UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Escola Tècnica Superior d'Enginyeria Agronòmica i del Medi Natural

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Escuela Técnica Superior de Ingeniería Agronómica y del Medio Natural

Desarrollo de un modelo de inteligencia artificial basado en Machine Learning para optimizar la selección embrionaria en reproducción humana asistida.

Trabajo Fin de Grado

Grado en Biotecnología

AUTOR/A: Salvo Jiménez, Aitana Eugenia

Tutor/a: Marco Jiménez, Francisco

Cotutor/a externo: CUEVAS SAIZ, IRENE

CURSO ACADÉMICO: 2022/2023

Desarrollo de un modelo de inteligencia artificial basado en *Machine Learning* para optimizar la selección embrionaria en reproducción humana asistida.

RESUMEN.

En las clínicas de reproducción asistida, tanto en el ámbito humano como animal, uno de los pasos fundamentales es la selección del embrión cultivado de mayor calidad para su posterior transferencia al útero de la futura gestante. La elección se basa en la interpretación de la calidad que el embriólogo considere al visualizar imágenes del embrión; basándose principalmente en que posea una morfología correcta. Es un método demostrado como subjetivo, variable y poco reproducible; por tanto, en este área se está indagando en la búsqueda de alternativas que sean técnicas no invasivas para el embrión y que proporcionen buenos resultados con el nacimiento de individuos sanos. Este es el objetivo del presente trabajo: que a través de la información recogida mediante la segmentación manual y etiquetado de imágenes de embriones en estadio de *hatching* (etapa posterior al blastocisto donde el embrión eclosiona y sale de la zona pelúcida para implantarse en el endometrio), se ha desarrollado un modelo de inteligencia artificial basado en un enfoque de *Machine Learning* supervisado capaz de identificar los embriones de mejor calidad según los criterios de clasificación embrionaria de ASEBIR.

La segmentación manual, una herramienta del procesado de imágenes, se ha realizado mediante la delimitación de los componentes más relevantes del blastocisto: células del trofoectodermo, masa celular interna y la zona pelúcida, entre otros; además, las etiquetas seleccionadas estaban referidas a la calidad del embrión fotografiado y a su estadio de desarrollo. Con esta información previa se ha entrenado y evaluado el modelo y finalmente, se ha testeado con un 10% de imágenes nunca vistas por el sistema para comprobar su funcionamiento. De esta forma, la inteligencia artificial ha podido detectar todos los componentes del embrión, reconocer el estadio de desarrollo en el que se encuentra y asignarle una calificación con respecto a su calidad con una sensibilidad de 0,64 , una precisión del 0,75 y un rendimiento reflejado como el área bajo la curva ROC del 0,79.

Estos resultados contribuyen a la finalidad de que el modelo pueda ser empleado en los laboratorios y clínicas de medicina reproductiva para la selección del embrión óptimo para transferir al útero de la mujer y reducir la carga subjetiva que supone la elección por parte del embriólogo en cuanto a su criterio. Por lo que este trabajo se relaciona con los siguientes ODS de la Agenda 2030: ODS 3. Salud y Bienestar y ODS 10. Reducción de las Desigualdades.

PALABRAS CLAVE: Fecundación *In Vitro*, blastocisto, *hatching*, Inteligencia Artificial, *Machine Learning*.

Autora: Aitana Eugenia Salvo Jiménez.
Tutor académico: Francisco Marco Jiménez.
Tutora externa: Irene Cuevas Saiz.
Grado: Biotecnología.

Valencia, Junio 2023.

Development of an artificial intelligence model based on Machine Learning to optimize embryo selection in assisted human reproduction.

ABSTRACT.

In assisted reproduction clinics, both in the human and animal field, one of the crucial steps is the selection of the highest-quality cultured embryo for subsequent transfer to the uterus of the intended gestational individual. The choice relies on the interpretation of the quality that the embryologist considers upon visualizing embryo images, mainly based on correct morphology. This method has been shown to be subjective, variable, and poorly reproducible. Therefore, in this field, efforts are being made to explore alternative non-invasive techniques for embryos that yield good results in terms of live births. This is the objective of the present study: where through the collected information from manual segmentation and labeling of hatching stage embryo images (the stage after blastocyst where the embryo hatches and exits the zona pellucida to implant in the endometrium), a supervised Machine Learning-based artificial intelligence model has been developed to identify embryos of better quality according to the ASEBIR embryo classification criteria.

Manual segmentation, an image processing tool, was performed by delineating the most relevant components of the blastocyst: trophectoderm cells, inner cell mass, and the zona pellucida, among others. Additionally, the selected labels referred to the quality of the photographed embryo and its developmental stage. With this prior information, the model was trained and evaluated, and ultimately tested with 10% of images never seen by the system to verify its performance. In this way, artificial intelligence has been able to detect all embryo components, recognize the developmental stage, and assign a quality rating with a sensitivity of 0.64, precision of 0.75, and performance reflected by the area under the ROC curve of 0.79.

These results contribute to the aim that the model can be used in reproductive medicine laboratories and clinics for the selection of the optimal embryo for transfer to the woman's uterus and reduce the subjective burden of the embryologist's choice of criteria. Therefore, this work is related to the following SDGs of the 2030 Agenda: SDG 3. Health and Well-being and SDG 10. Reduction of Inequalities.

Keywords: *In Vitro* Fertilization, blastocyst, hatching, Artificial Intelligence, Machine Learning.

Author: Aitana Eugenia Salvo Jiménez.
Academic Tutor: Francisco Marco Jiménez.
External Tutor: Irene Cuevas Saiz.
Degree: Biotechnology.

Valencia, June 2023.

AGRADECIMIENTOS.

"Me gustaría expresar mi más sincero agradecimiento a mi directora del proyecto Irene Cuevas Saiz que no solo me ha orientado y guiado durante el desarrollo del trabajo, sino que también me ha transmitido su pasión por el campo de la medicina reproductiva.

No puedo dejar de mencionar a mis familiares, pareja y amigos, quienes me han apoyado incondicionalmente a lo largo de mi carrera universitaria."

ÍNDICE.

GENERAL

FIGURAS

TABLAS

ABREVIATURAS.

Abreviatura	Aclaración
TRA	Técnicas de Reproducción Asistida
FIV	Fecundación *In Vitro*
ICSI	Inyección Intracitoplasmática
ASEBIR	Asociación para el Estudio de la Biología de la Reproducción
ZP	Zona Pelúcida
MCI	Masa Celular Interna
TF	Trofoectodermo
IA	*Artificial Intelligence*
PGT-A	Test Genético Preimplantacional de Aneuploidías
TL	*Time Lapse*
HGUV	Hospital General Universitario de Valencia
ML	*Machine Learning*
DL	*Deep Learning*
IoU	*Intersection over Union*
TP	*True Positive*
TN	*True Negative*
FP	*False Positive*
FN	*False Negative*
PPV	*Positive Predictive Value*
NPV	*Negative Predictive Value*
AUC	*Area Under Curve*
ROC	*Receiver Operating Characteristic*
AP	*Average Precision*

1.INTRODUCCIÓN.

Las Técnicas de Reproducción Asistida (TRA) son los procedimientos más efectivos para el tratamiento de la infertilidad y esterilidad humana (Farias et al., 2023). En el uso médico, la esterilidad generalmente se refiere a la incapacidad de originar un recién nacido vivo, mientras que la infertilidad se mide por la ausencia de gestación después de 1 año sin anticoncepción (Vander Borght & Wyns, 2018). En el uso demográfico, la esterilidad se refiere a la incapacidad no alcanzada quirúrgicamente para procrear, mientras que la infertilidad se refiere al rendimiento reproductivo en lugar de a la capacidad; es decir, la mujer es capaz de concebir, pero el embarazo no llega a término (Rochon, 1986).

Las TRA abarcan múltiples procedimientos entre los cuales se encuentra la Inseminación Artificial, basada en la disposición en el útero de la muestra de espermatozoides del varón para que la unión entre el óvulo y el espermatozoide tenga lugar de manera natural (Ombelet et al., 2003). Otra de las técnicas de empleo clínico rutinario, es la Fecundación *In Vitro* (FIV). La FIV se basa en la unión de gametos, pero a diferencia de la técnica anterior, se lleva a cabo *in vitro*, obteniendo un embrión que será transferido a la mujer (Edwards et al., 1981).

Con el avance e innovación de las TRA surgió la Inyección Intracitoplasmática de espermatozoides (ICSI) que consiste en la inyección directa de un único espermatozoide en el ooplasma del ovocito (PALERMO, 1992). De esta fecundación que no sucede de forma natural, sino que es generada por el embriólogo, se desarrolla un embrión que será cultivado y transferido al endometrio de la mujer.

Entonces, en los procedimientos donde la unión entre gametos se lleva a cabo fuera del cuerpo humano, se precisa inicialmente de la estimulación folicular de la mujer para que produzca múltiples óvulos o del empleo de donantes de este gameto; al igual que sucede con el individuo masculino, ya que los espermatozoides pueden provenir del cónyuge o de donante. A diferencia de la Inseminación Artificial, donde los gametos masculinos pueden provenir del marido o de donante, pero el gameto femenino pertenece a la mujer y futura gestante.

Posteriormente a la realización de la técnica, los embriones obtenidos serán cultivados, produciéndose su embriogénesis. El proceso de desarrollo comienza tras la fecundación, en la que los embriones experimentan divisiones de escisión que conllevan el paso de 2 células hasta alcanzar las 8; en esta etapa es donde el genoma embrionario se activa. En las divisiones consecutivas, se alcanza el estado de 16 células conocido como mórula. Las células internas de la mórula constituyen la masa celular interna y las células que rodean se denominan trofoectodermo; todo ello rodeado por la zona pelúcida (ZP), una fina capa glicoproteica que circunda exteriormente. La masa celular interna (MCI), originará los tejidos propios del embrión, mientras que la externa o trofoectodermo contribuirá a la formación de la placenta y otras estructuras extraembrionarias como el saco vitelino. Aproximadamente después de 4 días desde la fecundación, sucede la cavitación, proceso mediante el cual entra agua con iones sodio y se genera una cavidad interna llamada blastocele. Una vez generada la cavidad blastocística se denomina al embrión como blastocisto aproximadamente en el día 5 o 6 tras la fecundación. El blastocisto se caracteriza por continuar envuelto por la ZP, por poseer una masa de células central o MCI y unas células (blastómeros) en la parte externa, el trofoectodermo (TF). La siguiente etapa en la embriogénesis es la eclosión o *hatching*, en la que el blastocisto degrada la ZP y comienza a protruir para poder abandonarla e implantarse en el útero. Al blastocisto que ha conseguido salir de la ZP y que ya no se encuentra envuelto por esta, se denomina eclosionado o *hatched*. (Carlson, 2014; Shahbazi, 2020).

En las TRA, no hay una consensuación de cuál es el momento idóneo de transferencia del embrión al útero de la mujer durante su desarrollo. Siendo lo más habitual en la práctica clínica en el intervalo de entre 3 a 6 días; no obstante, cada embriólogo toma su decisión y existen estudios que explican las ventajas e inconvenientes de todas las posibilidades barajadas. La tendencia actual, al igual que en el centro donde se desarrolla este trabajo (Consorcio Hospital General Universitario de Valencia), es cultivar hasta el estadio de blastocisto y en esta fase transferir. Las ventajas de ello radican en que se observa el embrión en una etapa más avanzada de desarrollo, donde aquellos embriones que hayan bloqueado su crecimiento ya lo habrán hecho y donde se dispone de más marcadores morfológicos para determinar la calidad del embrión (Teh et al., 2016). Además, en esta etapa de transferencia hay mayor sincronía entre embrión y el endometrio de la futura mujer gestante. Por tanto, la transferencia en estado blastocisto mejora los resultados clínicos teóricamente. No obstante, estudios como el de Glujovsky y su equipo, concluye que hay una baja evidencia de calidad que demuestre que la transferencia de blastocistos en comparación a transferencias en días anteriores está asociada con una tasa de nacido vivo mayor (Glujovsky et al., 2022).

