,01%	83,10%	92,21%	71,57%
ML-AFM	91,33%	83,05%	86,99%	92,85%	77,19%

de manera consistente que nuestro modelo supera en todas las métricas evaluadas al modelo ResNet50 con la excepción de la métrica Precision. Esto sugiere que el ruido afecta mayormente al modelo de línea base, limitando su capacidad de reconocimiento en comparación con ML-AFM, evidenciado con un Recall de aproximadamente un 10% inferior que el modelo propuesto. Además, se interpreta que las predicciones otorgadas por el modelo de línea base se especializan en un subconjunto de ingredientes (por ejemplo, aquellos ingredientes menos propensos a ser ruidosos) y por este motivo la Precision es superior (alrededor de 3,5%). Estos hallazgos respaldan la efectividad de la solución propuesta para abordar el problema de la clasificación multi-etiqueta de alimentos con etiquetas ruidosas.

Con relación a la Figura 5 y Figura 6, se muestra el rendimiento de la métrica Recall a nivel de ingredientes tanto para el modelo base como para el modelo propuesto. Estas figuras nos brindan información sobre aquellos ingredientes que los modelos son capaces de identificar correctamente. Ambos modelos presentan un promedio de Recall por encima del 80%. Sin embargo, en el caso del modelo base se observa que algunos ingredientes no se clasifican en ninguna imagen, lo que indica que presenta dificultades para identificarlos. Por otro lado, el modelo propuesto muestra una mejora en el rendimiento de cada ingrediente y se evidencia que ofrece mejores resultados para todos los ingredientes, lo que indica que es capaz de identificar correctamente un mayor número de ingredientes en comparación con el modelo base.

En la Tabla 3, se presentan los ingredientes pertenecientes al Top 6 de los mejores porcentajes de Recall obtenido por cada modelo. En términos generales, ML-AFM logra un desempeño superior en términos de Recall al clasificar los ingredientes, obteniendo en el mejor de los casos un 99,6%. Además, se observa que en ambos hay coincidencia en los ingredientes mejores identificados en ambos modelos, lo que sugiere que existen algunos ingredientes que tienden a ser más simples de detectar en una imagen.

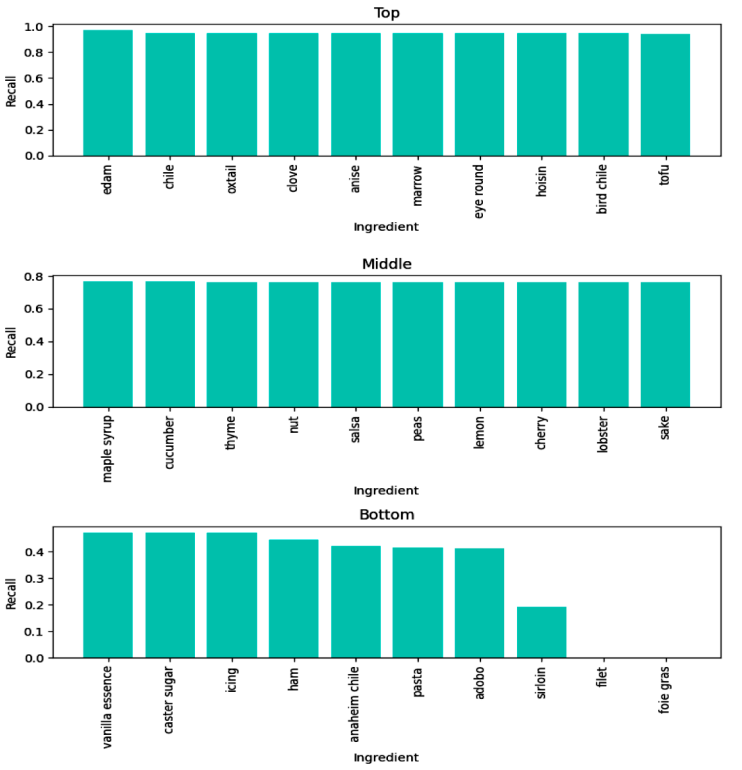

Figura 5. Mejores, medios y peores ingredientes reconocidos con el modelo ResNet50.

