# Exploración de las Redes Neuronales Artificiales para la detección de Salmonella *spp*. en aves de corral

Melissa Ortega Alzate

Monografía presentada para optar al título de Especialista en Analítica y Ciencia de Datos

Asesor
Nombres completos, Título académico más alto

Universidad de Antioquia
Facultad de Ingeniería
Especialización en Analítica y Ciencia de Datos
Medellín, Antioquia, Colombia
2023

| | |
|---|---|
| **Cita** | (Ortega Alzate, 2024) |
| **Referencia**<br><br>**Estilo APA 7 (2020)** | Ortega Alzate, M. (2024). Exploración de las R*edes Neuronales Artificiales para la detección de Salmonella spp. en aves de corral* [Trabajo de grado especialización]. Universidad de Antioquia, Medellín, Colombia. |



Especialización en Analítica y Ciencia de Datos, Cohorte VI.

Centro de Investigación Ambientales y de Ingeniería (CIA).

**Dedicatoria**

Texto de dedicatoria centrado.

**Agradecimientos**

Mi gratitud para Iluma Alliance, incubadora de sueños.

A Okuo, fuente inagotable de aprendizaje e inspiración.

## Tabla de contenido

## Lista de tablas

## Lista de figuras

## Siglas, acrónimos y abreviaturas

| | |
|---|---|
| **APA** | American Psychological Association |
| **Esp.** | Especialista |
| **IA** | Inteligencia Artificial |
| **ICA** | Instituto Colombiano Agropecuario |
| **INVIMA** | Instituto Nacional de Vigilancia de Medicamentos y Alimentos |
| **ML** | Machine Learning (Aprendizaje de máquinas) |
| **OMS** | Organización Mundial de la Salud |
| **RNA** | Red Neuronal Artificial |
| **RNC** | Red Neuronal Convolucional |
| **ROC** | Receiver Operating Characteristic (Característica Operativa del Receptor) |
| **UdeA** | Universidad de Antioquia |

**Resumen**

La presencia de cepas patógenas del microorganismo Salmonella *spp*. en alimentos, es uno de los contaminantes microbiológicos más comunes que afectan la salud de los consumidores de proteína de origen animal. Por lo tanto, el seguimiento microbiológico durante toda la cadena de producción es una prioridad para las compañías, pues se trata no sólo de garantizar alimentos salubres para la población sino prevenir pérdidas económicas debido a la muerte de las aves por Salmonelosis. Por lo anterior, la detección precisa y temprana de aves de corral enfermas en las granjas es un reto vigente para estas organizaciones.

En esta monografía se desarrolla un método alternativo basado en algoritmos de Inteligencia Artificial para la detección de salmonella *spp*. en imágenes. Utilizando un conjunto de datos de Kaggle con imágenes de heces de aves de corral tomadas en África entre 2020 y 2021, se realiza una clasificación binaria entre "Healthy" y "Salmonella". Tras una limpieza de datos, se preprocesan las imágenes y se dividen en trenes de entrenamiento y prueba. Se propone el uso de tres arquitecturas de Redes Neuronales Convolucionales (ResNet, Inception y GoogLeNet) y evaluar su desempeño con métricas como precisión, sensibilidad y F1-Score.

El código desarrollado durante el desarrollo de este proyecto se encuentra disponible en el repositorio de GitHub asociado (https://bit.ly/49EMvhb)

*Palabras clave*: Inteligencia Artificial, Seguridad alimentaria, Salmonella *spp.*, avicultura, Red Neuronal Convolucional, clasificación de imágenes, mortalidad en aves.

**Abstract**

The presence of pathogenic strains of the microorganism Salmonella *spp*. in food is a common microbiological contaminant, significantly impacting the health of consumers of animal protein. Therefore, microbiological monitoring throughout the production chain is a priority for companies, aiming not only to ensure food safety for the population but also to prevent economic losses due to poultry death from Salmonellosis. Consequently, the accurate and early detection of sick poultry on farms remains a current challenge for these organizations.

This project introduces an alternative method based on Artificial Intelligence algorithms for Salmonella *spp*. detection in images. Using a Kaggle dataset comprising images of poultry feces taken in Africa between 2020 and 2021, a binary classification between "Healthy" and "Salmonella" is performed. Following data cleaning, images are preprocessed and split into training and testing sets. The study proposes the use of three Convolutional Neural Network architectures (ResNet, Inception, and GoogLeNet), evaluating their performance using metrics such as accuracy, sensitivity, and F1-Score.

The code developed for this project is available in the following GitHub repository (https://bit.ly/49EMvhb).

*Keywords*: Artificial Intelligence, Food Safety, Salmonella *spp.*, Poultry farming, Convolutional Neural Network, image classification, Poultry mortality.

# 1. Descripción del problema

Según la Organización Mundial de la Salud (OMS), cada año se enferman alrededor de 600 millones de personas en el mundo por consumir alimentos contaminados y 420.000 mueren a raíz de enfermedades transmitidas por estos alimentos (OMS, 2020). Por lo anterior, las compañías involucradas en la producción de proteína de origen animal adquieren una responsabilidad social con los consumidores, y están en la obligación de implementar un plan de gestión de calidad en función de garantizar la inocuidad alimentaria. Normalmente, estos planes incluyen un programa de muestreo microbiológico, en donde se describen los procedimientos para la detección de microorganismos patógenos durante el proceso productivo (ICA, 2019).

