INTEGRACIÓN DE UNA RED NEURONAL CONVOLUCIONAL PARA LA DETECCIÓN DE ELEMENTOS DE PROTECCIÓN PERSONAL EN LOS LUGARES DE TRABAJO

Presentado por:
Iván Hernández Ruiz

Universidad del Sinú Seccional Cartagena

Facultad de Ciencias Exactas e Ingenierías

Escuela de Ingeniería de Sistemas

Cartagena - Colombia

Junio 2023

UNIVERSIDAD DEL SINÚ
Elías Bechara Zainúm
Seccional Cartagena

INTEGRACIÓN DE UNA RED NEURONAL CONVOLUCIONAL PARA LA DETECCIÓN DE ELEMENTOS DE PROTECCIÓN PERSONAL EN LOS LUGARES DE TRABAJO

Trabajo de grado presentado como requisito para optar el título de
INGENIERO DE SISTEMAS

Asesor disciplinar:
Ph. María Claudia Bonfante

Asesor metodológico:
Mg. Eugenia Arrieta

Universidad del Sinú Seccional Cartagena

Facultad de Ciencias Exactas e Ingenierías

Escuela de Ingeniería de Sistemas

Cartagena - Colombia

Junio 2023

AGRADECIMIENTOS

Queridos miembros del comité evaluador, profesores y compañeros de la universidad; con gratitud que me dirijo a ustedes en este momento para expresar mis más sinceros agradecimientos por su apoyo, acompañamiento y guía durante el desarrollo de mi proyecto de grado. Ha sido un camino lleno de retos y adquisición de nuevos conocimientos y no pudo haber sido posible sin su valiosa contribución.

Primeramente, agradezco a Dios, por brindarme la sabiduría, fortaleza, inspiración y la perseverancia necesaria para llevar a cabo este proyecto. Su guía misericordiosa ha sido mi luz en momentos de inseguridad y su amor inagotable ha sido mi motor para seguir adelante.

Agradezco de manera especial a mi respetada directora de proyecto Dra. María Claudia Bonfante cuyos conocimientos y dedicación han sido fundamentales para mi crecimiento académico y personal. Su apoyo constante, paciencia y orientación me han impulsado a alcanzar nuevas metas y superar mis propios límites. Cada consejo, corrección y sugerencia ha sido invaluable y ha contribuido en gran medida a la calidad de mi trabajo.

Asimismo, deseo expresar mi profundo agradecimiento a los profesores Eugenia Arrieta y Luis Murillo por contribuciones a este proyecto, a la universidad por brindarme el ambiente propicio para desarrollar mis habilidades y conocimientos.

Por último, pero no menos importante, quiero agradecer a mis seres queridos, familiares y amigos, quienes han estado a mi lado brindándome su apoyo incondicional, alentándome en los momentos difíciles y compartiendo mi alegría en cada logro alcanzado. Su amor y confianza han sido mi mayor motivación y fuente de inspiración.

CONTENIDO

LISTADO DE ILUSTRACIONES

LISTADO DE TABLAS

4 RESUMEN

La pandemia de COVID-19 ha generado una mayor comprensión acerca de la importancia de que los trabajadores lleven y utilicen adecuadamente equipos de protección personal. Esto busca no solo reducir los riesgos de contagio, sino también salvaguardarlos de posibles accidentes laborales a los que se enfrentan en sus lugares de trabajo. Asimismo, se aceleró la confianza en el uso y adopción de las tecnologías 4.0 para solucionar problemas reales; es nuestro caso de interés su aplicación a los entornos de Seguridad y Salud en el trabajo (SST), que se ha constituido como una normativa global que garantiza el bien de los empleados en todas sus dimensiones.

Este proyecto de culminación de pregrado persigue desplegar modelos soportados en Redes Neuronales Convolucionales que permitan detectar a trabajadores que no portan consigo elementos de protección personal como mascarillas faciales y cascos industriales.

Para la construcción de la solución, inicialmente se construyó un dataset propio de 2.000 imágenes etiquetadas con y sin elementos de protección personal como mascarillas faciales y cascos industriales, posteriormente se construyó el modelo de detección soportado en algoritmo (Yolo) You Only Look Once, el cual es muy utilizado para esta tarea según revisión de la literatura realizada. El experimento se realizó con las versiones Yolov5 y Yolov6, para el entrenamiento utilizó el 80 de las imágenes, el 15% para prueba y un 5% validación del modelo. Se seleccionaron las métricas de Precisión, exhaustividad, la precisión promedio de cada clase es equivalente al AP, con este valor se calcula la media de la precisión promedio (mAP) en ambas versiones de los algoritmos utilizados. Finalmente, se identificó que YOLOv7 logró una mayor tasa de éxito de mAP@.5 del 0,8.

Se obtuvo un modelo que puede ser desplegado como una aplicación que puede generar en un futuro alertas e informes estadísticos integrado al sistema de gestión de salud y seguridad en el trabajo.

Palabras Claves:

Redes Neuronal Convolucional, You Only Look Once (Yolo), Sistema de Salud y Seguridad en el Trabajo, Elementos de Protección Personal.

5 INTRODUCCIÓN

Tanto la Organización Internacional del Trabajo (OIT) como la Organización Mundial de la Salud (OMS) han insistido en múltiples ocasiones a los gobiernos a que establezcan políticas públicas de salud y seguridad que motiven a los empleadores a invertir en medidas preventivas para reducir los accidentes laborales y las enfermedades relacionadas con el trabajo. Esto se debe a que los costos económicos y sociales asociados con este problema son ampliamente subestimados. Esto se debe a que los accidentes laborales afectan la productividad y sostenibilidad de las compañías y de la sociedad en conjunto.En Colombia, ya sea por requerimientos regulatorios, calidad de procesos, requerimientos de empresas internacionales o la exportación de productos a mercados extranjeros, la necesidad de contar con un sistema certificado se ha incrementado en los últimos años. Anteriormente, gestionar el riesgo laboral en elpaís solo significaba seguir las reglas de tener un programa de medicina del trabajo; Sin embargo, desde 2012, las empresas deben tener un sistemade gestión para predecir, examinar, evaluar e inspeccionar los peligros que pueden repercutir la salud y la seguridad. De manera general, un Sistema de Gestión de Seguridad y Salud en el Trabajo (SGSST) constituye una herramienta que permite implementar medidas preventivas en una organización, brindando los recursos necesarios para administrar de manera coherente, organizada y estructurada la seguridad y la salud laboral. A través del uso de SGSST, una organización puede reducir el número de accidentes, así como aumentar la eficiencia del trabajo, lo que repercute directamente en los efectos monetarios de la empresa.

Uno de los principales aspectos que cubre el SGSST es el uso de Elementos de Protección Personal (EPP), los cuales se sugieren como caminos deseables para minimizar los riesgos de lesiones relacionadas con el mal uso de herramientas y la exposición a ambientes externos. Sin embargo, en la práctica, los trabajadores no están dispuestos a usarlos debido a varias razones, como: impactos negativos en su productividad; uso incómodo, particularmente enclimas cálidos; falta de capacitación para usar el EPP adecuado; subestimación de la utilidad de los EPP y falta de información sobre la frecuencia de lesiones yla

posibilidad de muertes causadas por el uso inadecuado de herramientas. Existen principalmente tres tipos de tareas de detección que benefician elmonitoreo de operaciones de seguridad en los trabajadores, a saber, detección de EPP, detección de interacción humano-objeto y detección de pose humana [1]. La detección de objetos utilizando el enfoque de Deep Learning y de la manode la Visión por Computadora, funciona para identificar y localizar objetos de determinadas clases en imágenes y vídeos en tiempo real, para una evaluación precisa en entornos laborales [2].

Esta investigación se enfoca en la detección de EPP, para abordar el problema de salud y seguridad en el trabajo, se propone implementar un modelo de Deep Learning soportado por una Red Neuronal Convolucional para la detección de mascarilla facial y casco industrial, utilizando el modelo Yolo el cual es uno de los algoritmos más extensamente utilizado para este propósito y disponible de forma pública en internet. El modelo entrenado, se construye a partir de una metodología estructurada en fases para el cumplimiento de los objetivos propuestos y se convierte en una importante herramienta tecnológica 4.0 para alertar a los trabajadores y supervisores en cargados de la SST sobre laausencia de estos elementos de sus trabajadores.

I. DISEÑO METODOLÓGICO

En este capítulo se describe el diseño metodológico del proyecto, el cual permite brindar un contexto de la intención y objetivos de este trabajo, teniendo en cuenta, descripción del problema, justificación, marco referencial, metodología y consideraciones éticas, etc. lo cual permite entender la importancia del proyecto, además también se define la organización de los procesos de investigación que se tuvieron en cuenta para guiar y/o orientar el proyecto.

1.1 Descripción del Problema

La Seguridad y Salud en el Trabajo (SST) es un área de gestión que su objetivo es evitar los daños y enfermedades derivados de las condiciones laborales, así como proteger y fomentar la salud de los trabajadores. Se enfoca en mejorar tanto las condiciones laborales como el entorno de trabajo, promoviendo la salud ocupacional y manteniendo el bienestar físico, mental y social de los trabajadores en todas las ocupaciones. [3]

Toda empresa de cualquier disciplina debe tener un departamento de gestión de seguridad y salud en el trabajo, los salubristas deben definir, observar, informar, intervenir y evaluar los peligros exasistentes en los sitios laborales y su incidencia en la compañía. En estos tipos de cargos puede estar el error humano en donde el salubrista no perciba que el empleado no lleva todos o parcial las herramientas de seguridad en labor, pudiendo repercutir deterioro de la salud del trabajador, disminución del tiempo en producción, sanciones disciplinarias y hasta perdidas de vida.

Sin embargo, las compañías y sobre todo el área de salud y seguridad en el trabajo no tienen la capacidad de construir herramientas tecnológicas que permitan detectar en tiempo real el porte de EPP de sus trabajadores soportadas por las funcionalidades que ofrecen las Redes Neuronales Convolucionales y sus modelos de detección. Ante ese panorama se propone como trabajo de fin

de pregrado propone inicialmente identificar los modelos de detección más utilizados y probados, comparar los modelos y seleccionar el de mejor desempeño que se pueda integrar a una solución tecnológica que permita la captura un trabajador en su entorno real, posteriormente la imagen pase por el modelo previamente entrenado con un dataset propio de imágenes de tapabocas y caso de seguridad, lo cual permite posteriormente generar alertas sobre la falta de EPP de tal forma que el supervisor pueda tomar la decisión pertinente, y al mismo tiempo con esta solución se pretende generar confianza en el sector producto sobre el uso de las aplicaciones soportadas por las CNN.

1.2 Justificación

La normativa legal sobre la SST están obligando a las empresas de diferentes sectores a implementar un programa que permita identificar, evaluar, controlar, medir, monitorear e informar los peligros en donde están expuestos sus trabajadores para evitar sanciones legales, sino a transformar sus procesos haciendo uso de tecnologías 4.0 como, Visión por Computador, Machine Learning, Big Data y Cloud Computing, ya que suministran datos, validación de algoritmos predictivos, de clasificación y de detección que permiten tomar decisiones a nivel organizacional.

Este trabajo de culminación de pregrado tiene el desafío de integrar estas tecnologías para abordar problemas relacionados con la protección y prevención de riesgos labores asociados con las lesiones y enfermedades producidas en el trabajo, específicamente detectar si los empleados llevan o no adecuadamente sus implementos de protección personal en su ambiente laboral.

Asimismo, esta propuesta de investigación se encuentra articulada con la Política Pública para la Transformación digital y la Inteligencia Artificial [3] y además con el programa 2022-2026 del gobierno actual y su objetivo *“DE UNA ECONOMÍA EXTRACTIVISTA HACIA UNA ECONOMÍA PRODUCTIVA*”, donde se desprende la línea: 2.3. La democratización del Saber: el conocimiento humano para la transformación productiva. El cual promueve la Investigación básica y aplicada para potenciar la innovación tecnológica del sector productivo y de la

sociedad. En consonancia con la planeación estratégica, este proyecto de investigación, también se encuentra enmarcado en el Plan de Desarrollo Departamental: Bolívar Competitivo, para la Inclusión Social: Bolívar Primero en Ciencia Tecnología e Innovación, y el plan de desarrollo municipal “Salvemos Juntos a Cartagena” y su línea estrategia Línea Estratégica: Competitividad e Innovación. Finalmente, esta propuesta le apartará a la línea de investigación institucional “Ciencia, Tecnología e Innovación. (Tics para la integración- Plan de desarrollo departamental)”, y a la línea en Inteligencia Artificial de la escuela de Ingeniería de Sistemas, al proyecto de investigación docente “INTEGRACIÓN DE TECNOLOGÍAS 4.0 PARA LA SALUD Y SEGURIDAD EN EL TRABAJO” financiado con recursos de vigencia 2022. de la dirección de investigación de la Universidad del Sinú Seccional Cartagena.

1.3 Alcance

El objetivo de este proyecto consiste en integrar y probar un modelo de Redes Neuronales Convolucionales (CNN) utilizando un conjunto de datos de imágenes que representan elementos de protección personal, específicamente tapabocas y cascos. El propósito principal de este modelo es asegurar la detección temprana y facilitar la toma de decisiones en los Sistemas de Gestión de Salud y Seguridad en el trabajo.