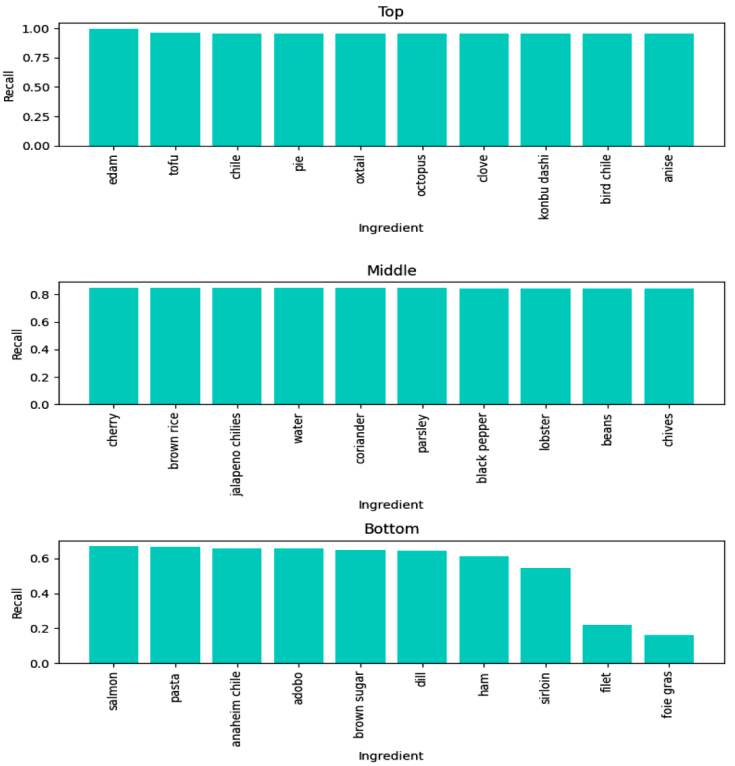

Figura 6. Mejores, medios y peores ingredientes reconocidos con el modelo ML-AFM.

Tabla 3. Ingredientes pertenecientes a los mejores 6 porcentajes de Recall obtenidos para cada modelo.

Ingredientes	**ResNet50**	**Ingredientes**	**ML-AFM**
Edam	96,8%	edam	99,6%
chile,oxtail,clove,anise,marrow, eye round, housing, bird chile	94,4%	tofu	96,4%
Tofu	94,0%	chile, pie, oxtail, octopus, clove, kombu dashi, bird chile, anise, hoisin, marrow, eye round,	95,6%
Sugar	92,3%	sashimi, mirin,	94,8%
turmeric,cress	92,1%	turmeric, cress,	94,4%
Salt	91,5%	oyster	94,2%

En el caso del modelo de la línea base, resulta llamativo que algunos de los ingredientes mejor identificados corresponden a ingredientes no visibles, tal es el caso de *salt* y *sugar*. Esto sucede en mejor frecuencia en el caso de ML-AFM. Por último, además de que ML-AFM proporciona un mayor porcentaje, este porcentaje también se obtiene en una mayor cantidad de ingredientes en comparación con ResNet50.

Por otro lado, la Tabla 4 muestra la lista de ingredientes pertenecientes a los peores resultados para cada modelo. Aquí se destacan los ingredientes con porcentajes de acierto más bajos. Se puede apreciar que ambos modelos enfrentan dificultades al clasificar ciertos ingredientes, siendo el modelo ML-AFM el que obtiene mejores resultados en esta comparativa. Interesante hay que destacar que existen 3 ingredientes (*sirloin*, *filet*, *foie gras*) que son mal clasificados por

ambos modelos, en los cuales se observa una amplia mejora (sobre un 15%) con el modelo propuesto.