Habitualmente, en la práctica *in vitro* se desarrollará más de un embrión de los cuales se deberá seleccionar el mejor para transferir al útero de la mujer (Farias et al., 2023; P. Saeedi et al., 2017). De esta forma, aparece la importancia de la selección del embrión óptimo que aumente las posibilidades de implantación y que dé lugar a un nacido vivo en el menor tiempo posible, ya que todo este proceso implicará el éxito de la técnica y será influyente en las cargas emocionales y económicas de las pacientes (Sawada et al., 2021).

En la mayoría del hábito clínico actual, el método de selección empleado es la Clasificación Morfológica, es decir, mediante la calificación visual por parte de un embriólogo experimentado, basándose en ciertas características del blastocisto (Chavez-Badiola, Flores-Saiffe-Farías, et al., 2020; Harun et al., 2019). Sin embargo, las características morfológicas no aseguran el éxito de implantación y embarazo, pues es un fenómeno multifactorial que depende también de la salud paterna y materna, del ambiente del endometrio, del ambiente del cultivo, etc. (Chavez-Badiola, Flores-Saiffe-Farías, et al., 2020).

Algunas de las guías (Balaban et al., 2011) más destacadas que se han desarrollado para poder determinar si un embrión se considera de buena calidad son por ejemplo Gardner (Gardner et al., 2000) o ASEBIR (Cuevas Saiz et al., 2018). No obstante, presentan ciertas desventajas, como que estas guías únicamente se basan en tres características del blastocisto que suelen ser una adecuada masa celular interna (MCI), el número y forma de las células del trofoectodermo (TF) y el grado de expansión que se encuentra ligado al adelgazamiento de la zona pelúcida (ZP) (Chavez-Badiola, Flores-Saiffe Farias, et al., 2020). Otro inconveniente del Método de Clasificación Morfológica se trata de la tediosa labor que supone la visualización de los embriones; así como, que la elección final depende de la valoración subjetiva por parte del embriólogo basada en su experiencia y precisión (Rocha et al., 2017). De hecho, esta subjetividad ha sido estudiada y se ha visto cómo existe una variación inter-observador e intra-observador (Storr et al., 2017), lo que disminuye la reproducibilidad de la técnica. Esta es la razón por la que se precisa el automatizado de este proceso y es donde la inteligencia artificial (IA) juega un papel prometedor.

La IA se define como la capacidad de una máquina para aprender y ejercer un trabajo inteligente como lo haría un humano. La IA abarca diferentes tipos de tecnologías como el *Machine Learning*, que consiste en entrenar al programa con un amplio conjunto de datos determinando por parte del usuario qué es lo que debe aprender del *dataset*, con la intención de que posteriormente con datos nunca vistos

por la tecnología, esta sea capaz de obtener resultados nuevos similares a los aprendidos (Afnan et al., 2021; Chen et al., 2019). En el campo en el que se engloba este trabajo, las bases de datos corresponderán a imágenes o vídeos de embriones. Por tanto, los métodos de procesado de imágenes presentan la ventaja de que están automatizados y permiten estandarizar el proceso de selección de embriones. Además, permiten la evaluación de parámetros que el ojo humano no puede apreciar extrayendo más características del blastocisto que podrían estar relacionadas con un éxito de transferencia. También reducen la variabilidad y proveen de un método más objetivo, rápido y no invasivo; a diferencia de otras técnicas como el Test Genético Preimplantacional de Aneuploidías (PGT-A) (Afnan et al., 2021; Chavez-Badiola, Flores-Saiffe-Farías, et al., 2020; Isa et al., 2023; P. Saeedi et al., 2017). No obstante, requieren de gran capacidad informática y de la disposición de una base de datos masiva (Isa et al., 2023).

Existen ejemplos de IA ya publicados para el uso en el procesamiento de imágenes de embriones que ayuden al día a día de los biólogos en los laboratorios de reproducción asistida. Uno de ellos es ERICA (Chavez-Badiola, Flores-Saiffe-Farías, et al., 2020), un algoritmo de clasificación inteligente que fue entrenado con 1231 imágenes de embriones y que es capaz de diferenciar entre un embrión de buen pronóstico (euploide) de uno que no lo es (aneuploide y no implantado). Otro ejemplo de un modelo de aprendizaje pre-entrenado con imágenes es *Cell-profiler* (McQuin et al., 2018); sin embargo, en este caso se trata de fotografías de embriones de ratón. En adición, se encuentra Blas-Net (Rad et al., 2019), un sistema de segmentación automatizada de blastocistos humanos a partir de imágenes de microscopio. Aunque en proceso de desarrollo, existen gran número de herramientas basadas en la inteligencia artificial por la necesidad que acarrea.

Pese a que la elección final del embrión se realiza en cuanto a su morfología, también se considera importante la evaluación del desarrollo del embrión durante todas sus etapas de crecimiento; generando que los parámetros morfocinéticos deban tenerse en cuenta. Para lo que el uso de los incubadores dotados de tecnología *Time Lapse* (TL) es recomendable (Armstrong et al., 2019) , ya que permiten la visualización de toda la embriogénesis: del paso del embrión de unas pocas células a mórula; la expansión que da lugar al blastocisto, el *hatching*, y finalmente, la obtención del embrión "*hatcheado"*.

Se considera un blastocisto en proceso de *hatching* cuando al menos una célula se encuentra fuera de la zona pelúcida mientras que un blastocisto "*hatcheado"* es cuando todo el embrión o la mayoría se encuentra fuera de la ZP (Farias et al., 2023). La zona de *hatching* o eclosión es un prerrequisito crítico para asegurar la implantación y éxito del embarazo temprano (Seshagiri et al., 2016). La razón de ello reside en que el embrión precisa de realizar el proceso de eclosión para tener capacidad de implantarse en el útero, porque son las células del TF las que poseen los receptores para el endometrio; en cambio, la ZP no. Por tanto, si se queda envuelto en la ZP y no consigue salir, no se dará nunca la implantación ni gestación. En ocasiones es el fallo en el *hatching* lo que puede explicar la baja tasa de éxito de las TRA (Fong et al., 2001). De hecho, se practica la eclosión asistida para ofrecer una mejora en las tasas de implantación. Esta implica la degradación artificial de la ZP, con el adelgazamiento o la perforación o la eliminación completa de la zona pelúcida (Hammadeh et al., 2011). Entonces la evaluación de este proceso es importante (Gerri et al., 2020).

2. OBJETIVOS.

El trabajo presentado forma parte de un proyecto multicéntrico aprobado por los comités de ética e investigación del Hospital General Universitario de Valencia (HGUV). En él, se va a desarrollar un modelo de inteligencia artificial (IA) basado en *Machine Learning* que permitirá predecir los resultados de calificación de los embriones humanos en estadio de *hatching*.

Los objetivos específicos son:

- Generar un modelo de IA que permita la calificación de embriones en estadio *hatching* de forma no invasiva.
- Incorporar una técnica que facilite y mejore la toma de decisiones de los embriólogos cuando se selecciona un blastocisto a transferir a una mujer.
- Comparar la evaluación del procesado de imágenes automatizado con el método convencional de selección de blastocistos para la transferencia.
- Promover y explicar las ventajas del empleo de la IA en un ámbito diferente como es el de la medicina reproductiva.
- Exponer el grado de relación del trabajo con los Objetivos de Desarrollo Sostenible (ODS) de la Agenda 2030 (Anexo I).

3.MATERIAL Y MÉTODOS.

3.1. PROTOCOLO GENERAL DE LAS TÉCNICAS DE REPRODUCCIÓN ASISTIDA.

El término TRA (técnicas de reproducción asistida) abarca todas las metodologías donde esté incluida la manipulación *in vitro* de ovocitos, espermatozoides y embriones, que tengan un fin reproductivo. Mientras que un término más amplio es la "Reproducción Médicamente Asistida" (MAR, por sus siglas en inglés), donde se incluye por ejemplo la Inseminación Artificial (De Geyter, 2019).

En cuanto a las TRA, estas pueden variar dependiendo de la clínica o centro donde se realizan. No obstante, todas siguen un proceso habitual que comienza con la evaluación de la fertilidad de los individuos que se van a someter a las técnicas.

La pareja o persona debe ser evaluada para determinar la causa de su infertilidad o esterilidad y ello puede incluir análisis de sangre, pruebas de ovulación o seminogramas, entre otros; para poder aplicar los métodos que garanticen la mayor probabilidad de consecución del embarazo. Entre esos métodos se encuentra la estimulación ovárica de la mujer que consiste en la toma de fármacos para la obtención de una cohorte homogénea de óvulos maduros y evitar el fenómeno de la dominancia que suele ocurrir de forma natural. Cuando tiene lugar la dominancia, implica que únicamente un ovocito será ovulado. Sin embargo, en las TRA se pretende la obtención de más de uno para poder aumentar las probabilidades de éxito. Una buena estimulación resulta en la recuperación de complejos cúmulo-ovocito (COC) bien expandidos, como se espera de cada folículo mayor a 14 mm de diámetro, con una alta proporción de ovocitos en metafase II (MII) (maduros).

Por tanto, cuando mediante una ecografía se detecta el tamaño de los folículos adecuado, se induce la ovulación y aproximadamente 36 horas después se extraen los óvulos de los ovarios mediante una aguja

guiada por ultrasonido. Destacando que se debe cumplir con los indicadores clave de rendimiento (denominados en inglés *Key Performance Indicators*), que señalan que tras la estimulación se espera recuperar entre el 70-80% de ovocitos ("The Vienna Consensus: Report of an Expert Meeting on the Development of ART Laboratory Performance Indicators," 2017).

De los óvulos extraídos se utilizan los que son maduros (estadio de MII), mientras que los inmaduros (Metafase I o Vesícula Germinal) son descartados. A los seleccionados se les realizará una fecundación controlada *in vitro* (considerado el día 0 del cultivo del futuro embrión resultante). Las técnicas más habituales son FIV o ICSI, aunque ICSI es la más empleada actualmente (Adamson et al., 2018), en especial en casos de infertilidad masculina pues solo se precisa un espermatozoide de calidad. Además, mediante ICSI se reduce el fallo de fecundación o la posible poliespermia (fecundación de un óvulo con más de un espermatozoide que originaría un embrión no viable), ya que el embriólogo introduce el espermatozoide en el ooplasma, aportándole más control sobre el proceso (Haddad et al., 2021).

Una vez que los óvulos son fecundados, se colocan en un incubador y se cultivan. La Ley 14/2006, de 26 de mayo, sobre técnicas de reproducción humana asistida, establece como el máximo de tiempo que se puede cultivar embriones es 14 días desde la fecundación (Ley 14/2006, de 26 de mayo, sobre técnicas de reproducción asistida humana publicado en el BOE). No obstante, en la práctica clínica general se mantienen mucho menos tiempo, pues entre las desventajas se encuentra que un cultivo prolongado puede aumentar el riesgo de anomalías. Será entre el intervalo de los 3 a 6 días cuando se finalice el cultivo y se transfiera el embrión al útero de la mujer para lo que habrá que realizar una selección del mejor. Normalmente, se realiza en estadio de blastocisto (día 5-6).