Sin embargo, las metodologías implementadas son aquellas conocidas como microbiología tradicional, en donde una muestra es enviada a un laboratorio para su diagnóstico. Este procedimiento demanda tiempo y recursos a las compañías, que, además, limitan su desempeño microbiológico.

Por lo anterior, explorar algoritmos de *Machine Learning* para la detección de microorganismos patógenos a partir del reconocimiento de patrones en imágenes desde una etapa temprana de la producción, es un factor determinante para reducir la carga microbiológica, mejorar la productividad de las compañías y garantizar el consumo de alimentos inocuos desde el principio de la cadena de producción.

## 1.1.Problema de negocio

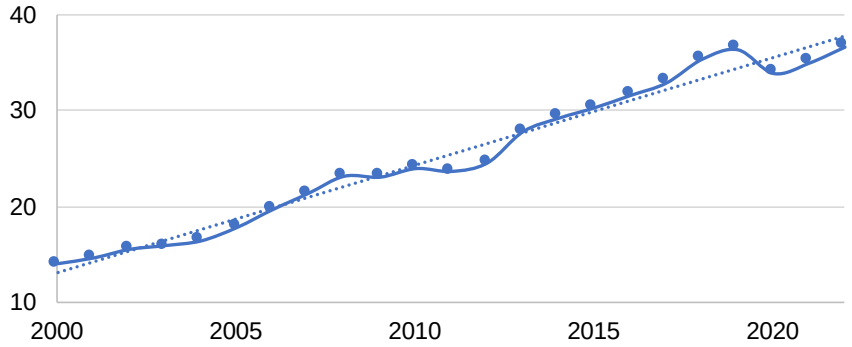
La inocuidad alimentaria es una prioridad en la agenda de organizaciones públicas y privadas alrededor del mundo, así como para las diferentes compañías involucradas en la cadena de producción de proteína animal. Un contaminante común es la presencia de especies enterohemorrágicas de Salmonella en alimentos, uno de los microorganismos que más afectan la salud de quienes consumen proteína de origen animal (OMS, 2020). El Grupo de Epidemiología Veterinaria del Instituto Colombiano Agropecuario, ICA estableció que “la Salmonelosis aviar es una enfermedad altamente contagiosa para aves de cualquier edad presentando una mortalidad muy

variable entre 4 y 50 % de la parvada" (Subgerencia de Protección y Regulación Pecuaria et al., 2007).

Precisamente, la producción per cápita de pollo durante el año 2022 en Colombia, fue de 32,07 kg; posicionándose ese año como el 14º país con mayor producción per cápita de pollo en el mundo (Federación Nacional de Avicultores de Colombia, 2023).

**Figura 1**

*Consumo per cápita de pollo (kilogramos por habitante) en Colombia.*




*Nota.* Fuente http://bit.ly/3QT15cC (Federación Nacional de Avicultores de Colombia, 2023). Elaboración propia en Microsoft Office Excel

El crecimiento constante en el consumo per cápita de pollo (**Figura 1**) ha traído oportunidades competitivas para la industria avícola del país (AviNews, 2023). Sin embargo, también representa grandes retos en términos de calidad, pues los productores asumen una mayor responsabilidad respecto a la salubridad, y están en la obligación de implementar planes de gestión de calidad para garantizar la inocuidad alimentaria en sus procesos y producir alimentos salubres para el consumidor final.

Normalmente, los planes de gestión que estas compañías llevan a cabo, incluyen un programa microbiológico, en donde se describen los procedimientos para la detección de microorganismos patógenos durante el proceso productivo. Específicamente para las granjas, el ICA estableció en 2019 el "Programa Nacional de Control y Disminución de Prevalencia[1] de las Salmonellas Paratíficas […] en aves de corral dentro del territorio nacional". Sin embargo, el problema radica en que la metodología implementada en granjas es aquella conocida como

[1] Prevalencia: proporción de casos de una enfermedad en un periodo de tiempo, respecto a la población existente en la zona de objeto de estudio (RAE, 2022).

microbiología tradicional, que consiste en tomar al menos 20 muestras cloacales de aves diferentes cada mes y enviarlas a un laboratorio de microbiología certificado por el ICA para su posterior tratamiento y diagnóstico (ICA, 2019).

En general, estas metodologías tradicionales han sido problemáticas porque implican tiempo, recursos humanos y costos a las compañías, pues deben contar con personal capacitado para la toma de muestras, así como con logística de almacenamiento y transporte para llevar las muestras refrigeradas hasta un laboratorio. Adicionalmente, tomar muestras cloacales a las aves puede causar estrés a los animales, lo que provoca pérdida de peso y consecuentemente afecta la productividad de las granjas. Se debe tener en cuenta, además, el tiempo que tarda el procesamiento y diagnóstico de las muestras (hasta 7 días calendario contados a partir de la fecha de recepción en el laboratorio) (Ma et al., 2022). Estos recursos, aumentan de manera considerable con el volumen de muestras tomadas, y su escalabilidad es cuestionable.