La implementación de esta solución, de bajo costo, está dirigida al sector productivo en el departamento de Bolívar. Su finalidad es contribuir a la reducción de los elevados índices de accidentes laborales y enfermedades, brindando una herramienta eficaz. Mediante el uso de esta tecnología, se busca favorecer la prevención de riesgos y promover un entorno laboral más seguro y saludable en la región.

1.4 Pregunta de Investigación

¿Cómo construir y entrenar un Modelo de Redes Neuronales Convolucionales que detecte la presencia o ausencia de mascarillas y cascos de seguridad?

1.5 Objetivos

1.5.1 Objetivo General

Construir y entrenar un modelo de Red Neuronal Convolucional (CNN) para la detección de elementos de protección personal.

1.5.2 Objetivos Específicos

- Definir la metodología para la construcción del modelo de detección de elementos de protección personal.
- Realizar un entrenamiento utilizando dos versiones del algoritmo Yolo para la detección de imágenes de personas con o sin mascarillas facialesy con y sin casos industriales.
- Comparar las métricas de desempeño de los algoritmos entrenados de CNN para la detección de elementos de protección personal.

1.6 Estado del Arte

Para responder a la pregunta de investigación ¿Cómo construir y entrenar un Modelo de Redes Neuronales Convolucionales que detecte el inadecuado uso de elementos de protección de personal en el trabajo?, se hace revisión inicial a los trabajos previos haciendo uso de bases de datos como Scopus, IEEE y Science Direct, a partir de cadena de búsqueda con los siguientes términos: (“Convolutional neural network” or “CNN”) and ("mask detection” OR "helmet detection") and (“detection methods” or “detection model”). Cómo criterios de inclusión se consideraron solo artículos y artículos de conferencia de los últimos 3 años (2020 al 2022). El resultado de la consulta preliminar se muestra en la tabla 1.

Tabla 1 Resultado de las cadenas de búsquedas

Base de Datos	2020	2021	2022	Total
Scopus	4	27	28	59
WoS	0	1	7	8
Science Direct	0	7	11	18

Cómo se puede observar en la tabla anterior, la base de datos Scopus arroja un total de 59 contribuciones y en los años de divulgación de estos artículosy se puede visualizar que en el 2021 y 2022 hay una tendencia constante de publicaciones sobre el tema, por lo que se puede concluir que la comunidad académica y científica demuestra un interés en las aplicaciones de las CNN para la detección de objetos en imágenes y videos en las áreas de transporte público, vigilancia, criminalística, salud y en la vida cotidiana.

En la tabla 2. Se visualiza un listado de las importantes contribuciones que muestran los modelos de detección utilizados.

Tabla 2 Listado de contribuciones que arrojo la búsqueda en bases de datos

Referencia	**Año**	**Modelo deDetección**	**Métrica Aplicada**	**Modelo con mejor desempeño**
[4]	2023	● ResNet50, ● VGG11, ● InceptionV3, ● EfficientNetB4, ● YOLOv2, ● YOLOv3, ● YOLOv4	Precisión	● EfficientNetB4 (95%) ● YOLOv 4 (93%)
[5]	2022	●YOLO-v2 with ResNet-50 ●YOLOV5	Precisión	● YOLO V5 (97.90%) ● Multi-
		● YOLOv4 ● Multi-StageCNN ● CNN, Procrustes Analysis		Stage CNN (99.49%)

[6]	2020	• CASP (propuesto porel autor) • Faster R-CNN ResNet + FPN • SSD • DSSD • RFBNe • YOLOv3	mAP	CASP (88,11%)
[7]	2022	• AlexNet • Mobinet • YOLO • Propuesto por el autor	• Precision Recall Accuracy	Propuesto por el autor: Precisión (97%) Recall 98% Accuracy 99%
[8]	2023	• ResNet 50 • Faster R-CNN CNN • TrVGG16 + BiLSTM • SVM • SSDMNV2 • FMD-Yolo • YOLO v2 • hybrid deep learning approach (ASMFO-HResMobileNet)	• Accuracy • Sensitivity • Specificity • Precision • 1-score	Propuesto por el autor: • ASMFO-HResMobileNet
[9]	2022	• CNN • ResNet50+SVM	Accuracy AP	YOLOv4-tiny mejorado por

		• YOLOv2+ResNet • YOLOv3 • Faster RCN FMD-Yolo • Efficient+YOLOv3 MarkHunter • (SE)-Yolo Improved YOLOv4 • SSD+MobileNetv2 • YOLOv4-tiny-SPP • YOLOv3 • YOLOv3-Tiny	mAP	el autor

Fuente: Elaboración propia

En [4] se hace una comparación de los ResNet50, VGG11, InceptionV3, EfficientNetB4, and YOLO, mediante el uso del conjunto de datos (MaskedFace-Net), los resultados mostraron que la arquitectura EfficientNetB4 tiene una mejor precisión con un 95,77 % en comparación con la arquitectura YOLOv4 de93,40 %, InceptionV3 de 87,30 %, YOLOv3 de 86,35 %, ResNet50 de 84,41 %,VGG11 de 84,38 %, y YOLOv2 de 78,75%, respectivamente.

En [7] se propone un sistema de detección de máscaras faciales basado en clasificadores profundos de CNN, se implementa un kit Raspberry Pi y herramientas de análisis de datos de código abierto. Utilizando un conjunto de datos en tiempo real con una colección de 1386 imágenes (con máscaras faciales e imágenes sin máscara facial). El modelo propuesto se comparó con otros modelos existentes como Alexnet, Mobinet y YOLO. Se realizaron las métricas de rendimiento, como la tasa de error, la velocidad de inferencia, la precisión, la recuperación, la exactitud y el análisis de sobreajuste. Los resultados revelan que el sistema propuesto ha superado a los modelos existentes con una precisión superior al 99 % y una tasa de error inferior al 2 %.

Otro aporte es un modelo de detección de máscara facial con IoT utilizando como base Single Shot Detector (SSD) y un método híbrido de aprendizaje profundo [8]. La novedad del modelo propuesto es que la mejora realizada en ladetección y clasificación de rostros es que utiliza un algoritmo de optimización denominado ASMFO. Posteriormente, EL modelo propuesto con IoT se compara con los algoritmos convencionales y los clasificadores existentes con varias medidas, el cual logra una precisión un 18,57 %, 15,67 %, 17,56 %, 16,24 % y un19,2 % superior a la de SVM, CNN, VGG16-LSTM, ResNet 50, MobileNetv2 y ResNet 50-MobileNetv2.

Asimismo, identificó trabajos que hacen mejoras a los modelos de detección existentes de elementos de protección como el SSDMNV2 model [10], que utiliza SSD como detector de rostros y la arquitectura MobilenetV2 alcanzando una puntuación de precisión de 0,9264 y una puntuación F1 de 0,93. Otro caso a resaltar es CASP [6], el cual está basado en el modelo SSD, este modelo demostró una precisión media (mAP) del 88,1 % utilizando un dataset de cascos de seguridad.

1.7 Marcos de Referencial

1.7.1 Marco Teórico

Redes Neuronales Convolucionales (CNN)

Las redes convolucionales son un conjunto de redes diseñadas siguiendo el funcionamiento del cerebro, con la capacidad de aprender en diferentes niveles de abstracción. En la primera capa, se identifican formas simples, colores y bordes. En la siguiente capa, se pueden distinguir combinaciones de bordes y colores, mientras que la última capa se enfoca en la forma precisa para determinar su naturaleza. Para lograr esto, los ordenadores utilizan filtros o lentes que permiten identificar distintas características, como bordes diagonales o colores.

Estas redes convolucionales procesan las imágenes mediante el escaneo y

aplicación de filtros en toda la imagen. A partir de este proceso, se definen y clasifican las características presentes. A diferencia del enfoque clásico de procesamiento de imágenes, que se basa en algoritmos definidos por humanos, las redes convolucionales extraen automáticamente características a partir de los datos de entrenamiento. Estas características se utilizan luego para la clasificación de objetos.

Las CNN se han tenido adelantos importantes en el sector del reconocimiento facial, detección de fatiga [11], detección de conductores distraídos [12] e infractores de normas de tránsito [13], la clasificación de imágenes y detección visual de objetos en tiempo real en el área de criminalística [14], vigilancia y transporte [15]. En sectores como la construcción las técnicas de Deep Learning hacen posible el seguimiento de equipos y detección de grietas [16], a partir de conjunto de datos de objetos que sirven de base para entrenar modelos de detección de objetos y probar el rendimiento de estos [17].

La función más básica y fundamental de las CNN es la clasificación de imágenes por diferentes categorías y agruparlas según sus atributos comunes en como muestra la Ilustración.1. e Ilustración 2. Este método se puede utilizar en la seguridad y salud en el trabajo para detectar con ayudas de cámaras de profundidad si los trabajadores llevan consigo mascarillas y/o cascos de seguridad.

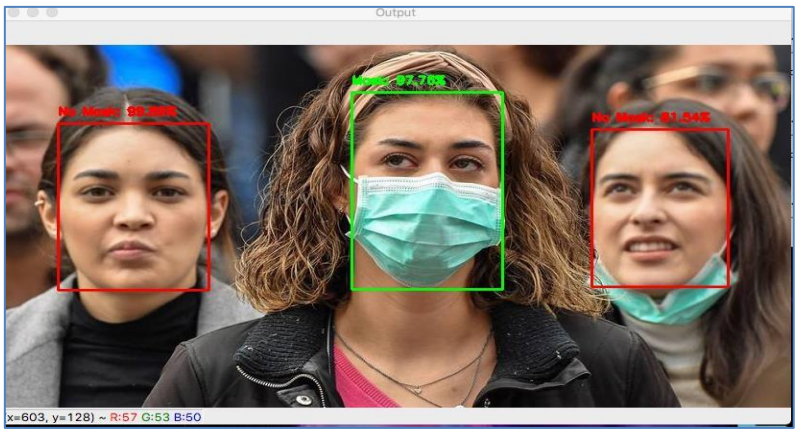

Ilustración 1 Clasificación de personas que llevan o no mascarillas. Fuente Búsqueda en internet

Ilustración 2 Clasificación de personas que llevan o no cascos de seguridad. Fuente Búsqueda en internet

Modelos de detección de objetos

Los modelos de detección de objetos están soportados por las arquitecturas backbone, que es uno de los componentes más importantes estos. Estas redes extraen características de la imagen de entrada utilizada por los modelos existentes [18], estás arquitecturas son: AlexNET, VCG-16, GoogleNET, ResNET-50, ResneXt-50, CSPresNext-50 y EfficientNetB4. Los modelos de detección basados en CNN se dividen en 3 categorías:

a) Método de una etapa

Son capaces de analizar una imagen con una sola valoración de red, estos se enfocan en todas las propuestas de regiones espaciales para la detección de objetos a través de una arquitectura relativamente más simple. Los algoritmos de una sola etapa son más rápidos, ejemplos de este tipo son:

- **Single Shot Detector (SSD)** [19]: Es el primer detector de una sola etapa que igualó la precisión de los detectores contemporáneos de dos etapas como Faster R-CNN, manteniendo la velocidad en tiempo real. SSD se construyó sobre VGG-16, con estructuras auxiliares adicionales para mejorarel rendimiento. Estas capas de convolución auxiliares, añadidas al

final del modelo, disminuyen progresivamente de tamaño. Es capaz de detectar objetos más pequeños antes en la red cuando las características de la imagen no son demasiado toscas, mientras que las capas más profundas son responsables del desplazamiento de los cuadros predeterminados y las relaciones de aspecto. Aunque SSD es significativamente más rápido y preciso que las redes de última generación como YOLO y Faster R-CNN, tenía dificultades para detectar objetos pequeños. SSD utiliza una estructura piramidal combinada con mapas de características de diferentes tamaños de campos receptivos, que realiza regresión de posición, tal cómo se muestra en su arquitectura de la ilustración 3.

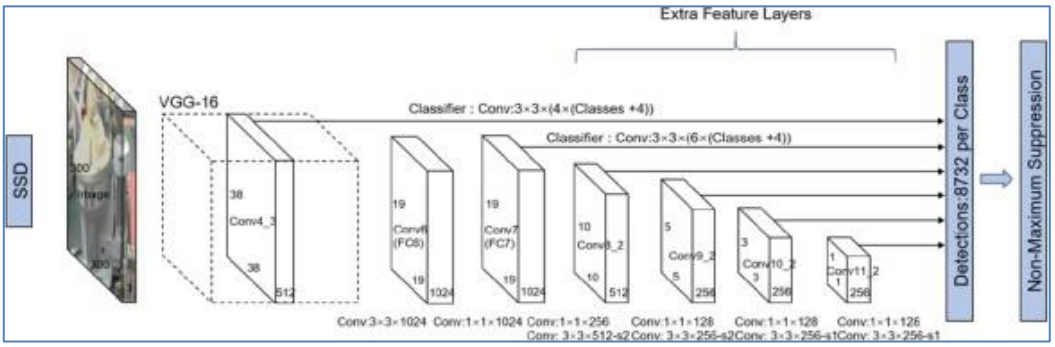

Ilustración 3 Arquitectura de Single Shot Detector. Tomado de [19]

- **RetinaNet** [20]**:** Es un modelo de red neuronal de una etapa popular en los últimos años que utiliza la función de pérdida focal (focal loss function). La pérdida focal es una mejora respecto a la pérdida de entropía cruzada, diseñada para abordar el problema del desequilibrio de clases en modelos de detección de objetos de una sola etapa. Esta forma de pérdida busca reducir el impacto de los valores correctamente predichos por la red, concentrándose en los casos en los que se ha producido una predicción incorrecta de clase.