El supuesto explicado ha sido partiendo de lo que se denominan ovocitos "frescos" y embriones también "frescos". Es decir, no ha habido necesidad de una crioconservación de ningún componente del proceso. En caso de que se partiera de algún gameto o un embrión criopreservado, previamente a su empleo, se requiere un proceso de desvitrificación (ovocitos y embriones) o descongelación (espermatozoides).

Después de la transferencia, se realiza un seguimiento para asegurarse de que el embrión ha sido implantado y se está desarrollando correctamente. Si el embrión se implanta en el útero, la mujer puede someterse a una prueba de embarazo para confirmar la gestación.

Es importante tener en cuenta que el proceso exacto de reproducción asistida puede variar según la técnica utilizada y las necesidades específicas de cada pareja. Además, este proceso puede requerir varios ciclos de tratamiento antes de lograr un embarazo exitoso.

3.2. CRITERIOS DE CLASIFICACIÓN EMBRIONARIA.

La clasificación embrionaria en cuanto a su calidad es una práctica necesaria en las clínicas de reproducción asistida para la elección de los mejores embriones cultivados. Existen diversos métodos para realizar la selección, aunque se suele escoger aquellos que impliquen técnicas no invasivas. Dentro de las técnicas no invasivas, la más habitual es la caracterización morfológica en el estadio de blastocisto; para lo que se han desarrollado varias guías. La que se emplea en este trabajo es la desarrollada por la Asociación para el Estudio de la Biología de la Reproducción (ASEBIR), gracias a su grupo de interés de embriología que se encarga de estudiar y validar los criterios de valoración

embrionaria a fin de obtener un lenguaje común entre profesionales y pacientes. Su última actualización de dichos criterios fue en 2018 y se espera una futura actualización en noviembre de este 2023.

Los criterios de clasificación para embriones en día 5 y 6 de cultivo (Tabla 1), lo que corresponde a estadio blastocisto, son los que se han empleado para la etiquetación manual de las imágenes de los embriones realizando el proceso de eclosión o *hatching*. Se basan en tres principales características. La primera es la calidad del trofoectodermo (TE), se espera según el número de células que hay en el plano ecuatorial que sea un embrión de tipo A (excelente) si contiene un número mayor o igual a 14 células. En cambio, será tipo B (bueno) si tiene entre 11 y 13 células; mientras que si posee 10 células o menos será de calidad tipo C (regular). La segunda característica que permite la clasificación orientativa está basada en la morfología de la masa celular interna (MCI). Tipo A se le considera a una MCI compacta de 1900 a 3800 μm^2; Tipo B será cuando no está compacta y mide de 1900 a 3800 μm^2; tipo C si su tamaño es menor a 1900 μm^2 y finalmente, tipo D (calidad pésima) si tiene signos de degeneración. La última característica es el grado de expansión orientativo según el diámetro del blastocisto sin zona pelúcida (ZP); considerando un blastocisto en expansión si el diámetro es menor o igual a 165 μm y un blastocisto expandido si es mayor a 165 μm. Finalmente, si el embrión analizado no ha alcanzado el desarrollo de blastocisto, se le puede calificar como E de *early* en inglés, es decir, temprano. Por tanto, juntando la valoración de las tres características, se determina si el embrión es desde tipo A (mejor calidad) hasta D (sería descartado) (Figura 1).

Tabla 1: Resumen de los criterios de ASEBIR para calificación de embriones en estadio de blastocisto (Cuevas Saiz et al., 2018).

	Trofoectodermo	**Masa Celular Interna**	**Cavidad blastocística**
Calidad A	>14 células	Compacta. Tamaño de 1900 a 3800 μm^2	Expandida si el diámetro es <165 μm
Calidad B	Entre 11 y 13 células	No compacta. Tamaño de 1900 a 3800 μm^2	En expansión si el diámetro es menor o igual a 165 μm
Calidad C	Menor o igual a 10 células	Tamaño menor a 1900 μm^2	Temprana o cavitando
Calidad D	Muy pocas células y degeneradas	No compacta. Tamaño menor a 1900 μm^2. Signos de degeneración	Mórula

En adición, en la guía se incluyen ciertas consideraciones como que son favorables aquellos embriones con el tamaño de las células del TE más homogéneo y la MCI de mayor tamaño y compactación. Además, se debe realizar la valoración entre embriones de misma cronología de desarrollo, sin tener en cuenta el desarrollo de los días anteriores hasta alcanzar el día 5 o 6.

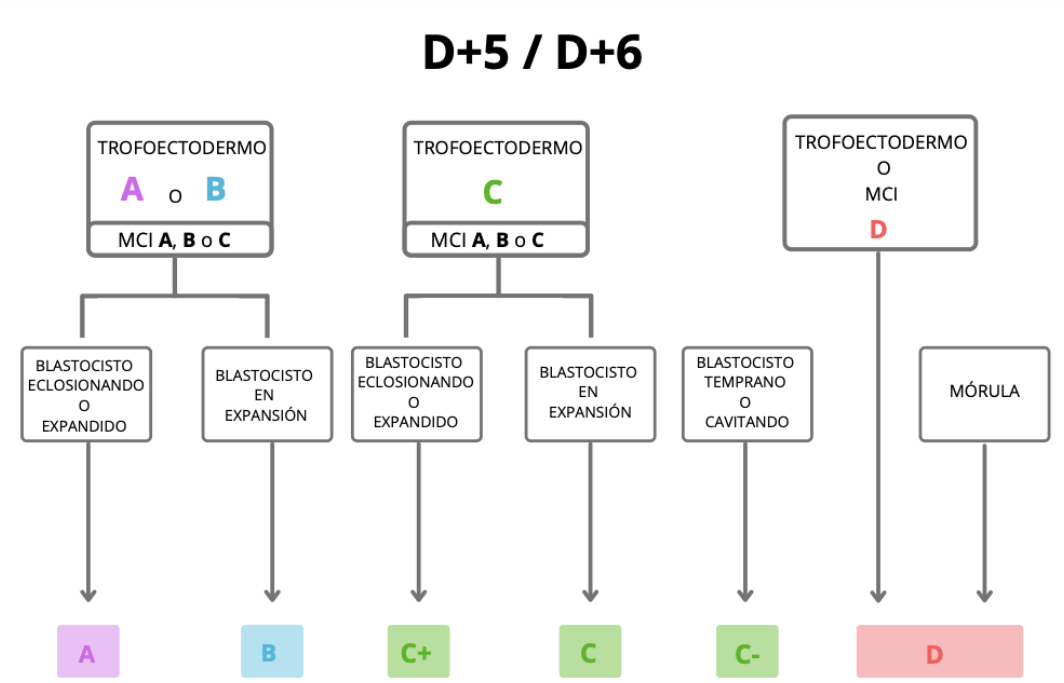

Figura 1: esquema de la fusión de los criterios de clasificación de las tres características morfológicas más importantes en embriones en día 5 o 6 de cultivo (trofoectodermo, masa celular interna y expansión de la cavidad blastocística) de ASEBIR para el veredicto de la calidad del blastocisto (Cuevas Saiz et al., 2018).

En la siguiente figura (Figura 2), se muestran ejemplos de las diferentes clasificaciones de calidad en fotografías de embriones empleadas en el desarrollo del modelo:

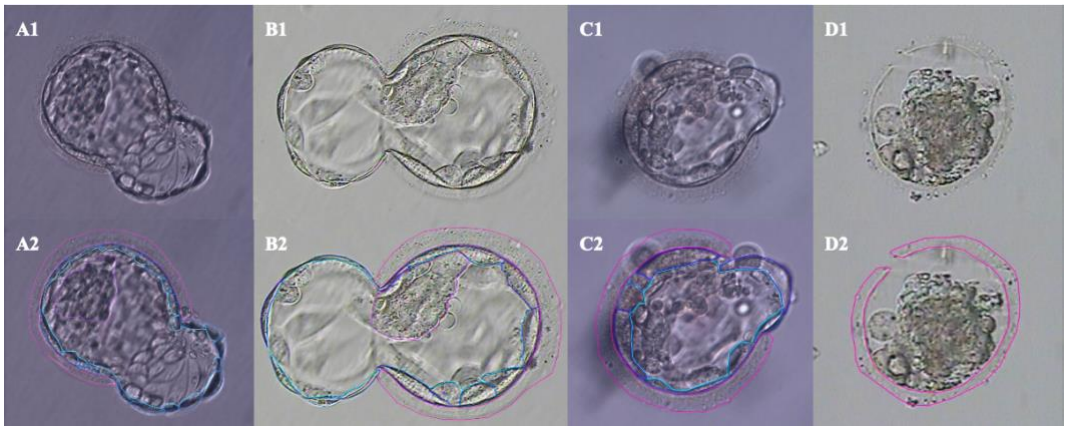

Figura 2: Ilustraciones de embriones empleados para el desarrollo del modelo en los que cada letra corresponde con la graduación de calidad que se les etiquetó. Las identificadas con el número 1 se trata de la imagen previa a la segmentación manual y las que contienen el número 2 son las segmentadas, en las cuales se puede visualizar las diferentes partes del blastocisto delimitadas.

3.3. DESCRIPCIÓN DE LA BASE DE DATOS.

El estudio retrospectivo multicéntrico realizado se ha basado en el desarrollo de un modelo de inteligencia artificial para el que se emplearon 810 imágenes de blastocistos en estadio de *hatching* , con 891 embriones analizados en total; por tanto, algunas de las imágenes contaban con más de un embrión (Tabla 2). Todas ellas provenían de 3 clínicas diferentes, siendo una de ellas en la que se desarrolla este trabajo, el Laboratorio de Embriología de la Unidad de Medicina Reproductiva del Consorcio HGUV. De las 810 imágenes totales, de forma aleatoria se separaron en imágenes que se emplearon para el entrenamiento del modelo, validación y para testear este mismo. Correspondiendo al 90,4% (726 imágenes) de entrenamiento y validación y el 9,6 % (81 imágenes) de prueba. Dentro de ese 90,4% de entrenamiento y validación, más exactamente se divide de nuevo en 90,28% de entrenamiento y 9,72% de validación. Por tanto, el *split* o división del conjunto de datos está siendo de aproximadamente del 10%.

Tabla 2: Descripción del *set* de partida de imágenes de embriones con las que se desarrolla el modelo. Cuenta con el número de imágenes, la cuantificación de embriones detectados por cada fotografía y las frecuencia y número de veces en la que estos aparecen.

	Dataset		
Número de embriones	891		
Número de imágenes	810		
Número de embriones por imagen	1	2	3
Veces en las que se detecta 1-3 embriones por imagen	734	71	5
Porcentaje sobre el total de embriones	90,62 %	8,76 %	0,62 %

Un factor interesante de este modelo que lo diferencia de trabajos previos similares es que el *set* de datos es randomizado, es decir, las condiciones de obtención de cada imagen son diferentes. Cada clínica ha realizado su práctica habitual, sin consensuar una forma de cultivo común, ni el tipo de paciente del que proviene el embrión, ni la máquina a través de la cual se ha obtenido la imagen (microscopio, *Time Lapse,* AINE, *etc*). Tampoco se han editado las fotografías ni se han excluido embriones a menos que las imágenes estuvieran repetidas o la fotografía presentara una calidad pésima tan considerable como para no poder extraer de ella ninguna información. Esto se puede demostrar en el hecho de que en la base de datos existe diversidad en el tipo de imágenes, pues algunas presentan más de un embrión por fotografía. Todo ello, con el objetivo de desarrollar una herramienta que verdaderamente sea aplicable y reproducible en diferentes bases de datos y clínicas. A diferencia del algoritmo de clasificación inteligente ERICA (Chavez-Badiola, Flores-Saiffe-Farías, et al., 2020) que normalizaron su *dataset* empleando únicamente fotografías nítidas de embriones fecundados mediante ICSI, con edad conocida de las pacientes y con microscopios de la misma marca (Olympus®).