En la Tabla 5, se presentan los resultados sobre el conjunto de entrenamiento, particularmente aquellos obtenidos sobre los datos validados (*noisy* o *clean*) que provee el conjunto Food-101N. Se puede observar que el modelo ResNet50 obtiene un $F1$ más bajo en comparación con el modelo ML-AFM en ambos conjuntos de datos ruidosos y limpios, lo que indica que al no tratar el ruido proveniente de las anotaciones el modelo se subajusta. Además, se muestra el porcentaje de datos memorizados con respecto al total de cada conjunto de datos, es decir la predicción del modelo es exactamente la mismas que las anotaciones. Se observa que tanto ResNet50 como ML-AFM son capaces de memorizar una cantidad considerable de casos en los conjuntos de datos ruidosos y limpios. Sin embargo, los valores específicos de memorización varían según el modelo y el conjunto de datos. Particularmente, resulta llamativo este porcentaje sobre los datos con anotaciones ruidosas, donde el modelo propuesto tiene una mayor memorización en comparación con ResNet50. De todos modos, como se observa en Tabla 1, el modelo ML-AFM exhibe un mejor rendimiento general en términos de clasificación a pesar de la memorización evidenciada.

Los resultados cualitativos brindan una representación visual del desempeño de ambos modelos. En la Figura 7 y Figura 8, se muestra la métrica F1 para cada modelo calculada promediando los resultados obtenidos sobre todas las imágenes del conjunto de prueba para cada clase por separado. Ambos modelos muestran un rendimiento sólido (por sobre el 50%), pero se puede observar claramente una mejora en ML-AFM en comparación con el modelo base, ya que ofrece una media más alta en cada una de las clases.

La Tabla 6 presenta el rendimiento del reconocimiento de ingredientes, agrupado por cada clase de alimento, en términos de F1 para el modelo ResNet50. Se puede observar que algunas clases, como “edamame” con un F1 del 98,0%, “macarons” con un 97,0% y “pho” con un 96,0%, tienen un desempeño sobresaliente. Estos resultados indican que ResNet50 es capaz de identificar y clasificar correctamente los ingredientes en las imágenes pertenecientes a esas clases que tienden a tener una menor variabilidad en cuanto a la información visual que las representa. Por otro lado, la Tabla 7 revela las clases con un rendimiento más bajo para el modelo ResNet50. Entre estas clases se encuentran “huevos_rancheros” con un 63,0% y “foie_gras” con un 41,0%.

Por otra parte, la Tabla 8 muestra las clases con el mejor rendimiento para el modelo ML-AFM. Destacan “edamame” y “salt, edam” con F1 perfectos del 100,0%. Además, “takoyaki” obtiene un alto F1 del 97,8%, seguido de “macarons” con un 97,2% y

Tabla 4. Ingredientes pertenecientes a los peores 6 porcentajes de Recall obtenidos para cada modelo.

Ingredientes	**ResNet50**	**Ingredientes**	**ML-AFM**
ham	44,4%	brown sugar	64,9%
anaheim chile	42,0%	dill	64,4%
pasta	41,6%	ham	61,2%
adobo	41,2%	sirloin	54,4%
sirloin	19,2%	filet	22,0%
filet, foie gras	0,0%	foie gras	16,0%

Tabla 5. Resultados obtenidos sobre los datos del conjunto de entrenamiento validados. como ruidosos y limpios.

Modelo	**F1 (Noisy)**	**F1 (Clean)**	**Memorización (Noisy)**	**Memorización (Clean)**
ResNet50	30,99%	84,42%	7,64%	63,63%
ML-AFM	36,94%	88,91%	13,47%	71,48%

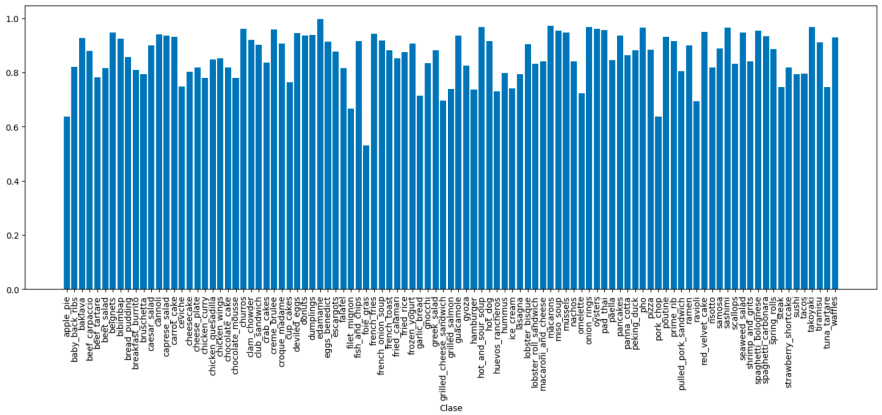

Figura 7. F1 media por cada clase para ML-AFM.