  Este modelo predice objetos mediante un muestreo denso de la imagen de entrada en ubicación, escala y relación de aspecto. Utiliza la arquitectura ResNet aumentado por Feature Pyramid Network (FPN), como la columna vertebral y dos subredes similares: clasificación y regresor de cuadro delimitador. Cada capa del FPN se pasa a las

subredes, lo que le permite detectar objetos en varias escalas. es fácil de entrenar, converge más rápidoy es fácil de implementar. Logró un mejor rendimiento en precisión y tiempo de ejecución que los detectores de dos etapas. RetinaNet también amplió loslímites en el avance de las formas en que se optimizan los detectores de objetos mediante la introducción de una nueva función de pérdida.

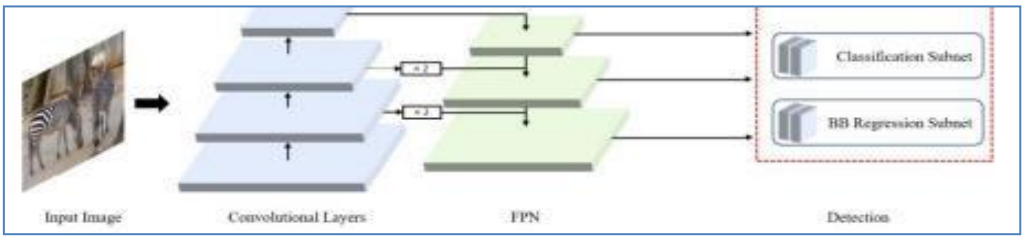

Ilustración 4 Arquitectura de RetinaNet. Tomado de [18]

- **EfficientDet [21]:** Estaba basado en optimizaciones claves para optimizar la eficacia de los algoritmos de detección. Inicialmente, se soporta en una red piramidal de características bidireccional ponderada (BiFPN), que logra una fusión de características multi escala fácil y rápida; la segunda optimización es un método de escalado compuesto que escala uniformemente la resolución, la profundidad y el ancho para todas las redes de predicción de red roncal, de características y de caja/clase al mismo tiempo. Sobre la base de estas optimizaciones y las redes troncales de EfficientNet, se desarrolló EfficientDet, que logra de manera consistente una eficiencia mucho mejor que la técnica anterior en un amplio espectro de limitaciones de recursos con mayor precisión y eficiencia.

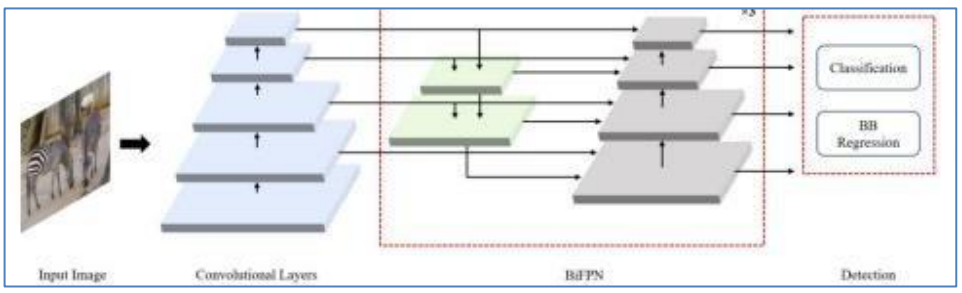

Ilustración 5 Arquitectura de EfficientDet. Tomado de [18]

- **You Only Look Once (YOLO)** [22]**:** El algoritmo YOLO popularizó el

enfoque de una etapa demostrando predicciones en tiempo real y logrando una velocidad de detección notable. En la arquitectura YOLO, se divide una imagen en una cuadrícula de tamaño GxG, y cada cuadrícula genera N predicciones de cuadros delimitadores. En cada predicción, se restringe a un cuadro delimitador a tener únicamente una clase, lo que limita la capacidad de la red para detectar objetos más pequeños.

YOLO se inspiró en el modelo GoogLeNet para la clasificación de imágenes, que utiliza módulos en cascada de redes de menor convolución. Está pre- entrenado en datos de ImageNet hasta el modelo logra una alta precisión y, en consecuencia, se modifica agregando una convolución inicializada aleatoriamente y capas completamente conectadas. En el momento del entrenamiento, las celdas de la cuadrícula predicen solo una clase a medida que la red converge mejor, pero se puede aumentar durante la entrada. La arquitectura de Yolo se muestra en la ilustración 6.

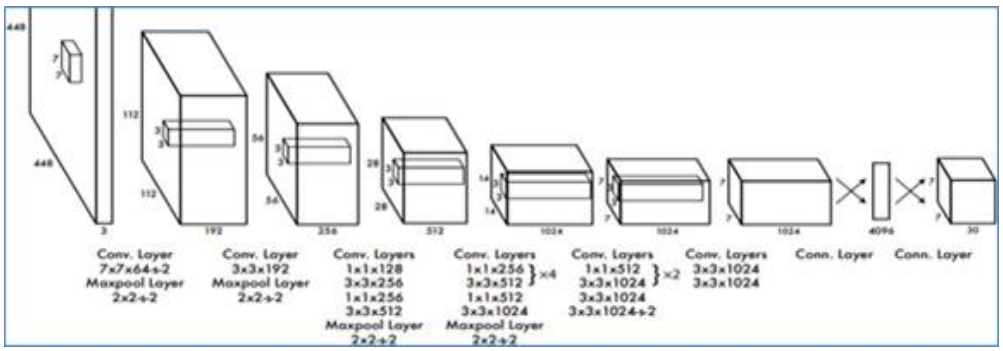

Ilustración 6 Arquitectura de Yolo.Tomado de [22].

El proceso de inferencia en general implica dividir la imagen en una cuadrícula de tamaño SxS. Si el centro de un objeto se encuentra en una celda de la cuadrícula, esa celda se encarga de detectar el objeto. Cada celda detecta B cajas delimitadoras y proporciona un nivel de confianza para cada una de esas cajas. Este nivel de confianza refleja cuánto confía el modelo en que la caja contiene un objeto y cuánto confía en que la caja detectada realmente corresponde al objeto

detectado. En otras palabras, la confianza se calcula como la probabilidad de que exista un objeto multiplicada por el índice de superposición (IOU) entre la caja detectada y la caja real. Si no hay objeto en esa celda, la confianza debería ser 0, mientras que en caso contrario debería ser el IOU entre la caja detectada y la caja real. Cada caja delimitadora está compuesta por 5 predicciones: las coordenadas (x, y) que representan el centro de la caja detectada, y las dimensiones (w, h) que representan el ancho y la altura de la caja.

Evolución del YOLO

Este algoritmo ha sido optimizado en diferentes versiones y varios trabajos, por ejemplo, en [23] la red se mejoró a YOLOv2 que incluía normalización por lotes, clasificador de alta resolución y cuadros anclaje. Posteriormente se mejora a YOLOv3 cuya estructura se conoce como Bloques Residuales [24], los cuales se emplean para el aprendizaje de características y están compuestos por conexiones convolucionales, la habilidad de detectar en tres niveles diferentes es su característica distintiva, lo cual hace que objetos de varios tamaños se reconozcan más correctamente como resultado de esto. Más adelante surge elmodelo YOLOv4, el cual puede reconocer múltiples objetos en un fotograma único, la red divide la imagen en diferentes regiones y predice cuadros delimitadores y sus probabilidades para identificar una clase de objetos, su versión mejorada YOLOv4-Tiny es más ligera y está diseñado para reducir el tiempo para la detección de objetos. Por lo tanto, YOLOv4-Tiny admite el análisisde imágenes en tiempo real también cuando se ejecuta en sistemas integrados o dispositivos móviles. En cuanto a las métricas de rendimiento, YOLOv4 es muypreciso en la detección de la correcta del casco de seguridad con un mAP@50 igual al 94,03% y un mAp@75 igual a 66,13%, mientras YOLOv4-Tiny, a pesar de ser caracterizado por el más alto rendimiento en términos de entrenamiento y velocidad de detección, pierde en precisión con respecto a YOLOv4[25]. Otro trabajo de

aplicación de Yolo propone un detector que aumenta los valores de AP (precisión promedio) en cada categoría del conjunto de datos de máscaras faciales públicas en comparación con el YOLOv4-tiny original. Se mejora el mAP(precisión media media) en un 4,56% y la velocidad alcanza los 92,81 FPS. El detector propuesto logra un mejor rendimiento de detección general en comparación con otros detectores como: YOLOv2, YOLOv3, YOLOv3-Tiny, YOLO4-Tiny, el modelo propuesto demostró la superioridad tanto con valor teórico como con significado práctico [9]. El sistema desarrollado también aportamayor flexibilidad a la aplicación de detección de mascarillas en hospitales, campus, comunidades, etc.

Seguidamente, la versión YOLOv5[5] es utilizado para entrenar tres clases diferentes al mismo tiempo: una persona con máscara, sin máscara y la identificación de una persona, en el cual se logra una precisión del 99,5%. Asimismo, utilizado para la detección de casos de seguridad el cual es otro de los objetos de interés en este estudio, FD-YOLOv5[26] integra un módulo de imagen basado en fuzzy, el cual que permite hacer diferencia entre varios tiposde cascos. Asimismo, en [27] se optimiza YOLOv5 para abordar problemas de baja precisión y poca robustez para objetos pequeños en entornos naturales complejos, y además la integración del módulo Ghost que exige menos parámetros y menos complejidad computacional para generar características. Posteriormente, se hace una comparación del rendimiento de los modelos YOLOv5, YOLOv6 y YOLOv7[28], se identificó que los modelos YOLOv6s y YOLOv7 funcionan mejor en condiciones de poca luz en comparación con el modelo YOLOv5s. Sin embargo, se observó que los tres modelos no han logradodiferenciar una gorra normal de un casco de seguridad. Tanto el modeloYOLOv5s como el YOLOv6s tienen dificultades para detectar la ausencia del casco de seguridad en una persona de piel oscura. El modelo YOLOv7 es capazde detectar la ausencia de cascos de seguridad en personas de piel oscura y encondiciones de poca luz, pero con bajos niveles de

confianza del 60% al 80%.

YOLOv7 ofrece una arquitectura de red más rápida y robusta que ofrece un enfoque de integración de características eficiente, rendimiento de reconocimiento de objetos más preciso, una función de pérdida estable y una asignación optimizada de eficiencia de entrenamiento de etiquetas y modelos. En consecuencia, en comparación con otros modelos DL, YOLOv7 utiliza hardware computacional mucho menos costoso. En conjuntos de datos pequeños, se puede entrenar significativamente más rápido sin el uso de pesos previamente entrenados. La arquitectura de Yolov7 (Ilustración 7), tiene con tres partes importantes: la columna vertebral (backbone), el cuello (neck) y la cabeza(head). La columna vertebral es responsable de la extracción de características de las imágenes de entrada dadas, el cuello genera principalmente las pirámides de características y la cabeza realiza la detección final como salida.

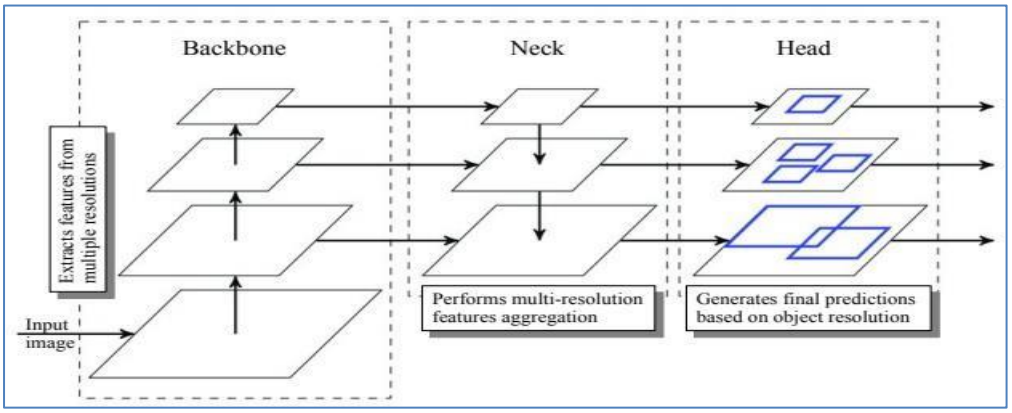

Ilustración 7 Arquitectura de Yolov7 Tomado de [29]

b) Método de dos etapas

Estos modelos intentan encontrar un número arbitrario de propuestas de objetosen una imagen durante la primera etapa y luego clasificarlos y localizarlos en lasegunda. Como estos sistemas tienen dos pasos separados, generalmente toman más tiempo para generar propuestas, tienen una arquitectura complicaday carecen de contexto. Entre estos

se encuentran entro otros los modelos:

- R-CNN The Region-based Convolutional Neural Network (R-CNN)
- Fast R-CNN
- Mask R-CNN

Las Redes Neuronales Convolucionales Basadas en Regiones (R-CNN) son lasprimeras CNN utilizadas en la detección de objetos. Los modelos R-CNNinicialmente proponen varias regiones adecuadas a partir de una imagen y luegolas pasan a través del modelo CNN de avance para extraer características relevantes, etiquetarlas en categorías y mezclarlas para crear cuadros delimitadores. La mayoría de los modelos R-CNN constan de los siguientes pasos [21]:

1. Se utiliza un algoritmo de búsqueda selectiva para proponer múltiples regiones de alta calidad en diferentes escalas, formas y tamaños a partir de la imagen de entrada.