3.4. DESCRIPCIÓN DE LA INTELIGENCIA ARTIFICIAL EMPLEADA.

La Inteligencia Artificial es un término genérico que se puede separar en subáreas específicas: como el aprendizaje automático (ML) y el aprendizaje profundo (DL) (Fernández et al., 2020). La diferencia radica en que el aprendizaje profundo o Deep Learning (DL), es un subconjunto dentro del Machine Learning (ML) (Figura 3); basado al igual que este en un aprendizaje automático, pero en lugar de emplear algoritmos de regresión o árboles de decisión, utiliza redes neuronales profundas que simulan las conexiones neuronales del cerebro humano para aprender grandes cantidades de datos (Fernández et al., 2020; Swain et al., 2020). Las redes neuronales profundas mencionadas (DL), son modelos compuestos por múltiples capas interconectadas por nodos, donde cada nodo procesa la información y la transmite a los nodos de la siguiente capa. De esta forma, la simulación del pensamiento humano tiene como objetivo generar modelos predictivos basados en datos y conocimientos específicos que permitan apoyar, mejorar o resolver problemas específicos.

Figura 3: Esquema ilustrativo que grafica como el *Deep Learning* es un subtipo de *Machine Learning*, siendo este a la vez un tipo de Inteligencia Artificial (imagen generada con BioRender ©).

El aprendizaje automático puede ser: supervisado, no supervisado o de refuerzo. Supervisado hace referencia al ajuste de un modelo de aprendizaje en el que los datos han sido previamente etiquetados por ejemplo por el usuario humano; entonces, se facilitan los datos de entrada y salida conocidos. Dentro de este enfoque supervisado se puede diferenciar entre algoritmos de clasificación, para clasificar objetos dentro de categorías; o de regresión, para predecir un valor numérico. Por otro lado, el aprendizaje no supervisado se basa en un modelo que identifica patrones y busca tendencias similares en el *dataset* sin etiquetar; por tanto, no hay una interferencia humana previa y se emplea para predecir resultados desconocidos. Finalmente, el de refuerzo funciona a base de ensayo y error, en el que, gracias al procesado de datos y obtención de resultados positivos o negativos, el modelo se construye a través de la retroalimentación de las experiencias previas (Baştanlar & Özuysal, 2014; Greener et al., 2022; Wang et al., 2019).

El aprendizaje automático puede emplearse en muchas áreas, pero destaca su aplicabilidad en el ámbito sanitario, por ejemplo, en el área de las técnicas de reproducción asistida (TRA) (Figura 4) (Wang et al., 2019). En la medicina reproductiva, uno de los aspectos claves es la visualización y selección de los mejores gametos o embriones y es en este campo en donde la IA está desarrollándose. Para ello, el *Machine Learning* se emplea en el procesado de imágenes de microscopio o de incubadores TL.

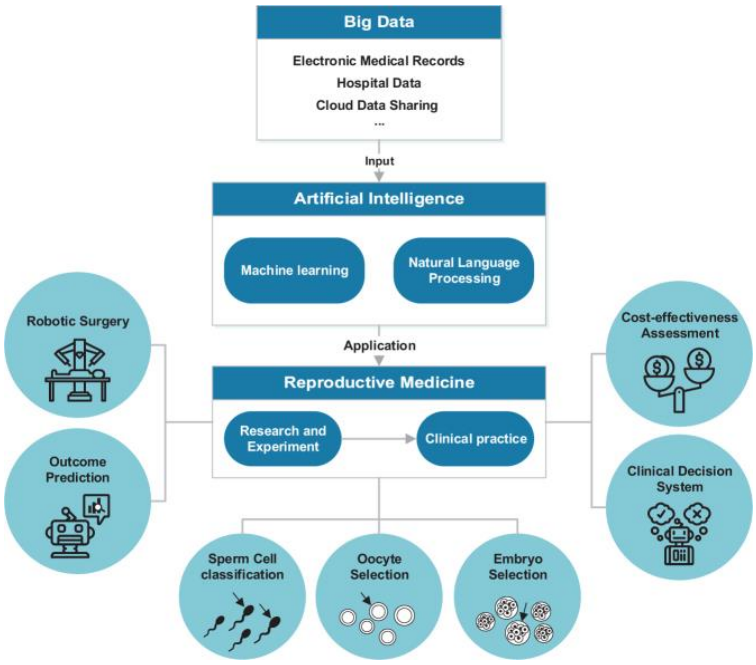

Figura 4: El papel de la inteligencia artificial en la medicina reproductiva, recogiendo las 7 aplicaciones principales (Wang et al., 2019).

Cuando el modelo ML procesa una imagen lo puede llevar a cabo de diferentes formas. Además, dependerá de si en la imagen hay un único objeto o varios. En la imagen, se puede identificar, es decir, la IA puede indicar que en la fotografía hay un objeto determinado (clasificación); también se puede clasificar y ubicar el objeto dentro de la imagen (clasificación y localización) o, por último, segmentar. Segmentar se basa en identificar el objeto, localizarlo y también delimitarlo exactamente. Dentro de los tipos de segmentación se encuentra la segmentación semántica que es el proceso de segmentación de los píxeles de la imagen en sus respectivas clases. Entonces, varios objetos de la misma clase se consideran como una sola entidad. Mientras que la segmentación de instancias es más exhaustiva ya que pese a que también segmenta por píxeles la imagen, trata a los varios objetos presentes en la imagen que son de la misma clase como entidades distintas.

Para el trabajo presentado, se emplea la IA de aprendizaje automático *Machine Learning* de forma supervisada, ya que las imágenes de los embriones se etiquetaron previamente según su calidad y según su estadio de embriogénesis como "*hatching*", "*hatched*" o sin etiqueta en caso de que se correspondiera a blastocistos de día 5 o 6 de cultivo sin eclosionar. Además, la forma de procesado de la imagen fue por segmentación de instancias, donde cada elemento del blastocisto (células del trofoectodermo (TE), masa celular interna (MCI), blastocele, zona pelúcida (ZP), etc.), se delimitaron exactamente y de forma independiente. La elección de este tipo de segmentación es debida a que de esta forma se puede contar el número de elementos segmentados y clasificarlos según su calidad; lo que permite determinar la condición de óptimo o no gracias a los criterios de ASEBIR (Cuevas Saiz et al., 2018).

Más exactamente el modelo empleado contiene dos componentes: un primer modelo de segmentación denominado Mask R-CNN y un segundo modelo de calificación embrionaria que es un modelo propio tabular. Mask R-CNN (*region-based convolutional neural network*) se basa en que, a partir de varios objetos dentro de una imagen, proporciona la etiqueta (clasificación), la detección de los distintos objetos (*bounding box*) y la máscara de segmentación (*instance segmentation*). La razón de la elección de esta arquitectura se basa en que otros modelos típicamente empleados no se ajustaban a los resultados que se quería obtener. Por ejemplo, existen modelos que únicamente clasifican (como ResNet o

DenseNet), dando etiquetas que se refieren a toda la imagen, pero no a objetos específicos. Por otro lado, también hay modelos que detectan los objetos y clasifican (como Faster RCNN, YOLO o DETR), pero no segmentan. Otra alternativa sería los modelos que solo hacen segmentación de toda la imagen asignando un valor a cada píxel; sin embargo, no permiten individualizar los distintos objetos de la imagen. Esta última opción planteada es conocida y ha sido mencionada anteriormente como segmentación semántica. Un ejemplo podría ser U-Net, que es un modelo frecuentemente empleado en medicina para la segmentación de vasos de la retina en el fondo del ojo, tumores en cerebro, etc. Por ello, la arquitectura que permite obtener los objetivos necesarios (detección de objetos, clasificación y segmentación) es la seleccionada (He et al., 2017) .

3.5. DESARROLLO DEL MODELO DE INTELIGENCIA ARTIFICIAL.

El proceso de aprendizaje automático desarrollado constó de los siguientes pasos (Figura 5):

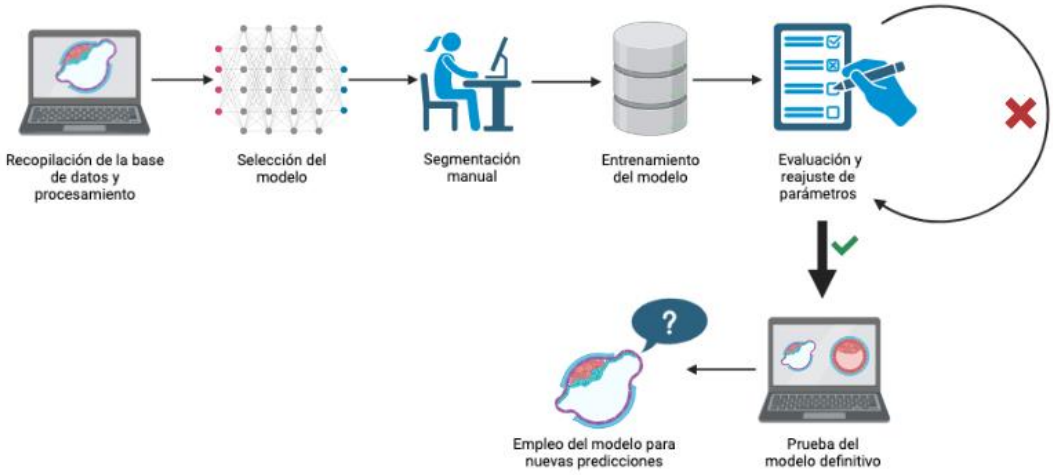

Figura 5: Esquema simplificado del proceso de desarrollo de un modelo *Machine Learning* supervisado para la predicción de la calidad y segmentación de embriones en estadio *hatching* (imagen generada con BioRender ©).

3.5.1. TRABAJO PREVIO.

El primer paso es la **recopilación de datos** relevantes para el problema que se desea resolver. En el caso del trabajo, se recopilaron las imágenes de embriones en estadio *hatching* provenientes de las clínicas participantes y de todo tipo de fuentes ya sean distintos microscopios como incubadores dotados de tecnología *Time Lapse*. Una vez que se recopilan los datos, se realizan **tareas de preprocesamiento** para garantizar que los datos estén en un formato adecuado para el modelado. Esto incluye la normalización de los píxeles de la imagen y la conversión en imágenes cuadradas que son las características que precisa el modelo como entrada. A continuación, **se selecciona un modelo de aprendizaje automático** que se adapte al problema en cuestión. Para el desarrollo del modelo se emplean las redes neuronales (*Deep Learning*), más concretamente el modelo utilizado está formado por dos componentes: el primero, de segmentación, basado en Mask R-CNN que recibe como entrada (*input*) una imagen y como salida (*output*), esa imagen con la máscara de segmentación. De esa máscara

de segmentación se puede extraer un resumen de las características obtenidas (número de blastómeros o células de TF buenas y malas, tamaño del blastocele, etc.). Posteriormente, utilizando estas características de segmentación como entrada, se utiliza un segundo modelo propio basado en datos tabulares, que devuelve como salida el valor del *grading* predicho (A, B, C, D o E). La letra E hace referencia a que el blastocisto se encuentra todavía en un estadio muy temprano, pero en este estudio no se contempla esta situación pues la base de datos la componen embriones eclosionando, que es una fase de desarrollo más avanzada.