Por otro lado, es esencial considerar que cada momento es susceptible a error humano: una omisión en la toma de muestras, un formato de remisión mal diligenciado, contaminación cruzada o demoras durante el transporte (más de 24 h), almacenamiento con refrigeración deficiente o errores en la implementación de los métodos en el laboratorio. Todo lo anterior puede afectar la confiabilidad de los resultados obtenidos.

De esta manera, el panorama actual es un procedimiento para la detección de Salmonella *spp.* en granjas lento y de difícil seguimiento, que podría entregar resultados poco representativos del desempeño microbiológico real de las granjas.

Este problema se evidencia en las granjas, cuando los productores de aves de corral (las granjas) reducen su productividad debido a aves muertas por Salmonelosis. Adicionalmente, cuando no cumplen con los estándares exigidos por sus clientes y no pueden vender las aves, disminuyendo su productividad y afectando su reputación. El problema se extiende a las empresas posteriores en la cadena de producción, pues deben aumentar su inversión en puntos críticos de control, desinfectantes, inhibidores, etc. para controlar la carga microbiológica en sus instalaciones.

Finalmente, en esta problemática repercute en personas enfermas o muertes causadas por la ingesta de alimentos contaminados (OMS, 2020).

Por lo tanto, es necesario desarrollar métodos alternativos para la detección de estos microorganismos en los animales de granja, que sean confiables y rápidos (Obe et al., 2023). Por ejemplo, mediante la exploración de algoritmos de *Machine Learning* (ML) para la detección de Salmonella *spp.* a partir del reconocimiento de patrones en imágenes de heces de aves de corral. Este reconocimiento es un factor determinante para mejorar la productividad de las compañías, reducir la carga microbiológica en la cadena de producción de proteína animal, y garantizar el consumo de alimentos inocuos desde el principio de la cadena de producción.

### 1.2. Aproximación desde la analítica de datos

En el marco del control de calidad como área de la industria de la Ciencia de Datos, se propone programar un algoritmo de *ML* que serviría para identificar patrones a partir de imágenes para detectar la presencia de Salmonella *spp*. en heces de aves de corral. Este método alternativo a la microbiología tradicional, podría disminuir los recursos invertidos en garantizar la inocuidad, pues habrá una reducción no sólo en el tiempo (desde 7 a 1 día hábil para conocer los resultados) y personal asociado a las metodologías actualmente usadas, sino también en las acciones correctivas que podrían generar sobrecostos a las empresas que compran las aves.

Las granjas, aumentarían su desempeño microbiológico, pues el modelo predictivo propuesto permitiría reducir la logística de toma, almacenamiento, transporte y diagnóstico de muestras microbiológicas, podría aumentar el número de muestras procesadas en un periodo de tiempo. Lo anterior, generaría conclusiones basadas en una mayor cantidad de información. Adicionalmente, habrá un impacto positivo en la reputación y competitividad de las empresas. Ser una granja que utiliza Inteligencia Artificial en sus procesos para responder en tiempo real a las necesidades del negocio, convierte a las compañías en proveedores confiables (Shaik Mazhar & Akila, 2022).

En una eventual implementación de los modelos predictivos desarrollados, en donde el algoritmo resulte en una alta cantidad de falsos negativos, es decir, en la no detección de Salmonella s*pp.,* el riesgo microbiológico sería subestimado lo que podría repercutir en la productividad de una granja y su reputación. Por el contrario, en el caso de falsos positivos, un negocio puede verse afectado si tomara decisiones como implementar un tratamiento médico en una granja cuando no sea necesario. En estos casos, debería considerarse un protocolo de seguimiento, que incluya una etapa preliminar en donde se corroboren los resultados obtenidos con el modelo predictivo por otro método de microbiología tradicional y asegurar la capacidad de generalización del modelo.

Además, la solución analítica que se propone:

- No incluye recomendaciones de cuantos datos son necesarios y/o representativos para la correcta detección de microorganismos patógenos en granjas.
- No incluye sugerencias sanitarias para el manejo de granjas que resulten identificadas como contaminadas con microorganismos patógenos.
- No incluye análisis de causas de posibles contaminaciones.
- No incluye el despliegue del algoritmo entrenado y ajustado en ninguna herramienta como celulares o tabletas para ser implementado en aplicaciones Web.

La solución analítica implicaría los siguientes riesgos:

Si el algoritmo entrenado resulta en falsos negativos, es decir, en la NO detección de microorganismos patógenos. El riesgo microbiológico sería subestimado lo que podría repercutir en la productividad de una granja y su reputación. Por el contrario, en el caso de falsos positivos, un negocio puede verse afectado si tomara decisiones como implementar un tratamiento médico en la una granja cuando no sea necesario.

Además, respecto a riesgos para desarrollar el proyecto de aprendizaje, se encuentra que:

- El data set disponible Kaggle descrito anteriormente, no sea suficiente para el entrenamiento del algoritmo.

- Los recursos computacionales disponibles (computador personal) no sean suficientes para ejecutar los algoritmos a programar.
- Otros estudios se han referido a la selección y complejidad de los modelos como un desafío significativo en soluciones propuestas anteriormente (Hossain et al., 2022).

### 1.3. Origen de los datos

El conjunto de datos fue extraído de la página Web de Kaggle, se titula "Chicken Disease Image Classification" y contiene un archivo .csv con las etiquetas de cada imagen y 8067 fotografías de heces de aves de corral. Las fotografías fueron tomadas originalmente en granjas de aves en Tanzania (África) entre septiembre 2020 y febrero 2021, utilizando la aplicación Open Data Kit[2] en dispositivos móviles (Kaggle, 2022). No se modificaría, combinaría o utilizaría el data set con propósitos diferentes a los enmarcados en este proyecto.