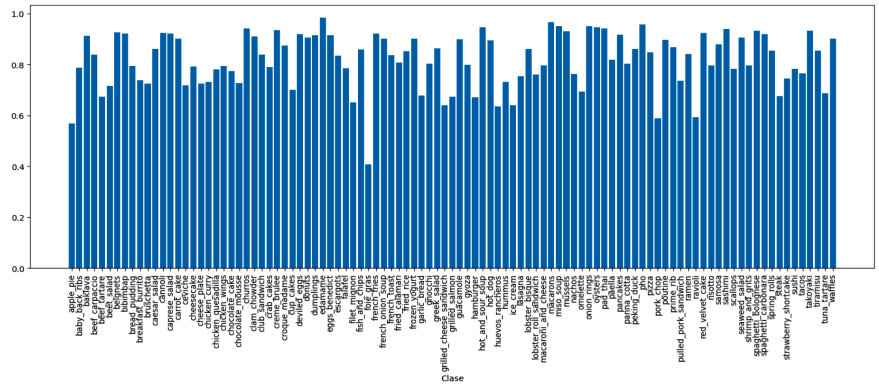

Figura 8. F1 media por cada clase para ResNet50.

Tabla 6. Clases que contienen las imágenes con mejor F1 para ResNet50.

Clase	F1	Ingredientes
edamame	98,0%	salt, edam
macarons	97,0%	sugar, almond, egg
pho	96,0%	marrow, oxtail, beef, clove, anise, cinnamon, cardamom, black pepper, coriander, fennel,gin, onion, shallot, fish, sugar, noodles, scallions, eye round, cilantro, basil, beans, lime, bird chile, chile, hoisin
onion_rings	95,0%	onion, flour, baking, egg, bread, oil
miso_soup	94,0%	noodles, miso, tofu, cress, onion, cilantro, pepper, miso, apple, gin, turmeric, water

Tabla 7. Clases que contienen las imágenes con peor F1 para ResNet50.

Clase	F1	Ingredientes
huevos_rancheros	63,0%	fat,steak, gin, shallot, parsley, capers, worcestershire, egg, black pepper, crostini
ravioli	59,0%	beef, cheese, mayonnaise, ketchup, mustard, dill, black pepper, onion, tomato, lettuce
pork_chop	59,0%	filet, oil, garlic, gin, salt, pepper
apple_pie	57,0%	bread, cheddar, tomato, onion, oil
foie_gras	41,0%	cocoa, brown sugar, milk, vanilla

Tabla 8. Clases que contienen las imágenes con mejor F1 para ML-AFM.

Clase	F1	Ingredientes
edamame	100,0%	salt, edam
takoyaki,	97,8%	flour, egg, cold water, salt, konbu dashi, dashi, soy, octopus, onion, pie, cheese
macarons	97,2%	sugar, almond, egg
hot_and_sour_soup	96,8%	broth, soy, pepper, shiitake, rice, corn starch, egg, tofu, gin, scallions
onion_rings	96,7%	onion, flour, baking, egg, bread, oil

"hot_and_sour_soup" con un 96,8%. Estos resultados indican que ML-AFM muestra una habilidad destacada para identificar y clasificar correctamente los ingredientes de las imágenes pertenecientes a estas clases. Sin embargo, en la Tabla 8 se presentan las clases con el peor rendimiento para ML-AFM. Por ejemplo, "ravioli" tiene un F1 del 69,0% y "foie_gras" registra un F1 de un 53,0%.