Como el modelo de ML se basa en un algoritmo supervisado con segmentación de instancias; **se segmentó, etiquetó y clasificó de manera manual** las 810 imágenes de embriones de la base de datos. Siendo este proceso supervisado por parte de una embrióloga experta. De esta forma, se indicó al modelo las *labels* o etiquetas que se espera predecir y las *features* o características que se deberán detectar. Las etiquetas incluidas fueron "*hatching*" o "*hatched*", si se encontraban en dicho estadio de desarrollo y una valoración de su calidad según los criterios de ASEBIR (tipo A-excelente, B, C y D- pésima calidad). Las características segmentadas fueron las células del trofoectodermo de forma individual, la masa celular interna (MCI), la cavidad blastocística y la zona pelúcida principalmente. De forma excepcional, se identificaron y delimitaron las células excluidas por el embrión y las "*string*" en caso de que hubieran (Figura 6). Las "*string*" son proyecciones largas que conectan la masa celular interna con las células de trofoectodermo (Eastick et al., 2021). La segmentación sigue la siguiente leyenda de la (Tabla 3).

Figura 6: Ejemplo de segmentación manual realizada de un blastocisto eclosionando donde se pueden observar marcadas en color azul las células del TF, en color rosa la ZP, en fucsia la MCI, en turquesa el blastocele y en marrón las *string*. Además de las etiquetas seleccionadas: *hatching* y blastocisto de calidad A.

Tabla 3: Código de colores que se ha seguido durante la segmentación y etiquetación manual de todos los embriones empleados para el desarrollo del modelo.

Feature	**Color**
Blastómero bueno	Azul claro
Blastómero pésimo	Azul oscuro
Blastómero intermedio	Morado
Blastocele	Turquesa
MCI buena	Fucsia
MCI pésima	Verde oscuro
String	Marrón
Zona degenerada	Granate
Célula excluida	Amarillo
Zona pelúcida	Rosa
Puntos de *hatching*	Verde claro
Colapso	Naranja

Label	**Color**
hatching	Azul oscuro
hatched	Gris
Blastocisto A	Morado oscuro
Blastocisto B	Rosa claro
Blastocisto C	Lila
Blastocisto D	Rosa oscuro

3.5.2. ENTRENAMIENTO Y VALIDACIÓN.

El siguiente paso es el **entrenamiento o *training* del modelo** que se lleva a cabo mostrando ejemplos de entrada etiquetados y caracterizados. De esta forma, se consigue que el modelo aprenda la relación entre las etiquetas y las características proporcionadas en el paso anterior; y tanto los criterios ideales como los penalizados. Ajustando así los parámetros del modelo para minimizar el error en la tarea que se desea realizar.

En adición, en esta fase se han empleado técnicas de aumento de datos (*Data Augmentation*) con el objetivo de aumentar el tamaño y diversidad del conjunto de datos del entrenamiento. Consiste en realizar distintas transformaciones a las imágenes originales para crear imágenes derivadas, obteniendo más variedad de ejemplos. Las transformaciones son aleatorias y se aplican a todas las imágenes del *training set*, pero no a las del *test set*, ya que en esta fase se emplean únicamente las originales. Incluyen:

- Volteo horizontal y vertical (50% de probabilidad).
- Desplazamiento (de un 20% máximo).
- Rotación de +/- 170º.
- Deformación elástica de 4%.
- *Zoom* de +/- 10% (tanto en eje X como Y).
- Alteración de la saturación y el brillo (5%).
- Introducción de desenfoque aleatorio (con un valor sigma de hasta 0.25).

Mediante estas transformaciones se han generado como máximo 8 imágenes nuevas a partir de cada imagen, dando lugar a un *dataset* aumentado de 6454 imágenes (90,28%) en el conjunto de entrenamiento y 729 imágenes (9,72%) en el de validación. Un ejemplo de este proceso aleatorio se visualiza en esta representación gráfica (Figura 7).

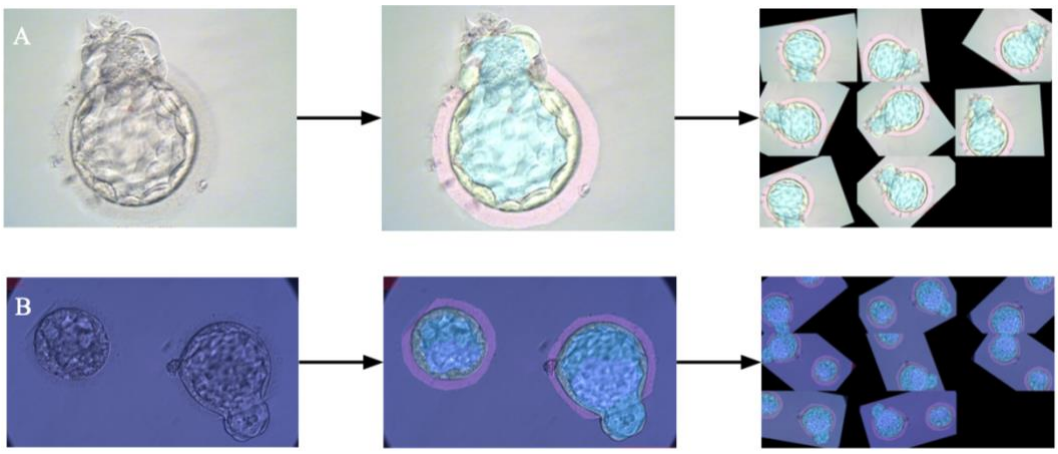

Figura 7: Imágenes reales del proceso de transformación aleatoria de las imágenes del *set* de entrenamiento, donde la primera imagen era la original y continuando la secuencia de flechas, se encuentra la imagen segmentada (con la máscara de segmentación encima de la imagen) y finalmente las imágenes transformadas, en las que se observa los 8 aumentos a partir de la original. A) Ejemplo en una fotografía que cuenta con un embrión. B) Ejemplo de dos embriones por imagen.

Después de entrenar el modelo con el conjunto de entrenamiento, **se evalúa** su desempeño con el conjunto de validación, un paquete de imágenes nunca vistas por el sistema. En caso de que no se desempeñe bien en este conjunto, **se reajustan sus parámetros** y se repite el proceso de validación hasta que se obtiene un modelo adecuado. Cuando funcione bien en dicho conjunto, se considera que ha aprendido la tarea correctamente y se puede usar para hacer predicciones en nuevos datos. Por tanto, un aprendizaje correcto será que la inteligencia artificial sea capaz de detectar todos los componentes del embrión, reconocer el estadio de desarrollo en el que se encuentra y en última estancia con el segundo modelo, también asignarle una calificación con respecto a su calidad.

Durante esta etapa habrá que tener especial cuidado con el *overfitting* o sobreajuste, que se produce cuando el modelo obtenido se ajusta tanto a los ejemplos etiquetados de entrada que no puede realizar las predicciones correctas en datos nuevos que nunca ha visto antes. Por tanto, para evitar esta situación se emplea el *data augmentation*, reajuste de hiperparámetros, validaciones repetitivas como se ha comentado anteriormente, balanceo del *set* de entrenamiento, etc.

Dichos parámetros mencionados permiten el aprendizaje óptimo y eficiente, por lo que deben ser correctamente ajustados. Entre ellos cabe destacar los que se encuentran a continuación; sin embargo, algunos de ellos serán explicados con más detalle en el apartado 6 de Estadística.

- *Batch size* de 4. El tamaño del *batch*, de cada lote del entrenamiento, lo que quiere decir que cada pequeño paso del entrenamiento se hace usando 4 imágenes.

- *"Detection_MIN_confidence"* de 0,7. Es la confianza mínima del modelo para decir que un objeto que ha detectado es real. Por ejemplo, en el caso de que el modelo indique que hay una célula del TF con un 60% de confianza, al estar por debajo del umbral que es el 70%, no se devolverá como real.

- "*Detection_NMS_threshold*" con valor de 0,3. Es el valor umbral de supresión de objetos. Las siglas "NMS" indican en inglés *Non-Maximum Suppresion*. Se emplea para cuando el objeto se

detecta varias veces y las máscaras de segmentación se superponen, entonces se eliminan las que estén solapadas porque se supone que corresponden al mismo objeto.

- *Epochs* de 500. Se trata del número de veces que el modelo se entrena con todo el *dataset*. Es necesario poner un valor suficiente para que el modelo aprenda, pero sin excederse para que el modelo no sufra un sobreajuste u *overfitting*.

- Tamaño de las imágenes (i*mage_shape)* que entran al modelo es de 256 x 256 píxeles y 3 canales puesto que la fotografía es en color, RGB. Si el tamaño es distinto, el modelo hace un reescalado automático.

- *IoU_threshold* de 0,5. Es el valor umbral para comprobar si un objeto se ha detectado correctamente. En caso de que la detección se solape más del umbral (50%) con el objeto real, quiere decir la detección ha sido exitosa.

- El número de clases que son 13. Las 12 *features* de la (Tabla 3) , incluyendo el fondo de la imagen el cual se corresponde con la clase número 0.

- Tasa de aprendizaje o *learning_rate* correspondiente a 0,001. Dicha tasa representa la velocidad a la que el modelo aprende los datos del entrenamiento. Por tanto, un valor bajo implicará que el modelo tarde más tiempo en converger en la solución; mientras que una tasa alta corre el riesgo de no proporcionar la respuesta adecuada. Entonces, la tasa de aprendizaje óptima será aquella que permita ir reduciendo el error (Figura 8).

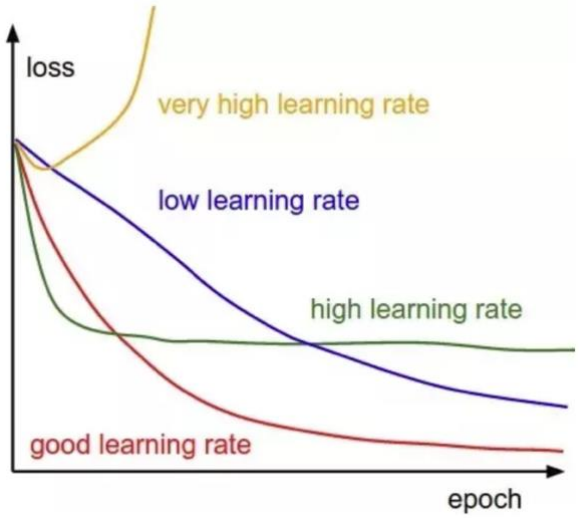

Figura 8: Ilustración de las diferentes posibilidades de tasas de aprendizaje y como estas influyen en la optimización del modelo representadas en el gráfico de épocas frente a pérdidas (KEEPCODING, 2023).

3.5.3. TEST.

Finalmente, con el modelo definitivo se emplea el *set* de imágenes de **prueba**, unas fotografías de embriones en estadio *hatching*, nunca vistas por la IA y seleccionadas aleatoriamente desde el inicio, para que desempeñe su función. Una vez se obtengan los resultados del modelo de segmentación, estos serán introducidos en el modelo de calificación embrionaria (como se ha comentado anteriormente)

siguiendo el mismo proceso, para que, a partir de las variables anotadas y las conclusiones de la segmentación, se determine el grado de calidad de los blastocistos. Todo ello, con el objetivo de poder usar el modelo en las clínicas de medicina reproductiva y que sirva como ayuda complementaria a los embriólogos en la elección del embrión de mejor calidad a transferir.