2. La región propuesta se transforma en las dimensiones de entrada requeridas de un modelo de clasificación CNN entrenado.

3. Esta región transformada se clasifica en una categoría específica mediante un cálculo directo, extrayendo las características relevantes.

4. La optimización del cuadro delimitador final y la localización del objeto se calcula combinando las características y el cuadro delimitador etiquetado decada región propuesta.

El mayor inconveniente de estos modelos R-CNN es la baja velocidad debido ala gran cantidad de regiones propuestas, aunque se utiliza CNN entrenada para extraer las características relevantes. Todas estas regiones pasan a través de la red CNN, lo que requiere de cálculos y de recurso de máquina, lo que lo haceinviable para las aplicaciones reales de

detección de objetos en tiempo real.

- **Fast R-CNN** [30]: Es una red de propuesta de región entrenable basada en CNNen lugar de una búsqueda selectiva de localización precisa. La red utiliza una ventana n × n para deslizarse sobre la capa convolucional. Esta ventana deslizante lo asigna a un mapa de características de menor dimensión. Luego se alimenta a dos capas completamente conectadas para una regresión de caja y capas de clasificación. Para una mejor optimización algorítmica, se implementa como una capa convolucional n×n seguida de dos capas convolucionales 1×1. Además de la red de propuesta de región entrenable, Faster R-CNN incluye las predicciones de cuadro delimitador y la categoría en la detección de objetos.
- **Mask R-CNN** [31]: Es un modelo de segmentación utilizado para etiquetar cada píxel en una categoría específica. La precisión de la detección de objetos se puede mejorar aún más mediante el uso de datos de entrenamiento etiquetados para cada píxel de una imagen. Mask R-CNN es una extensión de Faster R- CNN con una rama adicional para predecir máscaras de segmentación en cadaregión de interés (Roí), uno de los principales cambios en el modelo Mask R- CNN fue reemplazar la capa de agrupación de Roí con una capa de alineación de Roí. Esto permite que la información espacial de los mapas de característicasse conserve en la operación de interpolación bilineal, lo que la hace más adecuada para las predicciones a nivel de píxel, ya que genera mapas de características con la misma forma para todas las Roí. Esto predice las categorías y los cuadros delimitadores de las Roí y las posiciones a nivel de píxel de los objetos con una red totalmente convolucional adicional.
- **Métodos multietapa:** Se centran en la estrategia de propuestas de regiones selectivas a través de una arquitectura muy compleja [32].

Metricas de rendimiento

Los detectores de objetos utilizan varios criterios para medir el rendimiento de los detectores, a saber, fotogramas por segundo (FPS), precisión, Recall (exhaustividad), F1, Accuracy y la matriz de confusión.

Precisión (Precision): La precisión es una métrica utilizada para evaluar la calidad de un modelo de aprendizaje automático en tareas de clasificación. Se calcula mediante la fórmula: VP / (VP + FP), donde VP representa los Verdaderos Positivos y FP los Falsos Positivos. La precisión mide la confiabilidad de las detecciones o la proporción de casos Verdaderos Positivos entre todos los casos positivos detectados por la prueba.

La precisión permite determinar qué porcentaje de las detecciones realizadas por el modelo son realmente correctas y tienen valor. En esta fórmula, TP representa los Verdaderos Positivos, que son los casos correctamente identificados como positivos, y FP son los Falsos Positivos, que son los casos incorrectamente identificados como positivos. [33].

Recall (exhaustividad): La sensibilidad es una métrica que evalúa la capacidad de un modelo para detectar correctamente los casos positivos pertenecientes a las clases del conjunto de datos analizado. También se conoce como tasa de verdaderos positivos y se define como: VP / (VP + FN). Esta métrica determina el porcentaje de objetos existentes que fueron detectados correctamente por el modelo, mientras que FN representa los falsos negativos, es decir, los casos positivos que no fueron identificados por el modelo.[33]

Accuracy (exactitud): Mide el porcentaje de casos acertados, a partir de laformula siguiente:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

F1score (Valor F): Permite combinar las medidas de exhaustividad y precisión en un solo valor. La puntuación F1 es un valor que se utiliza

para representar un equilibrio entre la tasa de precisión y la tasa de recuperación delmodelo.

Matriz de Confusión: Una matriz de confusión es una herramienta utilizada para evaluar el desempeño de un modelo, generalmente en tareas de clasificación. Esta matriz cuadrada permite comparar los valores predichos por el sistema con los valores reales. Proporciona una visión adicional para evaluar la efectividad de nuestro modelo y determinar su rendimiento.

	Real Values		
Predicted Values		Positive	Negative
	Positive	TP = True Positive	FP = False Negative
	Negative	FN = False negative	TN = True Negative

Ilustración 8 Matriz de Confusión. Tomado de [20]

Si analizamos con más detalle, podemos observar que la variable objetivo del modelo tiene dos posibles valores: positivo o negativo. En la matriz de confusión, los valores reales se representan en la columna, mientras que los valores predichos se encuentran en las filas. Tomemos como ejemplo un sistema que contiene 1000 datos y genera una matriz de confusión con los siguientes valores: TP (Verdaderos Positivos) = 600, TN (Verdaderos Negativos) = 300, FN (Falsos Negativos) = 50 y FP (Falsos Positivos) = 50.

En este caso, el sistema demuestra un rendimiento generalmente bueno con una tasa de precisión del 90%, ya que identifica correctamente 900 de los 1000 datos. Sin embargo, al analizar más a fondo la matriz de confusión, también podemos obtener información valiosa sobre los tipos de errores que comete el sistema.

Dependiendo de la aplicación en particular, algunos tipos de errores pueden tener más relevancia que otros, y la matriz de confusión permite identificarlos y evaluar su impacto en el rendimiento general del modelo.

AP: Si representamos en un gráfico tomando como ejes y = precisión y eje x = recall, AP es el área bajo la curva, es decir, bajo el gráfico generado por ambasmétricas. Esta dada por la siguiente ecuación:

$$AP = \int_0^1 p(r)dr$$

donde p(r) es la curva formada por el gráfico precisión-Recall

Intersección sobre Unión (IoU): Métrica de evaluación utilizada en los puntos de referencia de detección de objetos [34], la cual mide el área de intersección entre dos cuadros delimitadoras, para lo cual se establece un umbral para determinar si la detección es correcta. Si el IoU supera el umbral, se clasifica como verdadero positivo, mientras si está por debajo se clasifica como falso positivo. Si el modelo no logra detectar se denomina falso negativo. Esta métrica está dada por la siguiente ecuación.

$$\text{IoU} = \frac{\text{área de intersección}}{\text{área total de la unión}}$$

Mean Average precisión (mAP): Finalmente definimos mAP como la media deAP para un umbral de IoU dado. Un umbral clásico para IoU es [0.5, 0.05, 0.95],de esta forma mAP será la media de los valores AP para cada valor de IOU, queva desde 0.5 a 0.95 con un paso de 0.05. El mAP incorpora el compromiso entreprecisión y recuperación y considera tanto los falsos positivos (FP) como los falsos negativos (FN). Esta propiedad hace que mAP sea una métrica adecuada para la mayoría de las aplicaciones de detección.
Esta dada por la expresión:

$$\text{mAP} = \frac{1}{N}\sum_{i=1}^{N} APi.$$

1.7.2 Marco Conceptual

Dataset: Colección estructurada de información que se utiliza para realizar análisis investigaciones o entrenar modelos de aprendizaje

automático. Consiste en un conjunto de observaciones, ejemplos o instancias, donde cada instancia representa un registro de datos y está compuesta por características o atributos que describen dicha instancia.

Deep Learning: Es una rama de la inteligencia artificial que utiliza redes neuronales, algoritmos inspirados en el cerebro humano, para aprender de grandes conjuntos de datos. Estos algoritmos realizan tareas repetitivas y mejoran gradualmente los resultados a través de capas profundas de procesamiento, lo que permite un aprendizaje progresivo. El Deep Learning es parte de un conjunto más amplio de métodos de machine learning basados en redes neuronales.

Machine Learning: Es un campo de la inteligencia artificial que se centra en el desarrollo de algoritmos y modelos que permiten a las máquinas aprender y tomar decisiones basadas en datos, sin ser específicamente programadas para cada tarea. En lugar de seguir instrucciones explícitas, los sistemas de Machine Learning aprenden patrones y reglas a partir de ejemplos y experiencias previas. Estos algoritmos pueden realizar tareas como clasificación, regresión, clustering y reconocimiento de patrones, entre otros, y se aplican en diversas áreas como la visión por computadora, el procesamiento de lenguaje natural, la recomendación de productos y la predicción de resultados. El objetivo principal del Machine Learning es permitir que las máquinas aprendan de manera autónoma y mejoren su rendimiento a medida que se les presenta más información y experiencia. Esta tecnología está presente en un sinfín de aplicaciones como las recomendaciones de Netflix o Spotify, las respuestas inteligentes de Gmail o el habla de Siri y Alexa.

Métricas de Rendimiento: Son utilizadas para evaluar los modelos se encuentran la precisión, Recall (exhaustividad), Accuracy (exactitud), Matriz de Confusión, F1score (Valor F), AP, Intersección sobre Unión

(IoU) y mean average precision (mAP).

Modelo de Detección de Objetos: Es un algoritmo de soportado una Red Neuronal Convolucional que detecta una cantidad limitada (o específica) de objetos, no pudiendo detectar objetos que antes no hubiera visto, o si están en tamaños que logra discernir y todas las dificultades de posibles “focos”, rotacióndel objeto, sombras y poder determinar en qué posición dentro de la imagen se encuentra. Se dividen en 3 categorías: a) método de una etapa que son capacesde analizar una imagen con una sola valoración de red, estos se enfocan en todas las propuestas de regiones espaciales para la detección de objetos a través de una arquitectura relativamente más simple, b) método de dos etapas que requieren de requieren miles de evaluaciones para una sola imagen y c) métodos multietapa, los cuales se centran principalmente en la estrategia de propuestas de regiones selectivas a través de una arquitectura muy compleja.

Redes Neuronales Convolucionales (CNN): Es un tipo de red multicapa que consta de diversas capas convolucionales y de pooling (submuestreo) alternadas, y al final tiene una serie de capas conectadas completamente comouna red perceptrón multicapa.

Sistemas de gestión: Un sistema de gestión es una herramienta que permite controlar, planificar, organizar y automatizar las tareas administrativas de una organización. Un sistema de gestión analiza los rendimientos y los riesgos de una empresa, con el fin de otorgar un ambiente laboral más eficiente y sostenible.Algunas empresas o Pymes cuentan con actividades que no están automatizadas, que con frecuencia se soportan en sistemas departamentales y casi siempre en hojas Excel desarrolladas individualmente por los usuarios implicados en cada una de las funciones. Un software de gestión unifica la operación de todas las áreas del negocio para alinearlas con los objetivos de laempresa.

Seguridad y Salud en el Trabajo: El Ministerio del Trabajo comprometido con las políticas de protección de los trabajadores colombianos y en desarrollo de las normas y convenios internacionales, estableció el Sistema de Gestión de Seguridad y Salud en el Trabajo (SG-SST), el cual debe ser implementado por todos los empleadores y consiste en el desarrollo de un proceso lógico y por etapas, basado en la mejora continua, lo cual incluye la política, la organización, la planificación, la aplicación, la evaluación, la auditoría y las acciones de mejora con el objetivo de anticipar, reconocer, evaluar y controlar los riesgos que puedan afectar la seguridad y la salud en los espacios laborales.

El sistema de gestión aplica a todos los empleadores públicos y privados, los trabajadores dependientes e independientes, los trabajadores cooperados, los trabajadores en misión, los contratantes de personal bajo modalidad de contrato civil, comercial o administrativo, las organizaciones de economía solidaria y del sector cooperativo, las empresas de servicios temporales, las agremiaciones u asociaciones que afilian trabajadores independientes al Sistema de Seguridad Social Integral; las administradoras de riesgos laborales; la Policía Nacional en lo que corresponde a su personal no uniformado y al personal civil de las Fuerzas Militares.

Visión por computadora: Hace referencia a un grupo de tecnologías o herramientas que permiten a los equipos captar imágenes del mundo real, procesarlas y generar información a través de ellas (análisis). Dicho de otra manera, la visión por computador es una propiedad de ciertas tecnologías que permiten a los equipos computarizados.

YOLO (You Only Look Once): Es un modelo de detección de objetos de una etapa que hace predicciones en tiempo real y logrando una velocidad dedetección notable, la red YOLO divide una imagen en una cuadrícula de tamañoGxG, y cada cuadrícula genera N predicciones para cuadros delimitadores. Se caracteriza porque **c**ada cuadro delimitador está limitado a tener solo una clase durante la predicción,

lo que impide que la red encuentre objetos más pequeños. La última versión al momento de este trabajo es Yolov7.