Ambos métodos muestran dificultades para clasificar con precisión los ingredientes en determinadas clases, como demuestran las bajas puntuaciones F1 de la Tabla 7 y la Tabla 9. Además del ruido en las anotaciones, los métodos se enfrentan a retos propios del análisis de alimentos. En el caso de determinados platos, existe una gran diversidad en la representación de estos. A veces resulta difícil distinguir dónde se encuentra el ingrediente principal dentro de la imagen. Esto se debe a que algunos platos presentan una combinación de diferentes ingredientes, colores y texturas, lo que dificulta su clasificación precisa. Además, hay platos que se sirven con varios ingredientes adicionales, lo que complica aún más la tarea de identificación. Otro aspecto para tener en cuenta son los platos que pueden presentar distintas preparaciones. Cada variante puede tener características visuales diferentes, lo que aumenta la complejidad del reconocimiento. La presencia de estos elementos adicionales dificulta el aprendizaje del modelo y, por tanto, complica la correcta asignación de etiquetas a cada ingrediente del plato. Sin embargo, es importante destacar que la solución propuesta, continúa demostrando una mejora notable en comparación con la línea base, ResNet50.

La Figura 9 y Figura 10 muestran los casos de éxito y fallo centrados en las clases que obtuvieron un rendimiento más bajo. Ambas figuras muestran que tanto el modelo ML-AFM como el modelo ResNet50 tienen casos de fallo en la clasificación de ingredientes, pero también tienen casos de éxito donde logran clasificar con precisión los ingredientes esperados. Es importante destacar que no todas las imágenes fueron clasificadas incorrectamente en las clases donde el modelo tuvo dificultades. Estas dificultades se deben a las características de las imágenes, como la textura, los ingredientes agregados, la preparación, entre otros factores. Por lo tanto, en ocasiones, el modelo puede clasificar correctamente ingredientes generales como "*oil*", "*sugar*" y "*salt*", pero puede fallar en la identificación de los ingredientes principales. Esto implica que los platos que contienen un ingrediente principal junto con varios ingredientes secundarios pueden tener un rendimiento generalmente alto, pero la clasificación no es la más adecuada debido a que no se identifica el ingrediente principal. Estos resultados destacan la importancia de evaluar y mejorar continuamente los modelos de clasificación para obtener resultados más precisos y confiables.

Tabla 9. Clases que contienen las imágenes con peor F1 para, ML-AFM.

Clase	F1	Ingredientes
ravioli	69,0%	cheese, pasta, cheese, cheese
filet_mignon	67,0%	filet, oil, garlic, gin, salt, pepper
apple_pie	64,0%	butter, flour, sugar, brown sugar, apple, cinnamon, nut
pork_chop	64,0%	pork, orange, lime, black pepper, cumin, cayenne pepper, garlic, oregano, oil, onion, garlic, white wine
foie_gras	53,0%	foie gras, salt, milk

Figura 9. Casos de éxito y fallo de ML-AFM. GT representa la salida esperada, PRED la salida predicha y F1 el resultado en la clasificación para la imagen. En verde, rojo y naranja se resaltan los ingredientes, encontrados, erróneos y no identificados, respectivamente.

Figura 10. Casos de éxito y fallo de ResNet50. GT representa la salida esperada, PRED la salida predicha y F1 el resultado en la clasificación para la imagen. En verde, rojo y naranja se resaltan los ingredientes, encontrados, erróneos y no identificados, respectivamente.

Adicionalmente, en algunas ocasiones el modelo puede reconocer ingredientes que son comunes en las preparaciones, pero no identificar correctamente el ingrediente principal. Por ejemplo, en el caso de "foie_gras", el modelo puede tener un rendimiento general alto, pero no lograr reconocer el ingrediente principal que es el foie gras en sí. Esto indica que, en las clases con peor rendimiento, es posible observar imágenes clasificadas correctamente en general, pero con la falta de reconocimiento de los ingredientes principales.