3.6. ESTADÍSTICA.

3.6.1. ESTADÍSTICA DESCRIPTIVA.

En primer lugar, se ha obtenido todo lo referente a la **estadística descriptiva** de la base de datos y el trabajo de segmentación manual; que correspondería con la evaluación cualitativa. Entre lo que se encuentra la cuantificación del total de embriones analizados, los blastocistos que han sido detectados por imagen y la frecuencia de aparición de cada tipo y etiqueta, etc.

Posteriormente, se continuó con la **estadística** de tipo más **cuantitativo.**

3.6.2. MÉTRICAS DE CLASIFICACIÓN BINARIA.

La clasificación binaria se refiere a una cuestión dentro del aprendizaje automático en la que se realizan predicciones entre dos categorías mutuamente excluyentes. Las métricas comunes de clasificación binaria son: exactitud o *accuracy*, sensibilidad, especificidad, el valor predictivo positivo o precisión, el valor predictivo negativo y la F1-*score*. Estas métricas proporcionan información sobre la capacidad del modelo para predecir las clases positivas o negativas; y de esta forma, poder concluir si el modelo desarrollado de aprendizaje automático es eficiente diferenciando cada categoría excluyente.

La **exactitud o *accuracy*** es el número de predicciones correctas del total de predicciones realizadas (Ecuación 1), tanto positivas como negativas. Es una medida intuitiva acerca del desempeño del modelo ya que indica las clasificaciones de instancias realizadas de forma apropiada en las regiones del blastocisto; pero esta métrica por sí sola no es suficiente para determinar la efectividad. La razón de ello reside en especial cuando las categorías del conjunto de datos están desequilibradas, es decir, hay mayor proporción de clases positivas o negativas. Si uno de los tipos es dominante sobre el otro, el modelo no estará aprendiendo realmente a diferenciar entre las dos categorías y se obtendrán resultados influenciados.

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad \text{(Ecuación 1)}$$

La **sensibilidad (o *Recall*)** y **especificidad** describen la proporción de verdaderos positivos y de verdaderos negativos (Ecuación 2 y Ecuación 3), es decir, la capacidad del modelo en solo positivizar en aquello que se busca. A diferencia de la métrica de exactitud, no dependen de la prevalencia de cada categoría. Aplicado a la segmentación de imágenes, el *Recall* mide la proporción de píxeles correctamente clasificados como positivos en relación con todos los píxeles de "*ground truth*" positivos (lo clasificado como correcto de forma confiable).

$$Recall = \frac{TP}{TP+FN}$$ (Ecuación 2)

$$specificity = \frac{TN}{TN+FP}$$ (Ecuación 3)

El **valor predictivo positivo (PPV) o precisión (**Ecuación 4**)** y el **valor predictivo negativo (NPV)** (Ecuación 5) se describen como la proporción de medidas positivas y negativas, respectivamente, que finalmente lo fueron. Es decir, aplicado al modelo, es la proporción de píxeles clasificados como positivos o negativos con relación a todos los píxeles de *ground truth* positivos o negativos respectivamente. Por lo que se puede emplear como un dato de probabilidad de acierto en la selección de instancias del modelo.

$$PPV = \frac{TP}{TP+FP}$$ (Ecuación 4)

$$NPV = \frac{TN}{TN+FN}$$ (Ecuación 5)

Finalmente, la **F1 *score***, es el promedio del valor predictivo positivo y la sensibilidad (Ecuación 6). Se emplea en bases de datos desequilibradas para buscar el equilibrio. Al realizar la media armónica, penaliza los casos en los que una de las dos métricas es muy baja, lo que resulta en un F1-*score* más bajo. Se mide del 0 al 1, siendo 0 un rendimiento deficiente y 1, un rendimiento óptimo.

$$F1\ score = 2 \cdot \frac{PPV \cdot Recall}{PPV+Recall}$$ (Ecuación 6)

Para poder obtener las métricas se precisa saber del modelo los verdaderos positivos (TP, true positives) o regiones extraídas de los píxeles correctamente, los verdaderos negativos (TN, true negatives) o regiones ignoradas correctamente de las imágenes, los falsos positivos (FP, false positives) o zonas de la imagen erróneamente extraídas y los falsos negativos (FN, false negatives) o regiones perdidas de los píxeles que se deberían haber identificado (Kragh & Karstoft, 2021; P. Saeedi et al., 2017).

Todo ello permite realizar una **matriz de confusión**. Es una representación tabular que permite comparar las predicciones realizadas con el modelo con las reales. En ella, como se observa en la (Figura 9), las filas representan la realidad (*outcome*), mientras que las columnas son las predicciones del modelo (Rocha et al., 2017).

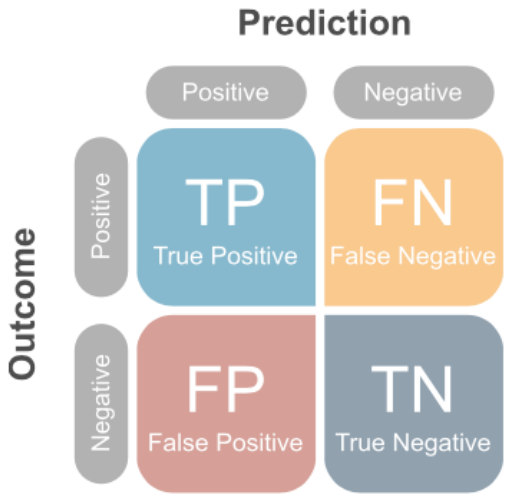

Figura 9: Esquema de la matriz de confusión (Kragh & Karstoft, 2021).

3.6.3. MÉTRICAS A NIVEL DEL MODELO.

Generalmente los modelos de IA proporcionan valores de salida continuos, de los que se debe establecer un punto de corte antes de determinar la predicción binaria. Un umbral que diferencie entre embriones de calidad positiva para transferir y embriones de pésima o menor calidad que no serían seleccionados por el embriólogo.

El **área bajo la curva (AUC) de la curva ROC** (*Receiver Operating Characteristic*) proporciona una medida global del rendimiento del modelo en todos los posibles umbrales de clasificación. Esta métrica es útil porque no se basa en un umbral específico y ofrece una evaluación general de la capacidad de discriminación del modelo, sin importar dónde se establezca el umbral de clasificación.

Para obtener la curva ROC se requieren la tasa de verdaderos positivos o sensibilidad y la tasa de verdaderos negativos o especificidad. Con estos datos se grafica los valores de sensibilidad por la proporción de falsos positivos dada por (1 - especificidad) para un conjunto diverso de puntos de corte (Fawcett, 2006). Una vez obtenida la representación, se calcula el área bajo la curva que es independiente de la prevalencia, porque proviene de datos de sensibilidad y especificidad también independientes como justifican Fly y Karstoft en su artículo (Kragh & Karstoft, 2021). La AUC podrá tener un valor entre 0 y 1; por lo que, cuanto mayor sea el valor, mejor rendimiento tendrá el modelo en la clasificación de instancias y en sus predicciones.

En adición, se puede extraer la puntuación de confianza o ***model confidence***, ya mencionada anteriormente. Es una métrica que hace referencia al nivel de certeza que asigna un modelo a una predicción específica. Por ejemplo, en el caso planteado, si la IA detecta un blastocele, devuelve una puntuación del grado de confianza en el que realmente el blastocele detectado lo sea. Este umbral de seguridad se puede establecer manualmente, por ejemplo, que solo se obtenga de *output* aquello con un grado de confianza mayor al 70%. De esta forma, se evitan predicciones erróneas que generalmente son ruido.

Otra métrica para obtener el rendimiento de precisión de detectores de objetos es **IoU**. Se emplea para medir la superposición entre el límite del objeto detectado previsto con el verdadero límite segmentado del objeto. También es conocido como *Jaccard Index* y se calcula dividiendo el área de intersección entre el área de unión. El área de intersección es el área que se superpone entre los dos conjuntos o regiones; mientras que el área de unión se trata del área total cubierta por ambos conjuntos (Figura 10). Este índice puede tener un valor entre 0 a 1, donde 0 significa que no hay superposición entre los conjuntos y 1 indica una superposición total o coincidencia perfecta. Por tanto, un resultado de IoU cercano a 1 en nuestro modelo, significará que es capaz de segmentar, es decir, delimitar el objeto detectado con buena exactitud (AHMED FAWZY GAD, 2020; MARK EVERINGHAM & JOHN WINN, 2012).

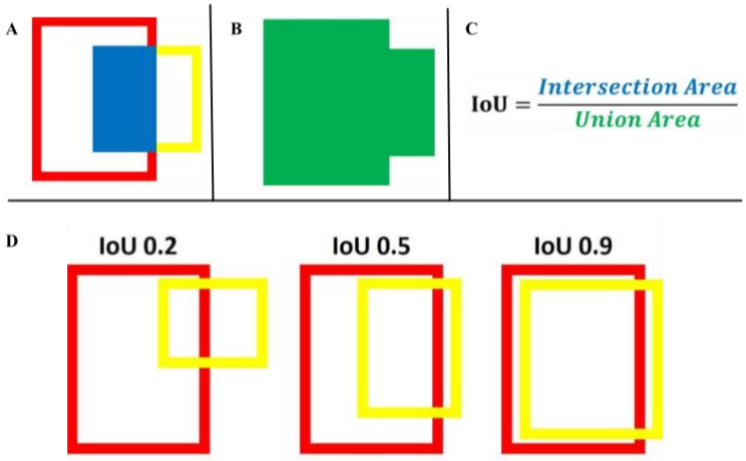

Figura 10: Explicación visual de la fórmula de IoU (*Intersection over Union)* (AHMED FAWZY GAD, 2020). (A) Representación de la segmentación de un objeto, donde en color rojo se encuentra la verdadera delimitación del objeto, en amarillo el límite que el modelo predice y en azul la superposición de ambos, conocida como el área de intersección. (B) El área que engloba todas las delimitaciones, denominada el área de unión. (C) La fórmula de la IoU donde cada componente se encuentra del color de su representación en las imágenes A y B. (D) Ejemplos de valores de IoU y sus representaciones de los límites reales y predichos, en los que se observa que a menor valor de IoU, peor predicción de la realidad hace el modelo.

4.RESULTADOS.