El data set está disponible para su uso abierto y pueden ser descargado desde la página Web de Kaggle (Kaggle, 2022). No hay ninguna restricción o condición de uso establecida para el conjunto de datos. Las imágenes pueden ser vistas con el visualizador de imágenes predeterminado y el .csv mediante Python en un Jupyter Notebook, estas herramientas permiten explorar el conjunto de datos de manera preliminar.

### 1.4. Métricas de desempeño

La métrica más utilizada en la literatura para evaluar el desempeño de modelos de *ML* es la exactitud (**Accuracy**) (Ma et al., 2022) que indica cuántas predicciones acertadas se realizan respecto al total de predicciones, se espera al menos una precisión alrededor del +-10 % comparada con la precisión del método realizado en los laboratorios de microbiología tradicional. La RNA propuesta también será evaluada utilizando otras métricas de desempeño, como el **Recall,** el cual da cuenta de la sensibilidad del modelo, indicando el número de imágenes clasificadas correctamente en relación con el número de imágenes pertenecientes a esta clase. El Recall es

[2] https://getodk.org/

especialmente importante, pues se espera detectar la mayoría de las aves contaminadas (al menos el 90 %) con Salmonella *spp.* para tomar acciones correctivas en las granjas.

Los tratamientos médicos se aplican normalmente en el agua de bebida del galpón, es decir, se realiza al galpón en su conjunto. En este sentido, alguna tolerancia a falsos positivos puede considerase. Sin embargo, se tendrá en cuenta el **F1-score** para encontrar un equilibrio adecuado entre precisión y Recall. Se utilizará, además, una **curva ROC** para comparar el desempeño de las diferentes arquitecturas de RNA desarrolladas entre ellas y respecto a un predictor aleatorio. Esta curva representa la tasa de verdaderos positivos (1 – especificidad) para diferentes umbrales de clasificación. Así, el área bajo la curva más cercana a 1 indicará el modelo con mejor rendimiento respecto a discriminación entre las dos clases (Hajian-Tilaki, 2013).

Se espera que el modelo predictivo tenga influencia sobre la **mortalidad** de cada galpón, este es un indicador clave del negocio y está directamente relacionado con productividad de las granjas. La mortalidad expresada en porcentaje da cuenta del número de aves que mueren respecto al total de aves iniciales en el galpón, puede expresarse por día o acumulado por galpón desde el comienzo de la cría. Cada día los galpones pierden aves, es decir, el lote reduce su tamaño y, no sólo se dejan de percibir ganancias por cada ave no vendida, sino que se pierde la inversión realizada en agua, alimento y logística durante el tiempo de vida de las aves que no podrán ser vendidas a una planta de beneficio. En este sentido, el modelo influiría en la mortalidad, pues se espera que se reduzca el tiempo de respuesta a eventos de Salmonelosis, aumente el número de muestras disponibles para tomar decisiones respecto a la salud de las aves y así, mejorar la precisión en la detección de aves enfermas.

Durante el 2023, 67.109 aves han sido afectadas, de las cuáles el 60,9 % corresponden a aves de engorde (Ministerio de Agricultura, 2023). Considerando un peso promedio del pollo entero fresco sin vísceras de 2 kg para cada ave y un precio de 10000 COP/kg (Boletín FENAVIQUIN, nov. 2023) se dejaron de comercializar 134.218 kilogramos de carne de pollo que corresponden a 1.342.180.000 COP (para diferentes causas de mortalidad).

Según la Subgerencia de Protección y Regulación Pecuaria, en el 2006, las pérdidas ocasionadas por mortalidad en las aves causada específicamente por Salmonelosis fueron de 18.452.195 COP (Subgerencia de Protección y Regulación Pecuaria et al., 2007).

En general, la mortalidad de aves de corral en Colombia por Tifosis aviar (enfermedad causada por Salmonella *Gallinarum*) es variable entre 10 % y 80 % (ICA, 2023). Así, un modelo predictivo podría generar tratamientos correctivos a tiempo y disminuir la mortalidad de las aves de corral y consecuentemente disminuir las pérdidas.

# 2. Objetivos

## 2.1. Objetivo general

Desarrollar un método alternativo basado en algoritmos de Inteligencia Artificial (IA) para la detección de salmonella *spp*. en imágenes de heces de aves de corral tomadas en África entre 2020 y 2021.

## 2.2. Objetivos específicos

- Preparar un conjunto de imágenes de heces de aves de corral y sus respectivas etiquetas para obtener trenes de entrenamiento y evaluación para su posterior uso en modelos de *Machine Learning*.

- Programar y entrenar tres arquitecturas diferentes de Redes Neuronales Convolucionales para la clasificación de imágenes: ResNet, Inception y GoogLeNet, y ajustar los hiperpárametros para maximizar la precisión y sensibilidad.

- Elegir el modelo con mejor desempeño entre tres arquitecturas diferentes de Redes Neuronales Convolucionales según su sensibilidad (Recall), precisión (Accuracy), F1-Score y ROC.