1.7.3 Marco Legal

El presente proyecto tiene sus bases legales sobre los siguientes pilares denormas, decretos y leyes del estado colombiano:

- Resolución 2400 de 1979: Mediante el cual se crea el estatuto de seguridad industrial.
- Resolución 2013 de 1986: Creación y funcionamiento de comités paritarios de salud ocupacional.
- Decreto 614 de 1984: Creación de bases para la organización de la salud ocupacional.
- Resolución 2013 de 1986: Establece la creación y funcionamiento de los comités de medicina, higiene y seguridad industrial en las empresas.
- Resolución 1016 de 1989: Establece el funcionamiento de los programas de salud ocupacional en las empresas.
- Decreto 393 del 26 de febrero de 1991 por el cual se dictan normas sobre asociación para actividades científicas y tecnológicas, proyectos de investigación y creación de tecnologías.
- Decreto 1295 de 1994: Mediante el cual se determina la organización y administración del sistema general de riesgos profesionales.
- Decreto 1530 de 1996: se define accidente de trabajo y enfermedad profesional con muerte del trabajador.
- Ley 776 de 2002: Se dictan normas de organización, administración y prestación del sistema general de riesgos profesionales.
- Resolución 1401 de 2007: Reglamenta la investigación de accidente e incidente de trabajo.
- Resolución 2346 de 2007: Regula la práctica de evaluaciones médicas ocupacionales y el manejo y contenido de las historias clínicas ocupacionales.
- Resolución 1918 de 2009: Modifica los artículos 11 y 17 de la resolución

2346de 2007 y se dictan otras disposiciones.

- Resolución 1956 de 2008: Se adoptan medidas para el consumo de cigarrilloy tabaco.
- Resolución 2646 de 2008: Se establecen disposiciones y se definen responsabilidades para la identificación, evaluación, prevención, intervención y monitoreo permanente de la exposición a factores de riesgo psicosocial en el trabajo y para la determinación del origen de las patologías causadas por elestrés ocupacional.
- Decreto 2566 de 2009: Se emite la tabla de enfermedades profesionales.
- Resolución 652 de 2012: Se establecen conformación de comités de convivencia laboral para empresas públicas y privadas y se dictan otras disposiciones.
- Circular 0038 de 2010: espacio libre de humo y sustancias psicoactivas en la empresa.
- Ley 1562 de 2012 por la cual se modifica el sistema de riesgos laborales y sedictan otras disposiciones en materia de salud ocupacional.
- Resolución 1356 de 2012: Por medio de la cual se modifica parcialmente la resolución 652 de 2012.
- Resolución 1409 de 2012: Por la cual se establece el reglamento de seguridad para la protección en caídas en trabajos en alturas.
- Resolución 4502 de 2012: Por la cual se reglamenta el procedimiento, requisitos para el otorgamiento y renovación de las licencias de salud ocupacional y se dictan otras disposiciones.
- Resolución 1903 de 2013: Por la cual modifica el numeral 5° del artículo 10 yel parágrafo 4° del artículo 11 de la Resolución 1409 de 2012, por la cual se estableció el Reglamento para Trabajo Seguro en Alturas, y se dictan otras disposiciones.
- Resolución 3368 de 2014: Modificación al reglamento para protección contracaídas de trabajo en alturas.
- Decreto 1443 de 2014: Por medio del cual se dictan disposiciones para la implementación del sistema de gestión de la seguridad y salud en el trabajo (SG-SST).

- Decreto 1072 de 2015, por el cual establece una serie de directrices de cumplimiento obligatorio para llevar a cabo la implementación del SG-SST.
- Decreto 293 del 2017; Por el cual se reglamenta el artículo 7 de la Ley 1753 de 2015 en lo relacionado con los Planes y Acuerdos Estratégicos Departamentales en Ciencia, Tecnología e Innovación y se dictan otras.
- Resolución 0312 de 2019, por el cual se definen los estándares mínimos del sistema de gestión de la seguridad y salud en el trabajo.

La metodología para construir un entorno de aplicación de la visión por computador para a la detección de elementos de protección para la seguridad de salud y trabajo está compuesta por fases y actividades concretas que permiten dar respuesta a los objetivos específicos del proyecto como se relacionan a continuación.

1.8 Metodología del estudio

La metodología para el desarrollo de este estudio se está estructura por fases y se describe en la tabla 3, indicado las actividades y los productos que se generan en cada una de ellas.

Tabla 3 Fases para resolver el problema

Fases	Actividades	Producto
Definir la metodología para la construcción del modelo	• Modelar las fases • Colección de imágenes • Preprocesamiento de Imágenes • Etiquetar las imágenes en categorías. • Dividir el Dataset en entrenamiento y pruebas.	Dataset de imágenes balanceado para el entrenamiento

Selección entrenamiento del modelo utilizando el algoritmo Yolo para la detección elementos	• Seleccionar las versiones del Algoritmo Yolo • Configurar las épocas y lotes • Entrenar el modelo	Modelo de redes neuronales entrenado
Generación y comparación de las métricas de desempeño de los algoritmos entrenados	• Generación de métricas Interpretación de las métricas	Generación y comparación de Métricas

Fuente: elaboración propia

II FASES PARA LA CONSTRUCCIÓN DEL MODELO DEDETECCIÓN DE ELEMENTOS DE PROTECCIÓN PERSONAL

Este capítulo describe las fases en detalle utilizadas para la construcción del modelo de detección, la cual se logró identificar luego de la revisión documental realizada para construir los antecedentes permitió identificar, las aplicaciones de deep learning y Redes Neuronales Convoluciones CNN en la detección de objetos, asimismo se logró identificar lasarquitecturas, la clasificación de los modelos, cuáles son los algoritmos utilizados para el reconocimiento de rostro u objetos. Se selecciono el algoritmo YOLO en sus versiones 5 y 7 por ser los más extensamente probados.

2.1 Fases para la construcción del modelo de detección

Para la construcción del modelo se adoptaron las fases de identificadas en [35],en la ilustración 9. se muestran las fases utilizadas.

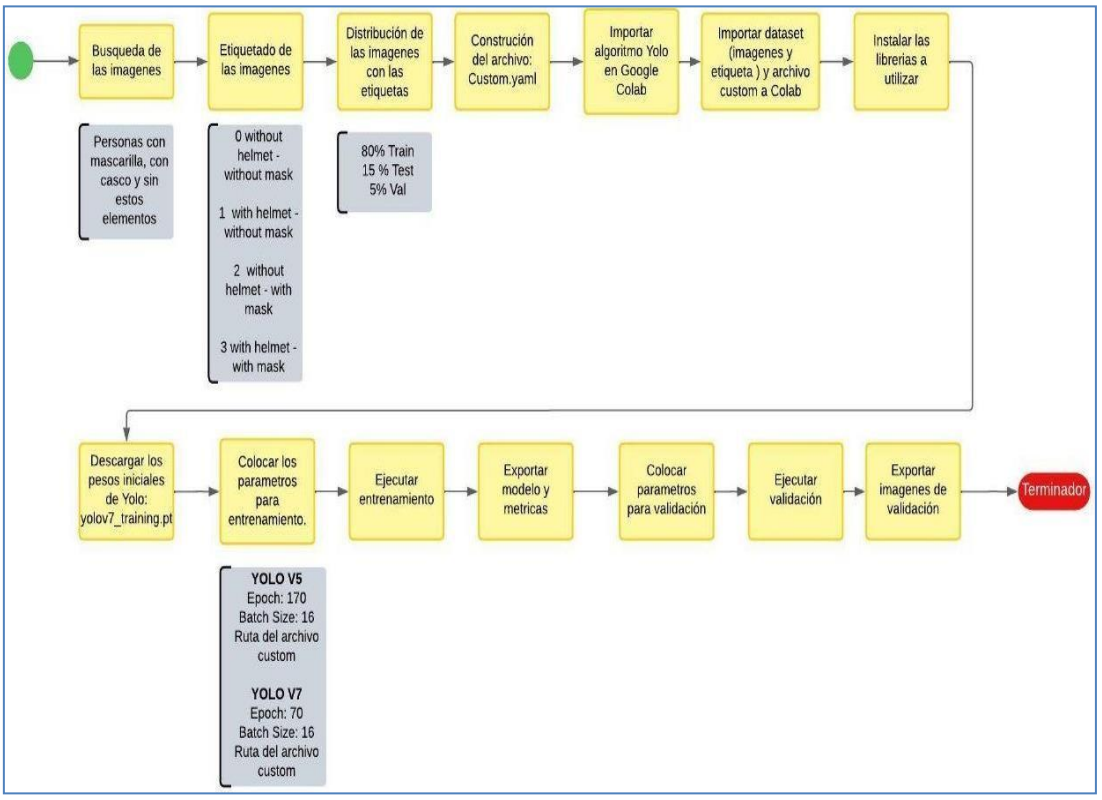

Ilustración 9 Diagrama de proceso para la construcción del modelo

2.2 Colección de datos (Búsqueda de imágenes): El conjunto de datos debe considerarse como una de las piezas importantes parala construcción de un modelo soportado por CNN, en este estudio el conjunto dedatos etiquetados fue construido por el autor agrupando 2.000 imágenes de personas con máscara y sin máscara y de personas con casco y sin casco, disponibles en internet de forma libre en https://www.kaggle.com/.

2.3 Etiquetado de imágenes*:* Antes de poder ingresar las imágenes al sistema, es necesario prepararlas adecuadamente. El primer paso consiste en normalizar el tamaño de las imágenes y los valores de los píxeles. Al normalizar el tamaño, se busca mejorar la eficiencia computacional y permitir que la red se adapte mejor a un tamaño de imagen específico en lugar de tamaños variables. Asimismo, se realiza la normalización de los píxeles para que sus valores estén en el rango de 0 a 1. Esta acción también contribuye a un entrenamiento más eficiente.

En el caso de la arquitectura YOLO, que es el enfoque que estamos analizando en este estudio, también se cargan los cuadros de anclaje junto con las imágenes y las transformaciones. Estos cuadros de anclaje consisten en anchos y alturas predefinidos seleccionados para que coincidan con los anchos y altos de los objetos presentes en el conjunto de datos.

La herramienta utilizada para la etiquetar las imágenes es https://www.makesense.ai/, la cual permite seleccionar el recuadro de la imagen que se desea detectar y etiquetar la (clase y coordenadas), para ello se identificaron las siguientes clases:

Tabla 4 Identificación de clases

INDEX	CLASE
0	without helmet - without mask
1	with helmet - without mask
2	without helmet - with mask
3	with helmet - with mask

Otro aspecto para tener en cuenta es que todas las imágenes deben tener el mismo tamaño de pixeles de alto y ancho y en los formatos JPG y PNG. El tamaño de imagen considerado es de 640 pp. En la figura 10 y 11 se muestra la imagen y su respectiva etiqueta la cual debe tener la siguiente estructura (<object-class-id> <x> <y> <width> <height>), la identificación de la clase, la posición x, y, el ancho y alto de la imagen respectivamente.

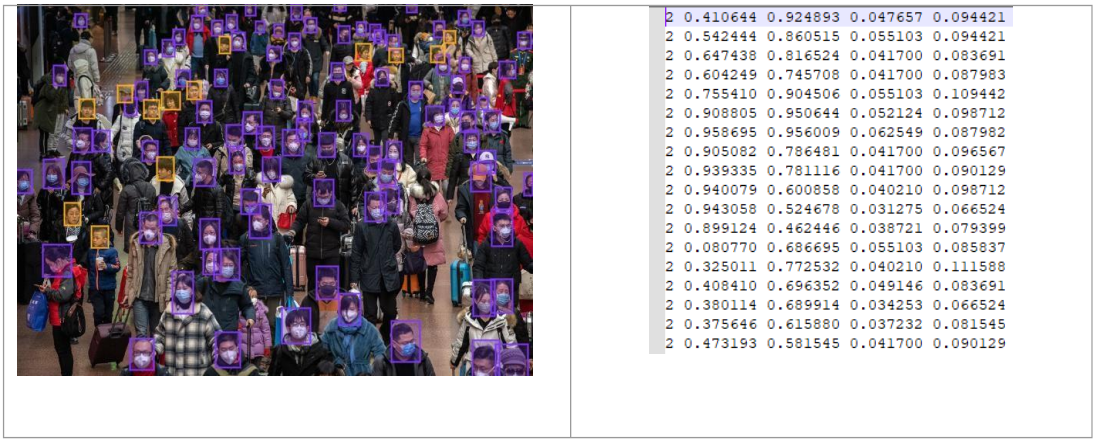

Ilustración 10 Imagen con, sin mascarilla y sin cascos y su etiqueta correspondiente.