CONCLUSIONES

En este trabajo, se diseñó y evaluó una variante modificada del modelo AFM nombrada ML-AFM para abordar el problema de la clasificación de alimentos múltiples con etiquetas ruidosas. Se comparó con el modelo ResNet50 (línea base), el cual es comúnmente usado en el campo de la clasificación de alimentos y además es parte del *backbone* del modelo propuesto. La evaluación y comparativa se realizó utilizando la base de datos Food-101N, que consiste en imágenes de alimentos con etiquetas que contienen un 20% de ruido. Se ampliaron las anotaciones de esta base de datos para que sea aplicable a un enfoque de múltiples etiquetas. Los resultados obtenidos demostraron que el modelo propuesto ML-AFM superó el rendimiento del modelo de línea base en la métrica de F1, AUPRC e índice de Jaccard, logrando un 76,56%, 84,93%, 77,19% respectivamente. Además, al analizar los resultados a nivel de ingredientes mejores y peores identificados, se evidenció la superioridad del modelo ML-AFM en comparación con el modelo de línea base. Se observó un aumento en el valor de F1 para todos los ingredientes, incluso en aquellos que ResNet50 no pudo detectar en las imágenes. Esto permitió que todas las imágenes pertenecientes a cada una de las clases fueran clasificadas en promedio con un F1 superior al 60%, con la excepción de la clase "foie gras". Como trabajo futuro, se pretende explorar en técnicas que permitan reducir el nivel de memorización cuando los modelos son entrenados con datos ruidosos, de este modo es posible ampliar la generalización y precisión en la clasificación multi-etiqueta de comida.

AGRADECIMIENTOS

Este trabajo es parcialmente financiado por el gobierno de Chile a través de su agencia nacional de investigación y desarrollo, ANID (No. Fondecyt iniciación 11230262).

REFERENCIAS

[1] E. Di Angelantonio *et al.*, "Body-mass index and all-cause mortality: individual-participant-data meta-analysis of 239 prospective studies in four continents," *Lancet*, vol. 388, no. 10046, pp. 776-786, 2016, doi: 10.1016/S0140-6736(16)30175-1.

[2] A. Afshin *et al.*, "Health effects of dietary risks in 195 countries, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017," *Lancet*, vol. 393, no. 10184, pp. 1958-1972, 2019, doi: 10.1016/S0140-6736(19)30041-8.

[3] E.K. Amine *et al.*, "Diet, nutrition, and the prevention of chronic diseases," World Health Organization (Who), Geneva, Switzerland. *Rep.* 916, pp. 11-12, 2002, https://apps.who.int/iris/bitstream/handle/10665/42665/WHO_TRS_916.pdf?sequence=1

[4] A.F. Subar *et al.*, "Addressing current criticism regarding the value of self-report dietary data," *The Journal of Nutrition*, vol. 145, no. 12, pp. 2639-2645, 2015, doi: 10.3945/jn.115.219634.

[5] C.J. Boushey, M. Spoden, F.M. Zhu, E.J. Delp, and D.A. Kerr, "New mobile methods for dietary assessment: Review of image-assisted and image-based dietary assessment methods," *Proceedings of the Nutrition Society*, vol. 76, no. 3, 2017, pp. 283-294, doi: 10.1017/S0029665116002913.

[6] J.E. Cade, "Measuring diet in the 21st century: Use of new technologies," *Proceedings of the Nutrition Society*, vol. 76, no. 3, pp. 276-282, 2017, doi: 10.1017/S0029665116002883.

[7] D. Pandey *et al.*, "Object Detection in Indian Food Platters using Transfer Learning with YOLOv4," *2022 IEEE 38th International conference on data engineering workshops*, 2022, pp. 101-106, doi: 10.1109/ICDEW55742.2022.00021.

[8] F.V. Fernández, "Diseño y entrenamiento de una red neuronal para reconocimiento de imágenes", Tesis de Grado, Ingeniería Electrónica, Robótica y Mecatrónica, Universidad de Sevilla, Sevilla, España, 2022. [En línea]. Disponible: https://idus.us.es/handle/11441/143564.