- Analizar el impacto económico del modelo en términos de reducción de pérdidas por mortalidad, considerando la posibilidad de ejecutar acciones preventivas cuando se cuenta con detección temprana de Salmonella *spp*. en las aves de corral.

# 3. Datos

## 3.1. Datos originales

El archivo descargado desde Kaggle tiene como nombre predeterminado “archive.zip” y un tamaño de 249,1 MB. Esta carpeta comprimida contiene 8067 imágenes y un archivo .csv con las etiquetas correspondientes.

El archivo train_data.csv”, pesa 220 KB y consta de dos columnas: “images” y “label”, en donde se asocia el nombre de cada imagen con su respectiva etiqueta. De acuerdo con la información mostrada en el archivo “ME03_data_preparation_local.ipynb” disponible en el repositorio de GitHub (Ortega A., 2023), el archivo con las etiquetas comprende 8067 instancias y 2 columnas, ambas son tipo “Object”. Las instancias se encuentran clasificadas originalmente en 4 enfermedades diferentes: “Salmonella”, “Coccidiosis”, “New Castle Disease” y “Healthy”. No hay valores nulos en ninguna columna, además, la base de datos cargada utilizando Python ocupa 126.2+ KB, y es fácil de procesar en términos de almacenamiento y carga. La **Figura 2** muestra una vista previa de este archivo.

**Figura 2** *Vista previa del archivo train_data.csv con las etiquetas de cada instancia.*

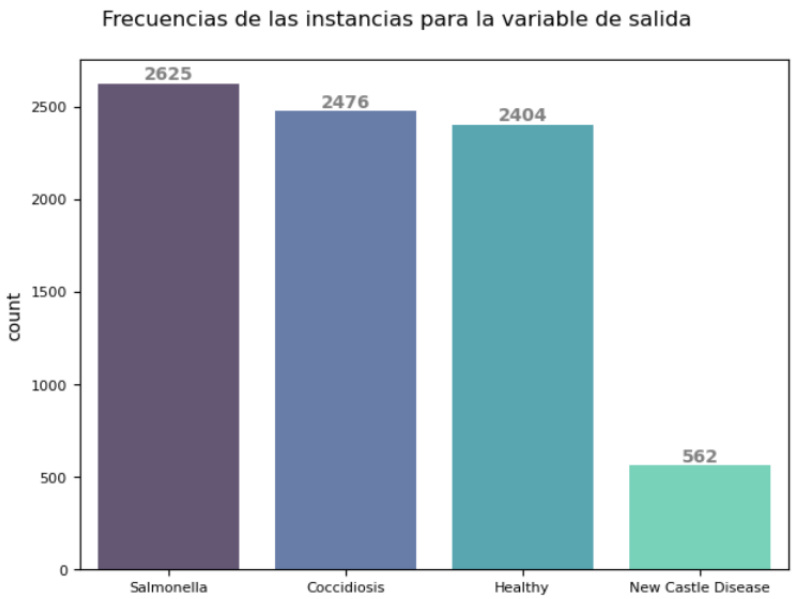
| | images | label |
|---|---|---|
| 0 | salmo.1558.jpg | Salmonella |
| 3 | salmo.1484.jpg | Salmonella |
| 5 | salmo.659.jpg | Salmonella |
| 6 | salmo.1386.jpg | Salmonella |
| 14 | salmo.2172.jpg | Salmonella |
| 16 | salmo.1291.jpg | Salmonella |
| 17 | salmo.1012.jpg | Salmonella |
| 18 | salmo.1532.jpg | Salmonella |
| 19 | pcrsalmo.129.jpg | Salmonella |
| 23 | salmo.1333.jpg | Salmonella |
| 25 | salmo.1407.jpg | Salmonella |
| 29 | salmo.1663.jpg | Salmonella |

*Nota.* Fuente: https://bit.ly/3MO6hx3 (Kaggle, 2022). Elaboración propia, código disponible en el repositorio asociado al proyecto (Ortega A., 2023).

La **Figura 3** muestra la frecuencia de las instancias para cada clase del archivo train_data.csv.

**Figura 3**

*Frecuencia de las instancias para la variable de salida.*




*Nota.* Fuente: https://bit.ly/3MO6hx3 (Kaggle, 2022). Elaboración propia, código disponible en el repositorio asociado al proyecto (Ortega A., 2023).

Adicionalmente, la carpeta descargada “archive.zip” contiene una subcarpeta llamada “Train” en donde se encuentran las imágenes en formato JPG. Todas tienen un tamaño de 224 x 224 pixeles y utilizan una codificación RGB, es decir, tienen 3 canales con tipo de dato “uint8”. Los datos almacenados como "uint8" son variables que almacenan números enteros de 8 bits sin signo. Por lo tanto, representan valores enteros no negativos. También indica que el número se almacena en 8 bits de memoria. Así, un número de 8 bits puede representar $2^8$ = 256 valores diferentes, que van desde 0 hasta 255, y es precisamente el rango de valores para cada píxel de la imagen a color (NumPy, 2022).