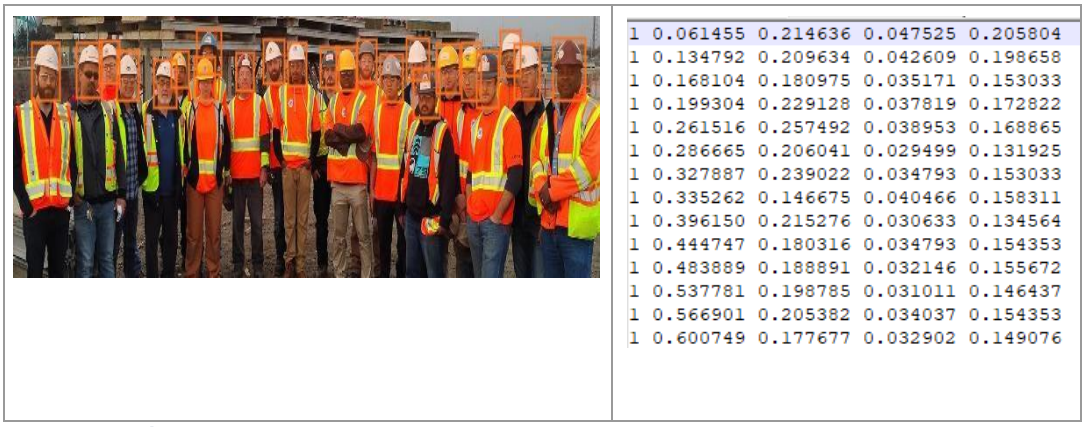

Ilustración 11 Imagen con casco y sin mascarilla su etiqueta correspondiente.

2.4. Distribución de Imágenes con etiquetas: Las imágenes etiquetadas fueron divididas en forma aleatoria con una proporcióndel 80% para el entrenamiento, 15% para pruebas y 5% para validación. La muestra de entrenamiento está conformada de la siguiente forma: imágenes de personas con casco y con mascarilla un 25%; sin casco y con mascarilla

un 25%;con casco y sin mascarilla un 25%; sin casco y sin mascarilla un 25%, lo cual sepuede afirmar que es una muestra balanceada. Esa misma proporción se utilizópara las pruebas y validaciones, tal como se muestra en la figura 9.

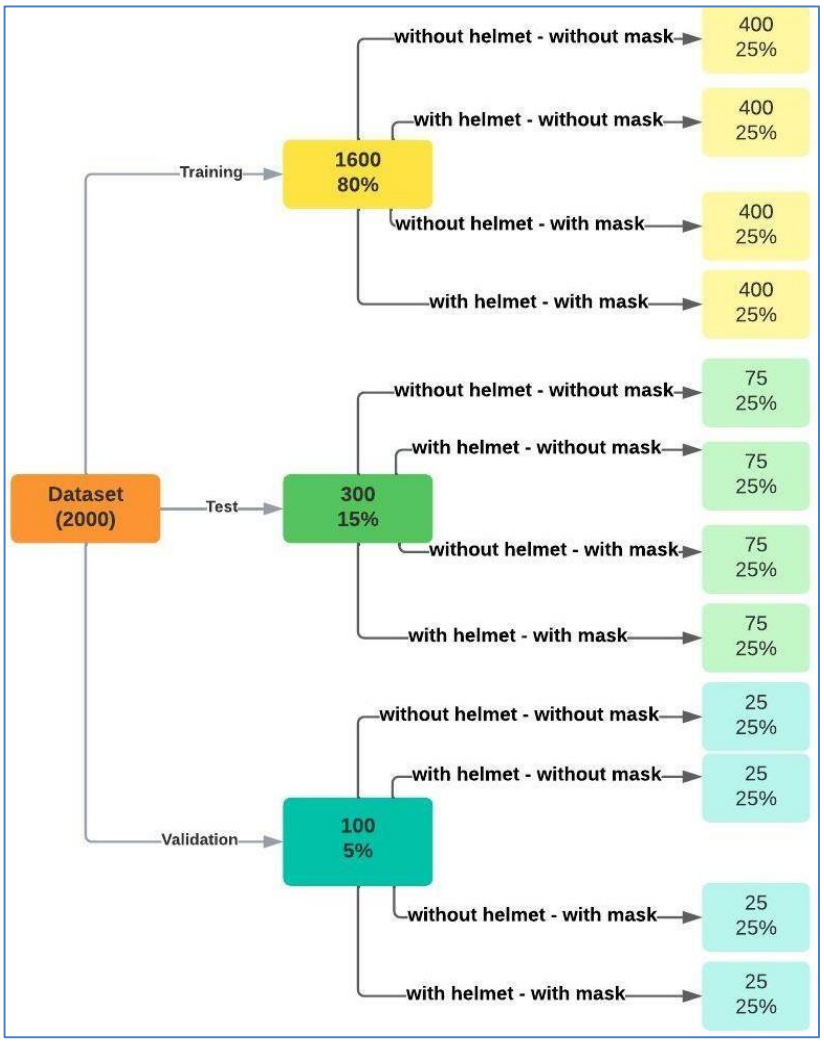

Ilustración 12 Distribución de los datos

III RESULTADOS

Este capítulo se explica el proceso de entrenamiento y sus parámetros de entrenamiento para el modelo de detección de imágenes de EPP, utilizando las versiones de yolov5 y yolov7 y además se describen las librerías utilizadas.

3.1 Construcción del archivo custom.yaml

El archivo con extensión .yaml se construyó con el fin de pasar las rutas en las que estaba el dataset de entrenamiento, pruebas y validaciones con sus respectivas etiquetas además allí se pasaron el número de clases que se iban autilizar para las predicciones con sus etiquetas.

```
train: /content/drive/MyDrive/data/images/train # train images
test:  /content/drive/MyDrive/data/images/test # test images
val: /content/drive/MyDrive/data/images/val  # val images

# Classes
nc: 4  # number of classes
names: ['without helmet - without mask', 'with helmet - without mask','without helmet - with mask', 'with helmet - with mask'] # class name
```

Ilustración 13 Archivo custom. yaml

3.2 Importar el algoritmo Yolo a Google Colab

Para el proceso de entrenamiento se utilizaron los siguientes recursos de Google Colab, adquiridos por el autor del trabajo:

- GPU de Colab con 13.7B en memoria
- tarjeta gráfica NVIDIA GeForce GTX 1080 TI
- Procesador Intel Core i5.

3.3 Importar dataset

El dataset luego de haber sido dividido entre entrenamiento, pruebas y validacióncon sus respectivas etiquetas estos fueron subido a Google drive y la ruta dondefue alojado se colocó en el archivo Custom.yaml. Para importar el dataset en el cuaderno de Google Colab se importó la librería de Google drive.

```
from google.colab import drive
drive.mount('/content/drive')
```

3.4 Instalar las librerías a utilizar

Se relacionan las librerías utilizadas y su descripción en el proceso de entrenamiento:

- **matplotlib>=3.2.2:** Se utiliza para la generación de los gráficos que fueron utilizados para visualizar las métricas y el comportamiento que fue teniendo el modelo a lo largo del entrenamiento, prueba y validación.
- **numpy>=1.18.5, <1.24.0:** Se utiliza para crear vectores y matrices multidimensionales.
- **opencv-python>=4.1.1:** Se requiere para implementar aplicaciones de visión por computadora en tiempo real.
- **Pillow>=7.1.2:** Permite abrir, manipular y guardar diversos formatos de archivo de imagen.
- **PyYAML>=5.3.1:** Es un analizador y emisor YAML recomendado para Python.
- **requests>=2.23.0:** Es un estándar que sirve para realizar solicitudes HTTP cuando se está desarrollando el lado del servidor de una página web.
- **scipy>=1.4.1:** Es una biblioteca de código abierto para Python que proporciona funcionalidades avanzadas de computación científica y análisis de datos. Incluye módulos para álgebra lineal, optimización, procesamiento de señales, estadísticas y más, siendo una herramienta fundamental para investigadores y científicos en diversas disciplinas.
- **torch>=1.7. 0,!=1.12.0:** Se utiliza paraaprendizaje automático.
- **torchvision>=0.8. 1,!=0.13.0:** Es una biblioteca de Python que extiende PyTorch, un popular marco de trabajo de aprendizaje profundo. Proporciona una amplia gama de herramientas y utilidades para cargar, preprocesar y visualizar conjuntos de datos de imágenes, así como modelos de visión por computadora preentrenados, facilitando el desarrollo de aplicaciones de visión artificial.
- **pandas>=1.1.4:** Es una biblioteca de Python que proporciona

estructuras de datos y herramientas para el análisis y manipulación de datos, especialmente útil en el ámbito de la ciencia de datos.

- **seaborn>=0.11.0:** Es una biblioteca de visualización de datos en Python que se basa en Matplotlib. Proporciona una interfaz de alto nivel para crear gráficos estadísticos atractivos y profesionales con menos líneas de código.

3.5 Descargar los pesos iniciales

Los algoritmos de YOLO cuentan pesos pre-entrenados lo que ayuda los tiempos de entrenamientos de nuevos modelos. Para descargar los pesos iniciales que se usaron en el entrenamiento para modelo se corrió la siguiente líneade código.

```
%cd /content/yolov7
!wget https://github.com/WongKinYiu/yolov7/releases/download/v0.1/yolov7_training.pt
```

Ilustración 14 Descarga de pesos preentrenados

3.6 Parámetros del entrenamiento Yolo

Para lograr un entrenamiento efectivo, es necesario especificar varios parámetros. En primer lugar, se definen las épocas, que indican cuántas veces se recorrerá todo el conjunto de datos durante el entrenamiento. El BatchSize determina cuántas imágenes se utilizan para actualizar los pesos en cada iteración. Por ejemplo, si el BatchSize es el 10% del número total de datos, se actualizarán los pesos 10 veces en una época.

Otro parámetro es el número de clases, que indica cuántas categorías distintas de objetos se están entrenando. La tasa de aprendizaje (learning rate) también debe ser especificada, ya que determina el tamaño de los pasos que se toman para llegar a una solución óptima. Este valor está relacionado con la función de coste, que mide el error entre el valor real y el valor predicho por el modelo. Durante el entrenamiento, se busca minimizar esta función de coste.

El optimizador es otro parámetro importante, ya que ajusta los parámetros del modelo para minimizar las pérdidas de la función de coste en el conjunto de entrenamiento. La función de pérdida nos permite evaluar el rendimiento del

algoritmo al procesar los datos y medir qué tan bien está funcionando. [20].

En esta tesis el proceso de entrenamiento fue relativamente sencillo, se consideraron 170 y 70 épocas y un tamaño de lote de 16, el optimizador seleccionado es Stochastic Gradient Descent (SGD), con logistic regression (Lr = 0.0), y con los parámetros 123 weight=decay (0,0) y 126 weight=decay (0.0005) y 126 Biases. La línea de código para entrenar el modelo se observa en la siguiente ilustración:

```
# ejecutar esta celda para comenzar el entrenamiento
%cd /content/yolov7
!python train.py --batch 16 --epochs 70 --data /content/drive/MyDrive/Metricas/custom.yaml --weights 'yolov7_training.pt'
```

Ilustración 15 Ejecutar entrenamiento

3.7 Metricas de desempeño YoloV5

La Tabla 5 muestra el tiempo de entrenamiento para YOLOv5 y YOLOv7. Se puede ver que YOLOv5 tiene un tiempo de entrenamiento más corto que YOLOv7, para la detección de las clases identificadas: sin mascarilla, sin casco, con mascarilla y con casco.

Tabla 5 Parámetros de configuración para el entrenamiento de los algoritmos.

Valores de hyperparametros	Yolov5	Yolov7
Epoch	170	70
Batch	16	16
Box Loss Gain	0.01951	0,02403
Class Loss Gain	0.000388	0.001092
Object Loss Gain	0.01857	0.009501
Learning rate	0.01	0.01
Warmup momentum	0.8	0.8
Warmup epochs	3.0	3.0
Optimizar weight decay	0.0005	0.0005
Warmup bias	0.1	0.1

Tabla 6 Resultados del entrenamiento de los algoritmos

Resultado del entrenamiento	Yolov5	Yolov7
Entrenamiento (HH:MM)	6:00	8:00
Tamaño del modelo (KB)	169,027	73046

3.8 Metricas de desempeño YoloV7

La ilustración 17 muestra la matriz de confusión de YOLOv7, donde se puede observar que el modelo entrenado es capaz de detectar imágenes without helmet - without mask con una puntuación de precisión de 88%, with helmet - without mask 95%, without helmet - with mask 88% y with helmet - with mask 96%.