4.1. EVALUACIÓN CUALITATIVA.

En primer lugar, se obtuvieron los resultados estadísticos descriptivos referentes a la segmentación manual del conjunto de datos, de la cual el modelo pudo aprender y desarrollarse. Las *labels* o etiquetas seleccionadas correspondían al estadio del embrión y a la graduación de calidad de este, recogiéndose en la siguiente tabla (Tabla 4):

Tabla 4: Cuantificación de las diferentes etiquetas de estadio y calidad impuestas a los 891 embriones fotografiados del *dataset.*

Calidad de los embriones del *set*	**Frecuencia de aparición de cada tipo de etiqueta de calificación: Número de veces [Porcentaje]**
Calidad A	359 [40,29 %]
Calidad B	354 [39,73 %]
Calidad C	162 [18,18 %]
Calidad D	2 [0,23 %]
Descartados	14 [1,57 %]

Estadio de los embriones del *set*	**Frecuencia de aparición de cada tipo de etiqueta de estadio: Número de veces [Porcentaje]**
Hatching	806 [90,46 %]
Hatched	12 [1,35 %]
Blastocisto	73 [8,19%]

Se puede observar como en la gran mayoría se trata de embriones en estadio *hatching* que es el objetivo del estudio. En cuanto a la graduación de la calidad, mayoritariamente son A y B, puesto que las imágenes provienen de embriones que fueron transferidos y las clínicas siempre que sea posible transferirán de este tipo, ya que son los que más probabilidades de éxito poseen. Además, se observa un número de imágenes en las etiquetas de calidad como descartadas. La razón de ello reside en que se han incluido todo tipo de imágenes para generar el modelo, por tanto, en ocasiones es posible que la imagen no posea la calidad suficiente como para poder determinar la graduación del embrión, pero que sí sea visible el estadio en el que se encuentra y se pueda segmentar algún componente en el plano focal que contribuya a continuar con el desarrollo del modelo. Cabe recordar que un embrión realizando la eclosión posee una morfología característica en forma de infinito sencilla de reconocer.

También se ha cuantificado las anotaciones de las características detectadas en cada blastocisto (Tabla 5). De las 810 imágenes originales, fueron anotadas 807, lo que corresponde al 99,63%. De nuevo, la causa de ello se debe a la pobreza en la calidad de algunas fotografías.

Tabla 5: Estadística descriptiva de la segmentación manual de instancias en el *dataset* original, por lo que se recoge la cuantificación de todos los elementos detectados en los embriones y la media de cada uno de ellos por cada embrión.

	Segmentación del *set* original	
Feature	**Número de detecciones**	**Media por embrión**
Blastómero bueno	15300	17,17
Blastómero pésimo	682	0,77
Blastómero intermedio	0	0,00
Blastocele	670	0,75
MCI buena	560	0,63
MCI pésima	52	0,06
String	146	0,16
Zona degenerada	1	0,001
Célula excluida	56	0,06
Zona pelúcida	879	0,99
Puntos de *hatching*	0	0,00
Colapso	0	0,00

En promedio en los datos originales, se detectan más de 14 células del blastómero (TF) correctas por embrión, al igual que casi una zona pelúcida por cada uno, porque sólo aquellos embriones ya eclosionados (*hatched*) no la poseen y estos se encuentran en minoría en la base de datos. En adición, aquellas características que disminuyen la graduación de calidad del embrión se encuentran en menor proporción como podría ser una MCI pésima. Un dato destacable podría ser que no se ha anotado el blastocele en todos los embriones, a causa de que no se haya podido visualizar correctamente. Además de otras características en menor proporción puesto que son elementos que opcionalmente se pueden visualizar o porque no corresponde su aparición con el estadio que se está estudiando.

4.2. MODELO DE SEGMENTACIÓN.

4.2.1. ENTRENAMIENTO Y VALIDACIÓN.

A medida que el modelo se entrena, las gráficas de pérdida y épocas proporcionan la información acerca del rendimiento que se está obteniendo (Figura 11).

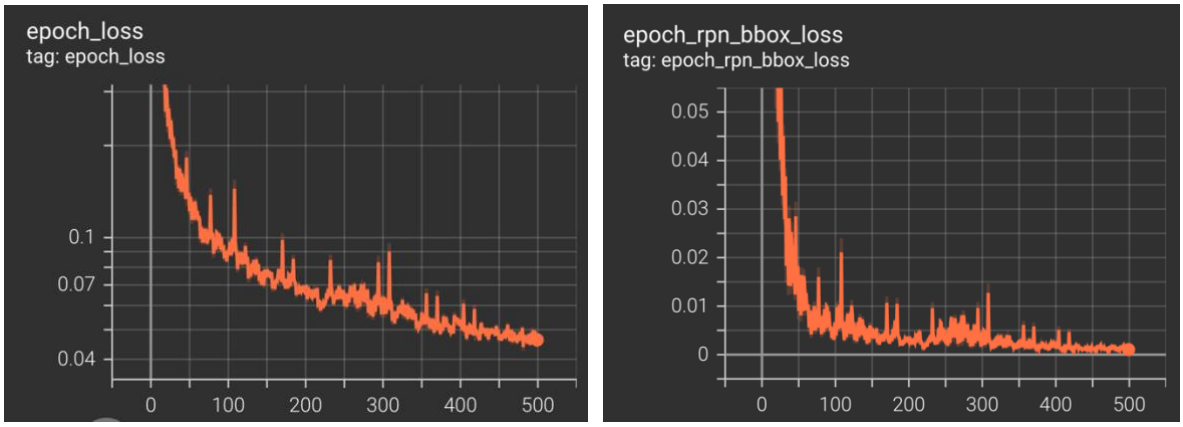

Figura 11: curvas resultantes del entrenamiento. Ambas curvas grafican el parámetro *epoch* o épocas en el eje X frente a la función de pérdida o *loss* en el eje Y. No obstante, la de la derecha representa el parámetro de función de pérdida en la eficiencia en la que la red neuronal (RPN, *Region Proposal Networks*) detecta los objetos y los distingue del fondo.

Un resultado esperado e ideal es aquel en el que se observa como la curva disminuye progresivamente, ya que la función de pérdida mide la discrepancia entre las predicciones del modelo y los valores reales. Por tanto, cuanto más disminuya, menor será la discrepancia y mejores predicciones será capaz de realizar el modelo. Esta disminución puede ser observada en las gráficas de la (Figura 11). Entonces tanto en la función de pérdidas general (gráfica de la izquierda) como en la función de pérdidas específica de la detección de objetos por segmentación (gráfica de la derecha), se afirma una reducción en la función de pérdida a lo largo de las épocas. En adición, la tendencia observada en estas curvas es indicativo de que se ha seleccionado una tasa de aprendizaje apropiada, como es explicado en la (Figura 8).

En la fase de validación, también se obtienen unas curvas similares que informan acerca del desempeño adecuado o no del modelo (Figura 12).

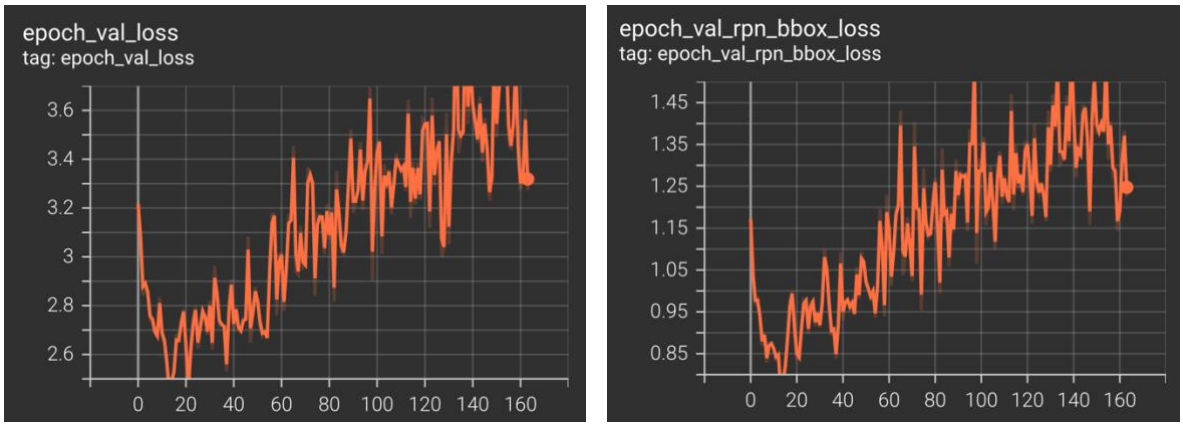

Figura 12: curvas resultantes de la validación. Ambas curvas grafican el parámetro *epoch* o épocas en el eje X frente a la función de pérdida o *loss* en el eje Y. Sin embargo, la de la derecha representa el parámetro de función de pérdida en la eficiencia en la que la red neuronal (RPN, *Region Proposal Networks*) detecta los objetos y los distingue del fondo.

Se observa como la curva de pérdida (*loss*) del *set* de entrenamiento va bajando progresivamente (Figura 11), lo que es indicativo de que el modelo está aprendiendo correctamente. Sin embargo, el *loss* del *set* de validación disminuye aproximadamente hasta el *epoch* 20 aproximadamente, y a partir de ahí vuelve a subir (Figura 12). Esta tendencia significa que se está produciendo un caso de *overfitting*, es decir, se ha sobreentrenado con respecto al *training set*. Entonces, el modelo en lugar de aprender los patrones, está aprendiendo características concretas de los ejemplos del *training set*. Como consecuencia de ello, cada vez generaliza peor en la validación y las curvas aumentan. Este es un comportamiento que se quiere evitar.

Idealmente las curvas de entrenamiento y validación deben disminuir al mismo tiempo, en el momento en el que esta armonía se altere, se debe detener el entrenamiento para que el modelo no empeore. Por tanto, esta es la medida que se ha tomado. El entrenamiento se ha parado antes de tiempo y se va a emplear un modelo entrenado hasta el *epoch* 21 (justo antes de que suceda el *overfitting*).

4.2.2. TEST.

Una vez ajustados los parámetros y obtenido el modelo definitivo, se prueba en su conjunto de imágenes nunca vistas y reservadas para esta función (Figura 13). De esta forma, se obtienen las métricas de la estadística cuantitativa que permiten evaluar el rendimiento y precisión del modelo de segmentación (Tabla 6).

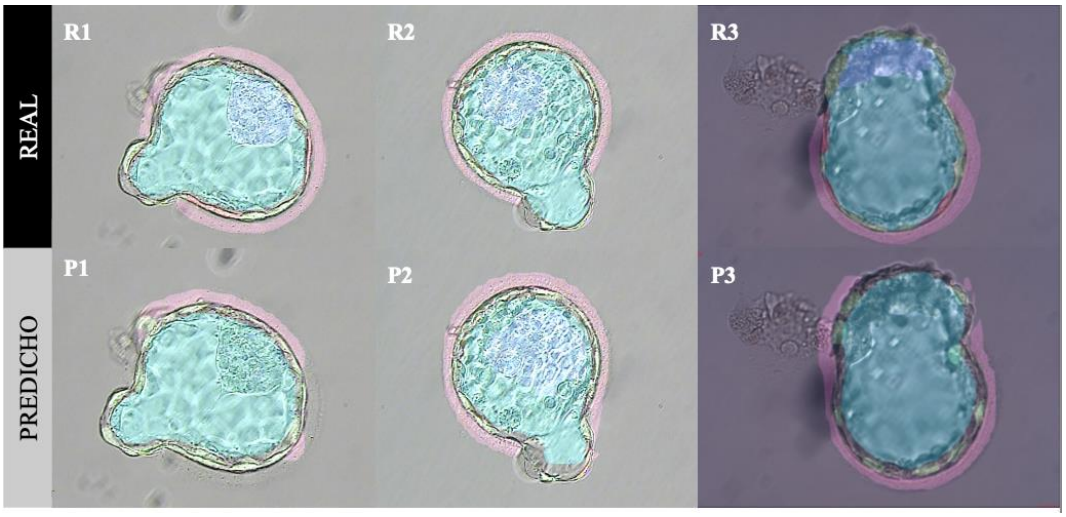

Figura 13: Ejemplo de comparación entre máscara de segmentación predicha (P) por el modelo en el test y la segmentación realizada manualmente considerada como real o *ground truth* (R) en imágenes de embriones de este *set* de datos.