Las imágenes están etiquetadas con la clase correspondiente tanto en el nombre del archivo, como en el archivo train_data.csv. Las 2625 imágenes etiquetadas con presencia de Salmonella están marcadas como “salmo” o “pcrsalmo” y las 2404 imágenes sin Salmonella *spp.* están marcadas como “healthy” o “pcrhealthy”. El nombre de cada imagen contiene la palabra “healthy.” o “salmo.” y un consecutivo, por ejemplo: “healthy.2053.jpg” o “pcrsalmo.1.jpg”. La **Figura 4** muestra ejemplos de las imágenes disponibles.

**Figura 4**

*Ejemplo de las imágenes disponibles. Arriba de izquierda a derecha: salmo.11, salmo.41, salmo.27. Abajo, de izquierda a derecha: healthy.67, healthy.761, healthy.2049.*

*Nota.* Fuente: https://bit.ly/3MO6hx3 (Kaggle, 2022).

Para el desarrollo del proyecto se asume que las etiquetas están correctas y que no hay imágenes repetidas. La **diversidad** del conjunto de datos es amplia, contiene imágenes tomadas durante un periodo de 6 meses, se espera que el algoritmo de *ML* a implementar pueda generalizar a partir del conjunto de imágenes.

Es importante destacar que las fotos fueron tomadas en África y pueden no reflejar el entorno o condiciones en Colombia, además no se han actualizado desde febrero 2021. Algunas imágenes se ven borrosas y otros estudios han concluido que “los conjuntos de datos pueden ser ruidosos o de baja calidad cuando se recopilan en entornos agrícolas interiores y exteriores hostiles” (Ma et al., 2022). El data set no presenta restricciones y puede ser descargado desde la página Web de Kaggle (Kaggle, 2022). Además, el archivo .csv original y una muestra de 20 imágenes se encuentran disponibles en el repositorio de GitHub asociado a este proyecto (Ortega A., 2023).

## 3.2. Datsets

### *Limpieza de datos*

Como se describió anteriormente, las imágenes y etiquetas pertenecen a cuatro clases diferentes. No obstante, se propone una clasificación binaria entre “Healthy” y “Salmonella”. Así, el primer paso es filtrar la tabla, para que sólo las instancias donde la etiqueta pertenezca a las clases de interés permanezcan. Además, el nombre de las columnas se cambia siendo “filepaths” y “salmonella” para “images” y “label”, respectivamente. La columa “filepaths” es modificada para que no sólo contenga el nombre de las imágenes, sino la ruta donde se encuentran almacenadas. Luego, se verificó que no existieran registros duplicados. Finalmente, se dumificó la variable de salida utilizando el método “get_dummies” de pandas. Con lo anterior, se obtuvo 1 para las imágenes clasificadas con Salmonella y 0 para las imágenes saludables. Con lo anterior, se obtiene un archivo .csv modificado con las clases y tensor (arreglo de NumPy multidimensional) con las imágenes. La **Figura 5**, muestra el archivo .csv transformado.

**Figura 5**
*Vista previa de la transformación del archivo .csv con nuevo nombres de las columnas y sus dimensiones.*

| | filepaths | Salmonella |
|---|---|---|
| 0 | Data/imgs/salmo.1558.jpg | 1 |
| 1 | Data/imgs/salmo.1484.jpg | 1 |
| 2 | Data/imgs/salmo.659.jpg | 1 |
| 3 | Data/imgs/salmo.1386.jpg | 1 |
| 4 | Data/imgs/healthy.1748.jpg | 0 |
| ... | ... | ... |
| 5024 | Data/imgs/healthy.22.jpg | 0 |
| 5025 | Data/imgs/healthy.935.jpg | 0 |
| 5026 | Data/imgs/salmo.1607.jpg | 1 |
| 5027 | Data/imgs/salmo.1641.jpg | 1 |
| 5028 | Data/imgs/healthy.1145.jpg | 0 |

5029 rows × 2 columns

*Nota.* Fuente: https://bit.ly/3MO6hx3 (Kaggle, 2022). Elaboración propia, código disponible en el repositorio asociado al proyecto (Ortega A., 2023).

***Preprocesamiento de imágenes***

No es consideró realizar un aumento de datos. El data set se encuentra balanceado y las imágenes disponibles son suficientes para avanzar con el proyecto. Se sugiere verificar esta hipótesis cada vez que sea necesario.

Mediante un ciclo en Python cada imagen se abre y verifica que no fuera un archivo corrupto. Además, se redimensiona a 224 x 224 pixeles en caso de que alguna imagen no tuviera las dimensiones mencionadas en la fuente original de los datos. En el ciclo, también se estandarizaron los valores de los pixeles (teniendo en cuenta los tres canales) y se convirtió cada imagen desde una matriz de (224, 224, 3) hasta un vector unidimensional de 224 x 224 x 3 = 150228 posiciones.

No se encontraron archivos corruptos durante la ejecución del ciclo y resulto una lista con 5027 elementos, en donde cada uno corresponde a arreglo de NumPy que contiene los 150528 pixeles de cada imagen.

Se encontró una discrepancia entre el número de imágenes y el número de instancias con etiquetas. Se identificaron y eliminaron las instancias que no tenían imágenes correspondientes en la carpeta descargada, específicamente, "Data/imgs/pcrhealthy.35.jpg" y "Data/imgs/pcrsalmo.202.jpg".