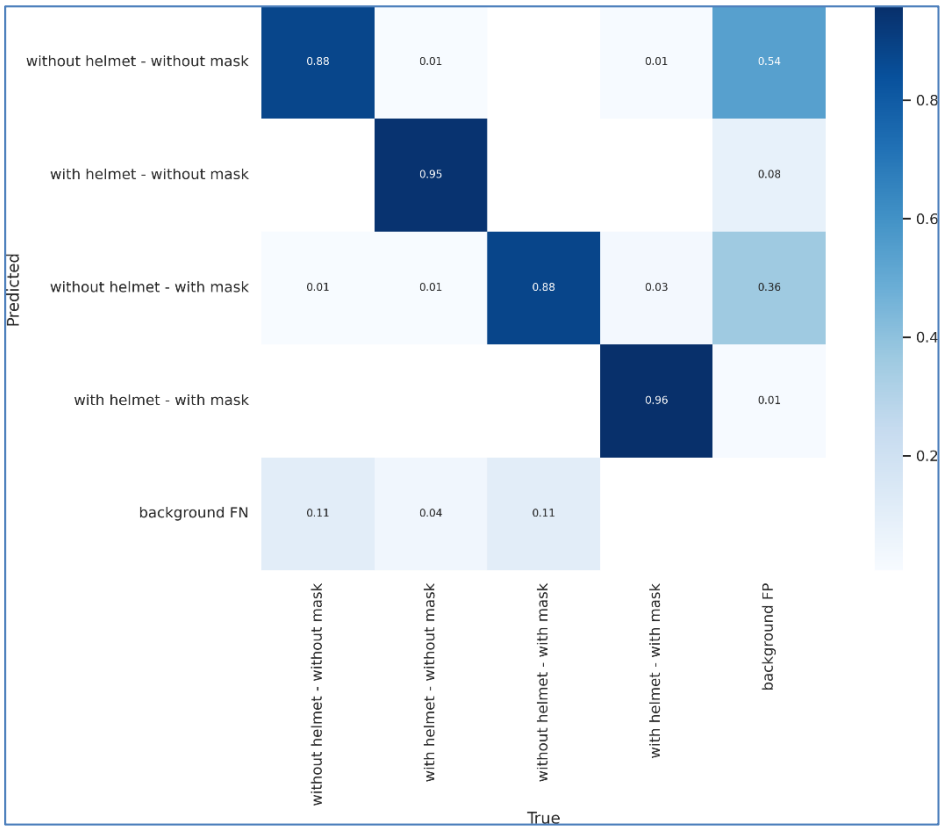

Ilustración 16 matriz de confusión del modelo YOLOv7

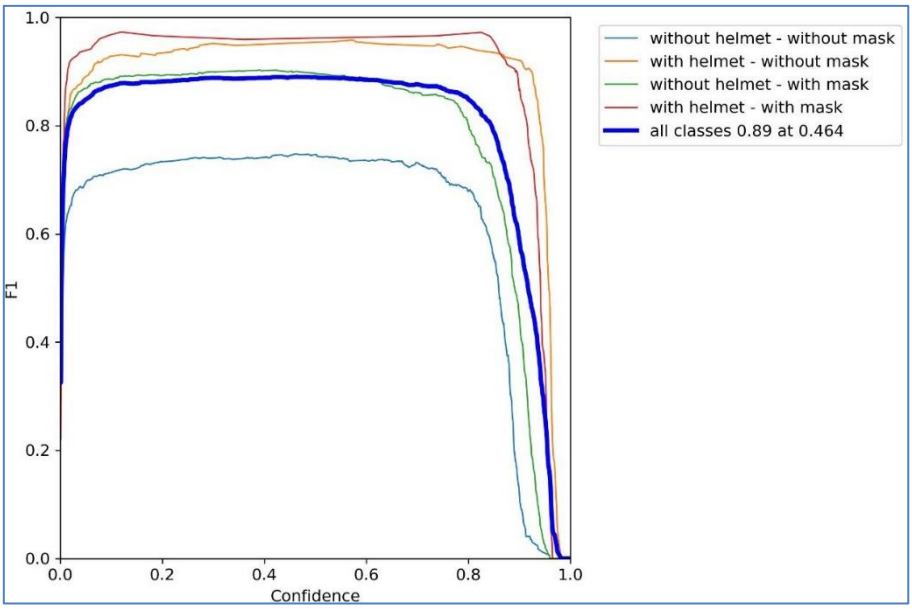

Ilustración 17 Curva de F1 contra la confianza Yolov7

Las métricas de desempeño para cada una de las clases se describen en la tabla 7.

Tabla 7 Métricas de desempeño Yolov7 para todas las clases

Clase	Precision	Recall	mAP@.5	Map50-95
without helmet - without mask	67%	84%	71%	39%
with helmet - without mask	96%	94%	97%	83%
without helmet - with mask	95%	85%	92%	59%
with helmet - with mask	95%	97%	99%	86%

3.9 Evaluación de desempeño YoloV5

En los modelos de detección de objetos, existen las siguientes pérdidas:

Pérdida de caja: Es una medida que evalúa la discrepancia entre las coordenadas de la caja predicha y las coordenadas de la caja real en problemas de detección de objetos, ayudando a mejorar la precisión de la localización.

Pérdida de clase: Es una métrica que mide la discrepancia entre la clase predicha y la clase real de un objeto en problemas de clasificación, permitiendo evaluar la precisión en la asignación de etiquetas a los objetos detectados.

Pérdida de objetos: Es una medida que cuantifica el error total en la detección

y clasificación de objetos en un modelo, combinando las pérdidas de caja y de clase para evaluar el rendimiento general del sistema de detección de objetos.

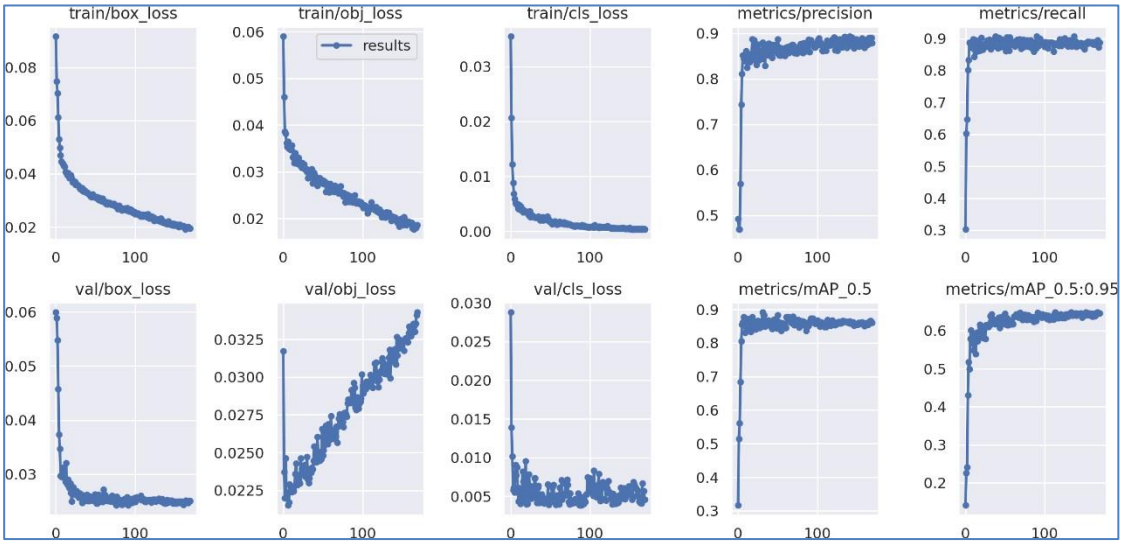

Ilustración 18 Pérdidas en el entrenamiento Yolo5

3.10 Evaluación de desempeño YoloV7

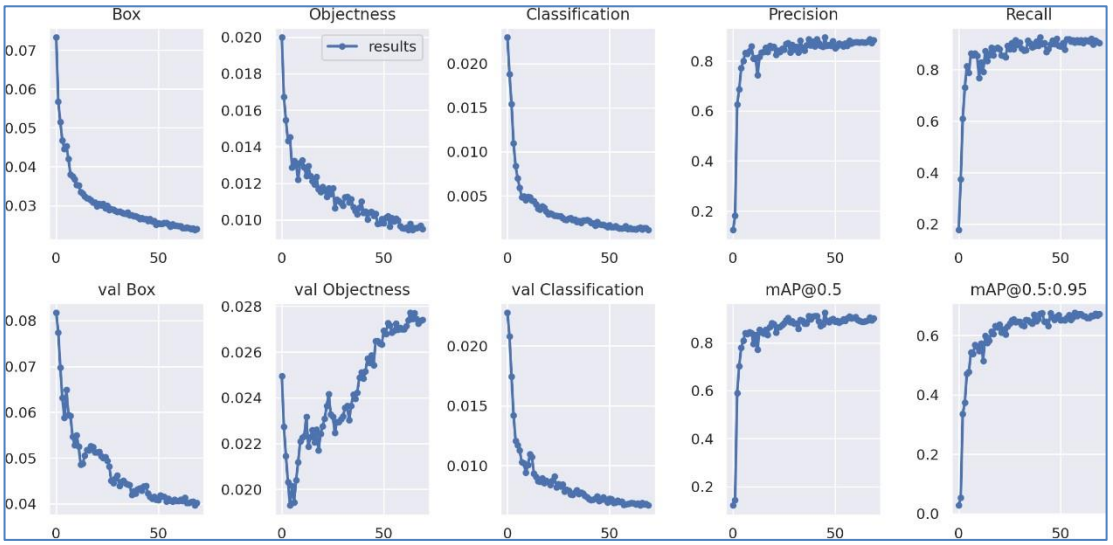

Ilustración 19 Pérdidas en el entrenamiento Yolo7

Luego de haber realizado el entrenamiento de los algoritmos de YOLOV5 y YOLOV7 se puede evidenciar que el primer algoritmo a pesar de que se usó un numero de 170 épocas siendo superior al otro tuvo una precisión del 87% y una duración del entrenamiento de 5.9 horas mientras que el algoritmo YOLOV7 con solo usar 50 épocas y una duración de entrenamiento de 8.4 con una precisión del 88%.

3.11 Validación

Se realizó una prueba de ambos modelos entrenados, pasando las imágenes destinadas para la validación, obteniendo los resultados que se observan en la ilustración 21 para Yolov5 y en la ilustración 22 la validación del modelo Yolov7 donde se nota una mejor detección de imágenes pequeñas.

Ilustración 20 Validación del modelo con Yolov5

Ilustración 21 Validación del modelo con Yolov7

Para realizar las métricas de la etapa de validación se hizo una matriz de confusión en donde se comparó las imágenes detectaba por el modelo con las

imágenes etiquetadas y se fue evaluando la precisión de cada clase en la imagen.

De acuerdo con la tabla 8 se puede concluir que la diagonal de la matriz de confusión tiene buenas predicciones ya que están por encima del 99% de precisión.

	0	1	2	3
0	99.4%	0	0.0060	0
1	0	100%	0	0
2	0	0	100%	0
3	0	0	0	100%

Tabla 8 Matriz de confusión en validación

IV DISCUSIÓN

En este capítulo se describe las comparaciones de las versiones de Yolo V5 y Yolo V7 en temas de arquitectura, en parámetros de configuración utilizados para su entrenamiento, además de los resultados arrojados en de cada uno de ellos en las métricas de precisión.

En primer lugar, este estudio manejó versiones superiores del algoritmo Yolo utilizadas en trabajos previos revisados como: [4], [5], [6], [7], [8] y [9]. Uno de los principales puntos de comparación entre YOLOv5 y YOLOv7 es su arquitectura. YOLOv5 utiliza una arquitectura basada en la red neuronal convolucional (CNN) de una sola etapa, que es más simple y rápida de entrenar en comparación con modelos más complejos como Faster R-CNN y Mask R-CNN. Por otro lado, YOLOv7 utiliza una arquitectura basada en la red neuronal convolucional de dos etapas, lo que permite una mayor precisión en la detección de objetos, pero con un costo de tiempo de procesamiento más alto (esto se puede validar con el entrenamiento realizado en este proyecto en donde con Yolo v5 se usó 150 épocas y tuvo una duración de 8 horas mientras que con Yolo v7se entrenó con 70 épocas y su duración de entrenamiento fue de 8 horas.).
Otro factor de comparación entre los dos modelos es su desempeño en términos de precisión y velocidad de detección de objetos en imágenes y videos. YOLOv5 se destaca por su velocidad de procesamiento rápido y su capacidad para manejar grandes cantidades de datos en tiempo real, lo que lo hace ideal para aplicaciones en tiempo real como la detección de objetos en vehículos autónomos y cámaras de seguridad. Por otro lado, YOLOv7 tiene una mayor precisión de detección de objetos en comparación con YOLOv5, lo que lo hace más adecuado para aplicaciones donde la precisión es crucial, como la detección de objetos en imágenes médicas.

En general, la elección entre YOLOv5 y YOLOv7 depende del contexto y los requisitos específicos del proyecto de visión por computadora en cuestión. Cada modelo tiene sus fortalezas y debilidades, y es importante considerar cuidadosamente las necesidades del proyecto antes de tomar una decisión sobre qué modelo utilizar.

	YOLO V5	YOLO V7
Fecha de lanzamiento	06/01/2020	06/07/2022
Estado de la comunidad	Activo	Activo
Tipo de modelo	Detección de objetos	Detección de objetos
Arquitectura	CNN, YOLO	YOLO, CNN
FPS	140	161
Marco utilizado	PyTorch	PyTorch

Tabla 9 Comparación de Yolo V5 y Yolo V7

Además, es importante tener en cuenta que cada nueva variante de YOLO puede presentar diferentes fortalezas y debilidades en términos de velocidad de procesamiento, eficiencia energética y capacidad de detección en diferentes entornos y condiciones de iluminación. Por lo tanto, es fundamental evaluar cuidadosamente cada variante de YOLO para determinar su idoneidad para un proyecto específico.

Otro factor para considerar es la disponibilidad de conjuntos de datos adecuados para entrenar y validar el modelo. Aunque los algoritmos de detección de objetosYOLO han demostrado un rendimiento impresionante en una variedad de aplicaciones, el rendimiento del modelo depende en gran medida de la calidad ycantidad de datos de entrenamiento disponibles. Por lo tanto, es importante seleccionar cuidadosamente los conjuntos de datos de entrenamiento para garantizar la eficacia del modelo.