[9] S. Adinugroho, P.P. Adikara, E. Santoso, R. Amara, K. Septiana, and K.D. Anggita, "Indonesian food identification and detection in the smart nutrition box using faster-RCNN," Proceedings of the 5th International Conference on Sustainable Information Engineering and Technology, 2020, pp. 113-117, doi: 10.1145/3427423.3427429.

[10] L. Deng, J. Chen, and S.T. Chong-Wah, and Q. Sun, "Mixed dish recognition with contextual relation and domain alignment," *IEEE Transactions on Multimedia*, vol. 24, pp. 2034-2045, 2021, doi: 10.1109/TMM.2021.3075037.

[11] Y. Zhu, X. Zhao, C. Zhao, J. Wang, and H. Lu, "Food det: Detecting foods in refrigerator with supervised transformer network," *Neurocomputing*, vol. 379, pp. 162-171, 2020, doi: 10.1016/j.neucom.2019.10.106.

[12] E. Aguilar, B. Remeseiro, M. Bolaños, and P. Radeva, "Grab, Pay, and Eat: Semantic Food Detection for Smart Restaurants," *IEEE Transactions on Multimedia*, vol. 20, no. 12, pp. 3266-3275, 2018, doi: 10.1109/TMM.2018.2831627.

[13] J. Li, J. Xiong, and Z. Chen, "Food-Agnostic Dish Detection: A Simple Baseline," *IEEE Access*, vol. 9, pp. 125375-125383, 2021, doi: 10.1109/ACCESS.2021.3108184.

[14] K. Lee and L. Y. He, Xiaodong, Lei Zhang, "Cleannet: Transfer learning for scalable image classifier training with label noise," *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5447-5456, doi: 10.1109/CVPR.2018.00571.

[15] C. Tan, J. Xia, L. Wu, and S.Z. Li, "Co-learning: Learning from Noisy Labels with Self-supervision," in *MM 2021 Proceedings of the 29th ACM International Conference on Multimedia*, Association for Computing Machiner, Oct. 2021, pp. 1405-1413, doi: 10.1145/3474085.3475622.

[16] K. Sharma, P. Donmez, E. Luo, Y. Liu, and I.Z. Yalniz, "NoiseRank: Unsupervised Label Noise Reduction with Dependence Models," *European Conference on Computer Vision*, 2020, pp. 737-753, doi: 10.1007/978-3-030-58583-9_44.

[17] Q. Li, X. Peng, L. Cao, W. Du, H. Xing, and Y. Qiao, "Product Image Recognition with Guidance Learning and Noisy Supervision,"*Computer Vision and Image Understanding,* vol. 196, 2020, doi: 10.1016/j.cviu.2020.102963.

[18] J. Han, P. Luo, and X. Wang, "Deep Self-Learning from Noisy Labels," *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5138-514, doi: 10.1109/ICCV.2019.00524.

[19] G. Algan and I. Ulusoy, "Image Classification with Deep Learning in the Presence of Noisy Labels: A Survey," *Knowledge-Based Systems,* vol. 215, doi: 10.1016/j.knosys.2021.106771.

[20] Y. Yao *et al.*, "Jo-SRC: A Contrastive Approach for Combating Noisy Labels," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5192-5201, doi: 10.1109/CVPR46437.2021.00515.

[21] H. Sun, C. Guo, Q. Wei, Z. Han, and Y. Yin, "Learning to Rectify for Robust Learning with Noisy Labels," 2022, arXiv:2111.04239.

[22] F.R. Cordeiro, R. Sachdeva, V. Belagiannis, I. Reid, and G. Carneiro, "LongReMix: Robust learning with high confidence samples in a noisy label environment," *Pattern Recognition*, vol. 133, 2023, doi: 10.1016/j.patcog.2022.109013.

[23] G. Algan and I. Ulusoy, "Meta Soft Label Generation for Noisy Labels," *2020 25th International Conference on Pattern Recognition (ICPR)*, 2021, pp. 7142-7148, doi: 10.1109/ICPR48806.2021.9412490.

[24] W. Zhang, Y. Wang, and Y. Qiao, "MetaCleaner: Learning to Hallucinate Clean Representations for Noisy-Labeled Visual Recognition," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7373-7382, doi: 10.1109/CVPR.2019.00755.