***División de los datos en trenes de entrenamiento y prueba***

Primero se convierte la lista de imágenes en un arreglo de NumPy, con tamaño (5027, 150528) y se define *y* como el vector con las etiquetas. A continuación, se dividen ambos en los trenes de entrenamiento y test utilizando el método train_test_split de scikit-learn, con una proporción de 80 % de datos para el entrenamiento, aleatoriedad y una semilla para conservar la división en las diferentes ejecuciones del código. Además, se utilizó el parámetro “stratify” para garantizar que se mantuvieran balanceados los trenes de entrenamiento y prueba. Al final, resultaron los siguientes elementos:

Tamaño del tren de entrenamiento: (4021, 150528)
Tamaño de las etiquetas de entrenamiento: (4021,)
Tamaño del tren de prueba: (1006, 150528)
Tamaño de las etiquetas de prueba: (1006,)

### 3.3.Analítica descriptiva

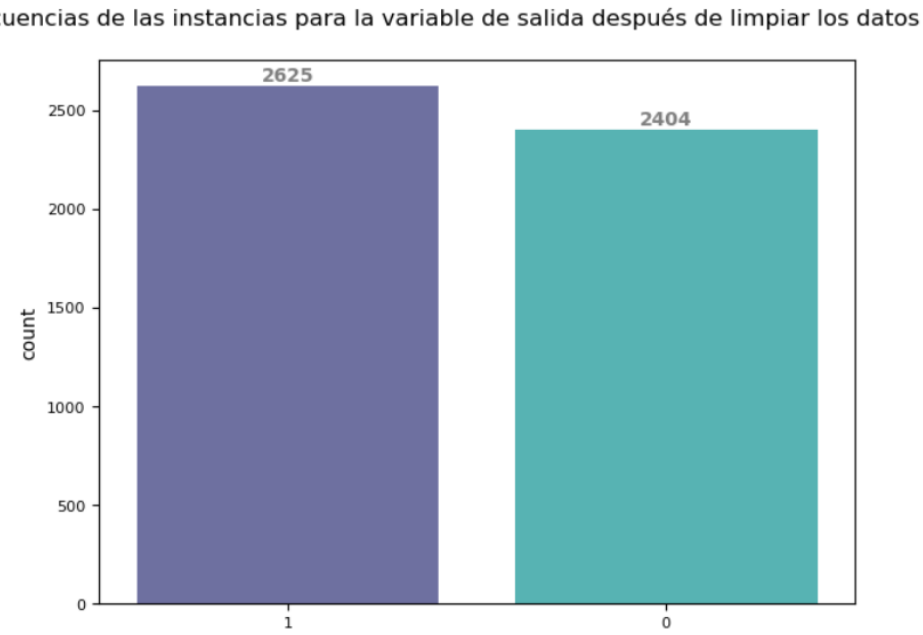
Los datos de entrada se encuentran en arreglo X, en donde cada posición contiene los pixeles de una imagen. Por ejemplo, la primera posición de la lista se ve así:

```
array([0.31372549, 0.3254902 , 0.38431373, ..., 0.6745098 , 0.67843137,
  0.64313725])
```

Respecto a las etiquetas almacenadas en el vector y, la variable objetivo presenta dos categorías muy equilibradas y el data set está balanceado con un 52,1 % de los registros en la categoría "Salmonella" y 47,9 % en la categoría "Healthy", representando aquellas imágenes que fueron confirmadas con presencia del microorganismo patógeno Salmonella *spp*. y las que no, respectivamente. La **Figura *6*** muestra el data set luego de limpiar los datos.

**Figura 6**

*Frecuencia de las instancias para la variable de salida luego de limpiar los datos.*




*Nota.* Fuente: https://bit.ly/3MO6hx3 (Kaggle, 2022). Elaboración propia, código disponible en el repositorio asociado al proyecto (Ortega A., 2023).

La

**Tabla *1*** muestra un resumen de los elementos resultantes de la preparación de los datos.

**Tabla 1**

*Resumen de elementos resultantes de la preparación de datos cuando se ejecuta el archivo ME03_data_prepatation.ipynb*

| # | Variable | Descripción | Tipo | Dimensión | Origen |
|---|---|---|---|---|---|
| **1** | df | Relaciona la ruta de cada imagen con su respectiva etiqueta dumificada. | DataFrame | (5027 , 2) | Archivo .csv “Test_data” |
| **2** | X | Cada posición contiene una imagen | Arreglo de Numpy | (5027, 150528) | Imágenes originales en la carpeta descargada |
| **3** | y | Cada posición contiene una etiqueta (0 o 1) para ausencia y presencia de Salmonella *spp*. respectivamente. | Arreglo de Numpy | (5027, ) | Archivo .csv “Test_data” |

## 4. Conclusiones

En este proyecto, se abordó la detección de Salmonella *spp*. en imágenes de heces de aves de corral. Se desarrolla un método alternativo utilizando Redes Neuronales Convolucionales (CNN) para clasificar las imágenes y predecir la presencia o no del microorganismo Salmonella *spp*.

El proceso de preparación de datos incluyó la transformación de las etiquetas en 0 y 1 para "Healthy" y "Salmonella", respectivamente. Se realizó una limpieza de los datos para garantizar la integridad de las imágenes y sus etiquetas. Se eligieron tres modelos de CNN como ResNet, Inception y GoogLeNet lo que permite explorar diferentes arquitecturas para la clasificación de las imágenes. Se ajustarán los hiperparámetros para maximizar la precisión y sensibilidad, evaluando el rendimiento con métricas clave como Recall, Accuracy, F1-Score y ROC.