En conclusión, cada nueva variante de YOLO representa un paso adelante en la detección de objetos, pero es importante evaluar cuidadosamente sus fortalezasy debilidades antes de seleccionar la variante adecuada para un proyecto específico. Al hacerlo, se pueden garantizar resultados precisos y confiables enuna variedad de aplicaciones de visión por computadora.

Con respecto a las métricas, en términos de MAP [6] y [9], YOLOv7 ha demostrado tener una mayor precisión promedio en comparación con YOLOv5. Esto se debe en parte a la arquitecturade dos etapas de YOLOv7, que permite una mayor precisión en la detección deobjetos, pero también puede afectar la velocidad de procesamiento del modelo.

Sin embargo, también es importante tener en cuenta que la elección entre YOLOv5 y YOLOv7 dependerá en gran medida del proyecto y sus requisitos específicos. Si la velocidad de procesamiento es una consideración importante, YOLOv5 puede ser una mejor opción, ya que se ha demostrado que es más rápido que YOLOv7 en muchos casos. Por otro lado, si la precisión de detección de objetos es la principal preocupación, YOLOv7 puede ser una mejor opción

Además, la elección del modelo dependerá de la cantidad y calidad de los datos de entrenamiento disponibles, así como del entorno y las condiciones de iluminación en las que se utilizará el modelo. Es importante evaluar cuidadosamente todas estas consideraciones antes de seleccionar un modelo de detección de objetos YOLO para garantizar que se logren los mejores resultados posibles en el proyecto.

V. CONCLUSIONES Y RECOMENDACIONES

Se logró utilizar una metodología estructurada por fases que permitió ir avanzando en cada una de las actividades exigidas para la construcción del dataset y el modelo de CNN utilizando las versiones del algoritmo Yolo. Un aporte significativo en esta etapa es la colección de 2000 imágenes etiquetadas con las clases identificadas. 2. Para el entrenamiento del modelo se logró configurar los parámetros de configuración de ambas versiones haciendo uso de recursos de GPU pagos de Google Colab. 3. En cuanto a la comparación de las métricas obtenidas, se demostró que el modelo detección YOLOv7 supera al algoritmo YOLOv5, se concluye que con los parámetros iniciales pero distintosa cada uno con los que se hicieron el entrenamiento se puede evidenciar el porcentaje tan parecido de precisión que cuentan los algoritmos, sin embargo, la versión de Yolo7 tiene mejores precisiones y métricas a comparación de YOLOV5, aunque el tiempo de entrenamiento es mayor. También se identificó que el valor de la época varía en el entrenamiento para producir los mejores resultados. Los resultados dependen en gran medida del conjunto de datos de las imágenes.

Para trabajos futuros, se realizarán modificaciones y mejoras en el conjunto de datos adicionando más imágenes de entrenamiento para la clase sin caso y sin mascarilla cuyas métricas resultaron más bajas en ambos experimentos, asimismo se adicionarán clases de otros objetos de protección personal como: con y sin chalecos, guantes y botas, igualmente se pretende en un futuro la aplicación del modelo en un entorno real de tal forma que se puede comprar su rendimiento, y de esta forma completar la solución y se pueda incorporar a la gestión integral de la salud y seguridad en el trabajo en las empresas que es el fin principal de este trabajo.

REFERENCIAS

[1] H. Li, J. Qiu, K. Yu, K. Yan, Q. Li, and Y. Yang, “Fast safety distance warning framework for proximity detection based on oriented object detection and pinhole model,” *Measurement*, vol. 209, no. March 2022, p. 112509, 2023, doi: 10.1016/j.measurement.2023.112509.

[2] W. Chern, J. Hyeon, T. V Nguyen, V. K. Asari, and H. Kim, “Automation in Construction Context-aware safety assessment system for far-field monitoring,” *Autom. Constr.*, vol. 149, no. July 2022, p. 104779, 2023, doi: 10.1016/j.autcon.2023.104779.

[3] CONPES, “Politica Nacional Para La Transformacion Digital e Inteligencia Artificial,” 2019. [Online]. Available: https://colaboracion.dnp.gov.co/CDT/Conpes/Económicos/3975.pdf

[4] M. V. Ramadhan *et al.*, “Comparative analysis of deep learning modelsfor detecting face Comparative analysis of deep learning models for detecting face mask,” in *7th International Conference on Computer Science and Computational Intelligence*, 2023, vol. 00, no. 2022. doi: 10.1016/j.procs.2022.12.110.

[5] M. M. A. Joy *et al.*, “Contactless Surveillance for Preventing Wind-Borne Disease using Deep Learning Approach,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 11, pp. 775–783, 2022, doi: 10.14569/IJACSA.2022.0131190.

[6] G. Han, M. Zhu, X. Zhao, and H. Gao, “Method based on the cross-layer attention mechanism and multiscale perception for safety helmet-wearing detection,” *Comput. Electr. Eng.*, vol. 95, no. March, p. 107458, 2021, doi: 10.1016/j.compeleceng.2021.107458.

[7] T. Kumar, R. Rajmohan, M. Pavithra, S. A. Ajagbe, R. Hodhod, and T. Gaber, “Automatic Face Mask Detection System in Public Transportation in Smart Cities Using IoT and Deep Learning,” *Electron.*, vol. 11, no. 6, 2022, doi: 10.3390/electronics11060904.

[8] R. Shaban, A. Kurnaz, and H. Farhan, “Optimized face detector-based intelligent face mask detection model in IoT using deep learning approach,” *Appl. Soft Comput.*, vol. 134, p. 109933, 2023, doi: 10.1016/j.asoc.2022.109933.

[9] Z. Han, H. Huang, Q. Fan, Y. Li, Y. Li, and X. Chen, “SMD-YOLO: An efficient and lightweight detection method for mask wearing status during the COVID-19 pandemic,” *Comput. Methods Programs Biomed.*, vol. 221, p. 106888, 2022, doi: 10.1016/j.cmpb.2022.106888.

[10] P. Nagrath, R. Jain, A. Madan, R. Arora, P. Kataria, and J. Hemanth, “SSDMNV2: A real time DNN-based face mask detection system using single shot multibox detector and MobileNetV2,” *Sustain. Cities Soc.*, vol. 66, no. December 2020, 2021, doi: 10.1016/j.scs.2020.102692.

[11] M. Ghazal, Y. Abu Haeyeh, A. Abed, and S. Ghazal, “Embedded fatigue detection using convolutional neural networks with mobile integration,” *Proc. - 2018 IEEE 6th Int. Conf. Futur. Internet Things Cloud Work. W- FiCloud 2018*, pp. 129–133, 2018, doi: 10.1109/W-FiCloud.2018.00026.

[12] H. Vin, J. Huang, C. Chow, and Y. Chang, “Engineering Applications of Artificial Intelligence Detecting and recognizing driver distraction through various data modality using machine learning: A review , recent advances, simplified framework and open challenges (2014 – 2021) ☆,” *Eng.Appl. Artif. Intell.*, vol. 115, no. July, p. 105309, 2022, doi: 10.1016/j.engappai.2022.105309.

[13] A. Tonge, “Traffic Rules Violation Detection using Deep Learning,” pp.1250–1257, 2020.

[14] P. Yadax, N. Gupta, and P. Kumar, “A comprehensive study towards high-level approaches for weapon detection using classical machinelearning and deep learning methods,” *Expert Syst. Appl.*, 2023.

[15] P. Singh, M. S. Suryawanshi, and D. Tak, “Smart Fleet Management System Using IoT, Computer Vision, Cloud Computing and Machine Learning Technologies,” *2019 IEEE 5th Int. Conf. Converg. Technol. I2CT 2019*, pp. 1–8, 2019, doi: 10.1109/I2CT45611.2019.9033578.

[16] R. Khallaf and M. Khallaf, “Automation in Construction Classification and analysis of deep learning applications in construction: A

systematic literature review," *Autom. Constr.*, vol. 129, no. May, p. 103760, 2021, doi:10.1016/j.autcon.2021.103760.

[17] R. Duan, H. Deng, M. Tian, Y. Deng, and J. Lin, "Automation in Construction SODA: A large-scale open site object detection dataset for deep learning in construction," *Autom. Constr.*, vol. 142, no. July, p. 104499, 2022, doi: 10.1016/j.autcon.2022.104499

[18] S. S. A. Zaidi, M. S. Ansari, A. Aslam, N. Kanwal, M. Asghar, and B. Lee,"A survey of modern deep learning-based object detection models," *Digital Signal Processing: A Review Journal*, vol. 126. 2022. doi: 10.1016/j.dsp.2022.103514.

[19] M. Boukabous and M. Azizi, "Image and video-based crime prediction using object detection and deep learning," *Bull. Electr. Eng. Inform.*, vol.12, no. 3, pp. 1630–1638, 2023, doi: 10.11591/eei. v12i3.5157.

[20] S. Rozada and I. Jiménez, "Estudio de la arquitectura YOLO para la detección de objetos mediante deep learning," 2021. [Online]. Available: https://uvadoc.uva.es/bitstream/handle/10324/45359/TFM-G1316.pdf;jsessionid=B24BF9563F5F12CFCC850463C05EA3BB?sequence=1

[21] H. M. Ahmad and A. Rahimi, "Deep learning methods for object detectionin smart manufacturing: A survey," *J. Manuf. Syst.*, vol. 64, no. April, pp. 181–196, 2022, doi: 10.1016/j.jmsy.2022.06.011.

[22] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once:Unified, real-time object detection," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-Decem, pp. 779–788, 2016, doi: 10.1109/CVPR.2016.91.

[23] S. Sethi, M. Kathuria, and T. Kaushik, "Face mask detection using deep learning: An approach to reduce risk of Coronavirus spread," *J. Biomed.Inform.*, vol. 120, no. June, 2021, doi: 10.1016/j.jbi.2021.103848.

[24] N. K. Anushkannan, V. R. Kumbhar, S. K. Maddila, C. S. Kolli, B. Vidhya, and R. G. Vidhya, "YOLO Algorithm for Helmet Detection in Industries forSafety Purpose," *3rd Int. Conf. Smart Electron. Commun. ICOSEC 2022 - Proc.*, no. Icosec, pp. 225–230, 2022, doi:

10.1109/ICOSEC54921.2022.9952154.

[25] G. Gallo, F. DI Rienzo, P. Ducange, V. Ferrari, A. Tognetti, and C. Vallati, “A Smart System for Personal Protective Equipment Detection in Industrial Environments Based on Deep Learning,” *Proc. - 2021 IEEE Int. Conf. Smart Comput. SMARTCOMP 2021*, pp. 222–227, 2021, doi: 10.1109/SMARTCOMP52413.2021.00051.

[26] M. Sadiq, S. Masood, and O. Pal, “FD-YOLOv5: A Fuzzy Image Enhancement Based Robust Object Detection Model for Safety Helmet Detection,” *Int. J. Fuzzy Syst.*, vol. 24, no. 5, pp. 2600–2616, 2022, doi: 10.1007/s40815-022-01267-2.

[27] L. Zhao, D. Zhang, Y. Liu, J. Guo, and Z. Shi, “Improved YOLOv5s Network for Multi-scale safety Helmet Detection,” *11th Int. Conf. Commun. Circuits Syst. ICCCAS 2022*, pp. 262–266, 2022, doi: 10.1109/ICCCAS55266.2022.9825037.

[28] F. H. Juwono and Z. A. Sim, “Safety Helmet Detection Using Deep Learning: Implementation and Comparative Study Using YOLOv5, YOLOv6 and,” *IEEE Explor.*, pp. 164–170, 2022.

[29] I. Gallo, A. U. Rehman, R. H. Dehkordi, N. Landro, R. La Grassa, and M.Boschetti, “Deep Object Detection of Crop Weeds: Performance of YOLOv7 on a Real Case Dataset from UAV Images,” *Remote Sens.*, vol. 15, no. 2, pp. 1–17, 2023, doi: 10.3390/rs15020539.

[30] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN : Towards Real- Time Object Detection with Region Proposal Networks,” vol. 39, no. 6, pp.1137–1149, 2017.

[31] P. Doll, R. Girshick, and F. Ai, “Mask R-CNN ar,” in *IEEE International Conference on Computer Vision*, 2017. doi: 10.1109/ICCV.2017.322.

[32] M. Guerrieri and G. Parla, “Real-time social distance measurement and face mask detection in public transportation systems during the COVID- 19 pandemic and post-pandemic Era: Theoretical approach and case study in Italy,” *Transp. Res. Interdiscip. Perspect.*, vol. 16, no. June, 2022,doi: 10.1016/j.trip.2022.100693.

[33] M. Massiris, C. Delrieux, and J. Á. Fernández, “Detección de equipos de protección personal mediante red neuronal convolucional YOLO,” in *Actas de las XXXIX Jornadas de Automática, Badajoz*, 2020, pp. 1022– 1029. doi: 10.17979/spudc.9788497497565.1022.

[34] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese,“Generalized intersection over union: A metric and a loss for bounding box regression,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2019-June, pp. 658–666, 2019, doi: 10.1109/CVPR.2019.00075.

[35] K. Saini, S. Bharadwaj, and V. Gupta, “Face Mask Detection: A Deep Learning Concept,” *Proc. 3rd Int. Conf. Intell. Eng. Manag.ICIEM 2022*, pp. 437–441, 2022, doi: 10.1109/ICIEM54221.2022.9853014