[25] M.K. Xie and S.J. Huang, "CCMN: A General Framework for Learning With Class-Conditional Multi-Label Noise," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 154-166, 2023, doi: 10.1109/TPAMI.2022.3141240.

[26] W. Zhao and C. Gomes, "Evaluating Multi-label Classifiers with Noisy Labels," 2021, *arXiv: 2102.08427.*

[27] M. Bolaños, A. Ferrà, and P. Radeva, "Food ingredients recognition through

multi-label learning," *New Trends in Image Analysis and Processing–ICIAP 2017: ICIAP International Workshops*, 2017, pp. 394-402, doi: 10.1007/978-3-319-70742-6_37.

[28] N. Inoue, E. Simo-Serra, T. Yamasaki, and H. Ishikawa, "Multi-label Fashion Image Classification with Minimal Human Supervision," *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 2261-2267, doi: 10.1109/ICCVW.2017.265.

[29] M. Hu, H. Han, S. Shan, and X. Chen, "*Multi-label Learning from Noisy Labels with Non-linear Feature Transformation," Computer Vision – ACCV 2018. Lecture Notes in Computer Science.* vol. 11365, C.V. Jawahar, H. Li, G. Mori, and K. Schindler, Eds. 2019, doi: 10.1007/978-3-030-20873-8_26.

[30] T. Burgert, M. Ravanbakhsh, and B. Demir, "On the Effects of Different Types of Label Noise in Multi-Label Remote Sensing Image Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, 2022, doi: 10.1109/TGRS.2022.3226371.

[31] A. Ghiassi, R. Birke, and L. Y. Chen, "Trusted Loss Correction for Noisy Multi-Label Learning," *Asian Conference on Machine Learning*, 2023, pp. 343-358. [Online]. Available: https://proceedings.mlr.press/v189/ghiassi23b.html

[32] Z. Sun *et al.*, "PNP: Robust Learning from Noisy Labels by Probabilistic Noise Prediction," IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022, pp. 5301-5310, doi: 10.1109/CVPR52688.2022.00524.

[33] Y. Zhang, S. Zheng, P. Wu, M. Goswami, and C. Chen, "Learning with Feature-Dependent Label Noise: A Progressive Approach," 2021, arXiv: 2103.07756.

[34] I. Papathanail, Y. Lu, A. Ghosh, and S. Mougiakakou, "Food Recognition in the Presence of Label Noise," *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, Part V, Jan. 10-15, 2021,* pp. 617-328, doi: 10.1007/978-3-030-68821-9.

[35] X. Liang, L. Yao, X. Liu, and Y. Zhou, "Tripartite: Tackle Noisy Labels by a More Precise Partition,"2022, *arXiv: 2202.09579.*

[36] H. Song, M. Kim, D. Park, Y. Shin, and J.G. Lee, "Robust Learning by Self-Transition for Handling Noisy Labels," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery*, Aug. 2021, pp. 1490-1500. doi: 10.1145/3447548.3467222.

[37] X. Peng, K. Wang, Z. Zeng, Q. Li, J. Yang, and Y. Qiao, "Suppressing Mislabeled Data via Grouping and Self-Attention," *Computer Vision–ECCV. Lecture Notes in Computer Science,* vol. 12361, A. Vedaldi, H. Bischof, T. Brox and J.M. Frahm, Eds. Aug. 23-28, 2020, pp. 786-802, doi: 10.1007/978-3-030-58517-4_46.

[38] H. Zhang, M. Cisse, Y.N. Dauphin, and D. Lopez-Paz, "MixUp: Beyond empirical risk minimization," in *6th International Conference on Learning Representations*, 2018.

[39] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101 - Mining discriminative components with random forests," in *Computer Vision–ECCV 2014. Lecture Notes in Computer Science, vol. 8694, D. Fleet, T. Pajdla, B. Schiele and T. Tuytelaars, Eds. Sept. 6-12, 2014*, pp. 446-461, doi: 10.1007/978-3-40319-10599-4_29.