Adicionalmente, se realizó un análisis descriptivo para entender el impacto potencial en la reducción de pérdidas económicas por mortalidad en aves de corral, y se espera que la detección temprana y precisa de Salmonella *spp*. implementado modelos de ML se valorada como una herramienta útil para la implementación de medidas preventivas, contribuyendo así a la salud de las aves y los consumidores, mientras se mejora la productividad en la cadena de producción de proteína animal.

Se recomienda realizar actualizaciones periódicas de la preparación de los datos y considerar mejoras continuas en la calidad de las imágenes, por ejemplo, técnicas para aumentar el contraste de las imágenes, encontrar si usar sólo un canal de color sería suficiente, realizar zoom a las imágenes para reducir el ruido alrededor, etc.

## Referencias


Food and Agriculture Organization (2019). La carga de los alimentos insalubres para la salud pública: la necesidad de un compromiso mundial. Disponible en: https://www.fao.org/3/CA3056ES/ca3056es.pdf

Federación Nacional de Avicultores de Colombia (3 de septiembre de 2023). Estadísticas. Disponible en: https://fenavi.org/estadisticas/consumo-per-capita-mundo-pollo/

FENAVI. (2023, 15 de noviembre). Boletín FENAVIQUIN, Edición 390. Disponible en: https://fenavi.org/boletin-fenaviquin/fenaviquin-edicion-390-noviembre-15-de-2023/

Hajian-Tilaki K. Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. Caspian J Intern Med. 2013 Spring;4(2):627-35. PMID: 24009950; PMCID: PMC3755824. Disponible en: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3755824/

Hossain, Md E., Kabir, M.A., et al. (2022). A systematic review of machine learning techniques for cattle identification: Datasets, methods and future directions. Artificial Intelligence in Agriculture, vol. 6, pp. 138-155. https://doi.org/10.1016/j.aiia.2022.09.002

Instituto Colombiano Agropecuario (ICA). Resolución No. 00017754 de 2019. Disponible en: https://www.ica.gov.co/getattachment/0b7d2fae-f96a-4b0d-8459-2b10be4142da/20194R1775.aspx

Instituto Colombiano Agropecuario (ICA). (2023). Salmonella. En Subgerencia de Protección Animal, Enfermedades Animales. Disponible en: https://www.ica.gov.co/areas/pecuaria/servicios/enfermedades-animales/salmonella.aspx

Kaggle (2022). Chicken Disease Image Classification. Disponible en: https://www.kaggle.com/datasets/allandclive/chicken-disease-1

Luyao Ma, Jiyoon Yi, Nicharee Wisuthiphaet, Mason Earles, Nitin Nitin (2022). Accelerating the Detection of Bacteria in Food Using Artificial Intelligence and Optical Imaging. Applied and Environmental Microbiology, vol. 89, No. 1. https://doi.org/10.1128/aem.01828-22

Ministerio de Agricultura - el campo es de todos (2023). Boletín interactivo epidemiológico pecuario. Disponible en: https://bit.ly/47ni4uC

NumPy. (2022). numpy.uint. En Reference, Arrays, Scalars. Disponible en: https://numpy.org/doc/stable/reference/arrays.scalars.html#numpy.uint

Ortega Alzate, M. (2023). Redes Neuronales Artificiales para la detección de Salmonella *spp*. en aves de corral. https://github.com/melissaortegaa/monografia

Organización Mundial de la Salud (OMS) (30 de abril de 2020). Inocuidad de los alimentos. Disponible en: https://www.who.int/news-room/fact-sheets/detail/food-safety

Real Academia Española (RAE) (2022). Prevalencia. Disponible en: https://dle.rae.es/prevalencia

S. A. Shaik Mazhar & D. Akila (2022). Machine Learning and Sensor Roles for Improving Livestock Farming Using Big Data. Cyber Technologies and Emerging Sciences, vol 467, pp. 181-190. https://link.springer.com/chapter/10.1007/978-981-19-2538-2_17

Subgerencia de Protección y Regulación Pecuaria. Grupo de Epidemiología Veterinaria. (2007). Informe Técnico: Sistema de Información y Vigilancia Epidemiológica Colombia, Sanidad Animal 2006 (Código: 00.03.12.07, Edición: diciembre de 2007). ISSN: [número de ISSN]. Bogotá, D.C. Disponible en: https://www.ica.gov.co/getattachment/e2e4ba97-a885-4b85-ba55-19188b37d6de/2.aspx

T. Obe, T. Boltz, M. Kogut, et al. (2023). Controlling Salmonella: strategies for feed, the farm, and the processing plant. Poultry Science, vol. 102. https://doi.org/10.1016/j.psj.2023.103086

# Anexos

## Anexo 1. Requerimientos de paquetes para ejecutar archivos .ipynb

Los siguientes paquetes y sus respectivas versiones deben ser instaladas para la correcta ejecución de los Notebooks referenciados en el desarrollo de este proyecto:

ipykernel == 6.22.0

ipython == 8.12.0

keras == 2.14.0

matplotlib-inline == 0.1.6

pandas == 2.1.1

numpy == 1.26.1

scikit-image == 0.22.0

scipy == 1.11.3