Tabla 6: Resultados de las métricas del test del modelo de segmentación. A la izquierda se recoge en esta tabla la cuantificación de los verdaderos positivos (TP), falsos positivos (FP) y falsos negativos (FN) de todo el conjunto. Mientras que a la derecha se encuentran los resultados de las métricas del modelo de segmentación en su totalidad expresados sobre 1.

GLOBALMENTE			
Métrica	**Cuantificación**	**Métrica**	**Resultado sobre 1**
TP	352	*Precision*	0,265
FP	976	*Recall*	0,200
FN	109	*F1 – score*	0,228
		IoU	0,381
		Model Confidence	0,959

De forma visual en la (Figura 13), parece observarse que el modelo predice y segmenta los diferentes componentes de los blastocistos con bastante certeza. No obstante, en la (Tabla 6) que incluye los valores de las métricas obtenidas por el modelo en general, se puede comprobar que la eficiencia no es tan buena. Una precisión de valor 0,265 sugiere que el modelo posee un rendimiento bajo clasificando de forma precisa. Un valor de *Recall* de 0,200 indica una baja sensibilidad. Al igual que una puntuación de F1 de 0,228 también demuestra un rendimiento deficiente, ya que un valor de 1 sería lo óptimo. En los parámetros de IoU, se estableció un umbral de 0,5; lo que implicaba que en caso de que la detección se solape más del umbral con el objeto real, quiere decir la detección ha sido exitosa. En este caso se ha obtenido un valor de 0,381, cercano a 0,5 pero continúa por debajo demostrando dificultades por el modelo en determinar una superposición significativa. Finalmente, tras haber impuesto como umbral de confianza 0,7; el haber obtenido un valor de 0,959 significa que el modelo muestra un alto nivel de seguridad en sus predicciones.

Las causas de la obtención de estos resultados no tan satisfactorios pueden ser varias. En primer lugar, el pequeño tamaño de la base de datos de la que se ha partido; lo que genera que algunas características no se muestren con prácticamente ninguna frecuencia como se observa en la estadística descriptiva (Tabla 5). Además, todas las etiquetas tienen el mismo peso en el modelo, aunque unas sean más importantes que otras, y este hecho tampoco se ha tenido en cuenta.

En segundo lugar, la calidad de las fotografías mediante las cuales se desarrolla el estudio es importante. Una mejor calidad, que las características estén controladas y que más balanceada se encuentre la base de datos; permiten la obtención de mejores resultados. No obstante, cuanto más regulado se encuentre, más se aleja de la realidad de la práctica rutinaria de las clínicas y del objetivo principal de estos tipos de estudios; que es el acercar el uso de la IA a todos los laboratorios de embriología. Por tanto, el haber empleado un *dataset* multicéntrico tan variante tiene sus consecuencias reflejadas en peores resultados, pero con mayor utilidad real.

Entonces, se está comentando los resultados del modelo global que incluye todas las características: tanto las más comunes y fáciles de detectar como podría ser el blastocele; como las más complicadas de detectar en los embriones, ya sea por ejemplo porque su aparición es opcional en los blastocistos como el caso de las *strings*. Por tanto, a la hora de obtener las métricas globales de la fusión de los resultados de todas las *features*, parece que el rendimiento del modelo se ve mermado. Mientras que si se observan estas mismas métricas en cada una de las características por separado se puede extraer conclusiones más favorables (Figura 14) (Anexo II). Una posible solución para poder evaluar correctamente todas las características, incluso las que se encuentran en menor proporción, sería aumentando la base de datos con más imágenes de embriones.

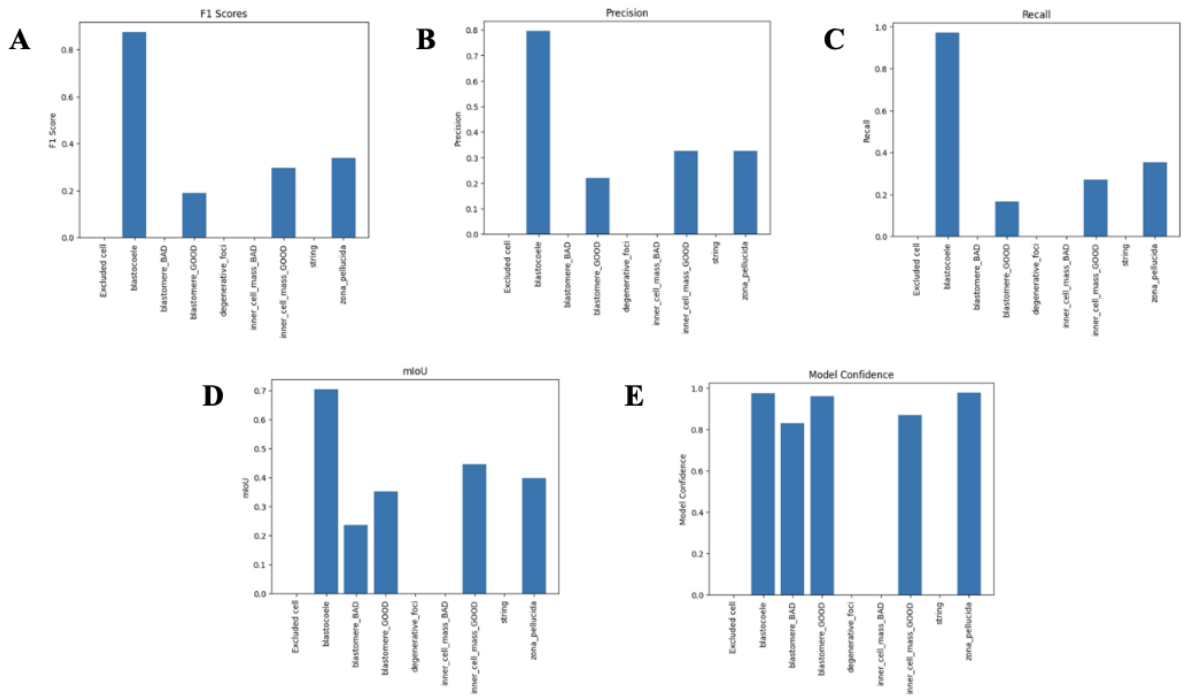

Figura 14: Gráficas de barras que resumen y comparan los resultados de cada *feature* para cada métrica. Las características incluidas en el eje X son en este orden: células excluidas, blastocele, blastómero pésimo, blastómero bueno, zona degenerada, MCI pésima, MCI buena, *string* y zona pelúcida. Mientras que en

el eje Y se encuentran las métricas: A) puntuación de F1, B) precisión, C) *recall*, D) media de IoU y E) confianza del modelo.

Comparando entre características de la (Figura 14), se puede afirmar que el blastocele es aquel componente de los blastocistos que el modelo detecta con la mayor precisión, sensibilidad, superposición, etc. En cambio, hay otras características como las células excluidas o las zonas degeneradas que, al haber sido anotadas en muy baja proporción, no poseen resultados. Destaca también como la confianza de predicción es alta en casi todas las características, mientras que no sucede lo mismo con la precisión y la sensibilidad. Todo ello alega la necesidad de un futuro estudio más completo y ampliado que incremente la precisión del modelo obtenido.

4.3. MODELO DE CALIFICACIÓN O *GRADING* EMBRIONARIO.

Mediante los resultados de la segmentación, el modelo tabular determinó las calidades de los embriones (A, B, C, D o E) del *set* de test final, así como las métricas de evaluación (Tabla 7).

Tabla 7: Resultados finales del modelo de calificación embrionaria. Incluyen la cuantificación de las diferentes etiquetas de calidad impuestas a los 81 embriones fotografiados del conjunto de datos de prueba. También los valores de las métricas (precisión, *recall* y la puntuación F1) para cada tipo de calidad y las medias. Se obtienen dos tipos de medias diferentes de las métricas: la *weighted average* es la media que tiene en cuenta la proporción de los elementos de cada grupo en el conjunto de datos; mientras que *la macro average*, se calcula sin ponderar y sin tener en cuenta el no balanceo de las clases. Finalmente, como un anexo a la tabla se visualiza el valor de la exactitud o *accuracy* del conjunto total de test de 80 embriones (ignorando el descartado).

Calidad de los embriones del test *set*	**Frecuencia de aparición de cada tipo de etiqueta de calificación: Número de veces [Porcentaje]**	**Precisión**	***Recall***	***F1-score***
Calidad A	33 [40,74 %]	0,78	0,64	0,70
Calidad B	32 [39,51 %]	0,77	0,62	0,69
Calidad C	15 [18,52 %]	0,67	0,67	0,67
Calidad D	0 [0,00 %]	0,00	0,00	0,00
Calidad E	0 [0,00 %]	0,00	0,00	0,00
Descartado	1 [1,23 %]	-	-	-
macro average		0,44	0,39	0,41
weighted average		0,75	0,64	0,69
accuracy			0,64	

En cuanto a la cuantificación de la (Tabla 7) se destaca que, en este *dataset* de imágenes de test, no hay ninguna correspondiente a la calidad D. Debido a que, en todo el conjunto de 891 imágenes originales,

solo 2 de ellas correspondían a esta calidad (Tabla 4). Entonces, a la hora de la división aleatoria de las imágenes en entrenamiento y test, este segundo conjunto no ha recibido ningún ejemplo de este tipo D. El mayor número que los laboratorios de embriología esperan tener es de A y B, ya que son los que más probabilidades de éxito poseen. Además, otro motivo por el cual la mayoría de las imágenes eran de las calificaciones óptimas es que, si un embrión es capaz de realizar la eclosión, se considera un indicio positivo de su desarrollo y potencial de implantación, por tanto, generalmente suelen tener mejor calidad. Por otro lado, una imagen fue descartada por estar duplicada, por tanto, se considerarán 80 imágenes en este conjunto y no 81.

Son las métricas de evaluación las que permiten determinar el rendimiento y la eficiencia del modelo desarrollado. En esta situación donde se observa un claro desequilibrio en el *dataset*, adquiere más sentido el valor de *weighted average***,** para poder tener en cuenta la importancia relativa de cada clase. De esta forma, un valor de exactitud en conjunto de 0,64 es un resultado intuitivo acerca del desempeño del modelo porque indica el porcentaje de clasificaciones realizadas de forma apropiada del total. Se considera intuitivo, pero no exacto, debido a que le influye el desbalance de las categorías de la base de datos. Por otro lado, el resultado final de precisión indica que el modelo es capaz de clasificar correctamente el 75% de las muestras positivas. Una sensibilidad o *recall* de 0,64 muestra una buena capacidad de identificar correctamente aquello que se busca. Finalmente, la puntuación F1 es una combinación de la precisión y el *recall*. El valor obtenido de 0,69 significa que el modelo tiene un buen desempeño en la clasificación de las muestras positivas, al tiempo que minimiza los falsos positivos y los falsos negativos. Entonces, el modelo de DL generado posee un rendimiento aceptable ya que todas las métricas globales superan el 50%.

Considerando los datos del conjunto de test, se ha obtenido la matriz de confusión original y normalizada (Figura 15). En ellas, los recuadros de más intensidad de color (diagonal principal) muestran los aciertos de la clasificación embrionaria tomando como la realidad la etiquetación manual realizada, es decir, los verdaderos positivos y negativos. Mientras que el resto de los recuadros de color más claro indican los errores, ósea, los falsos positivos y falsos negativos. De esta forma se puede comparar entre la predicción realizada por la IA y la manual supervisada por una embrióloga experta.

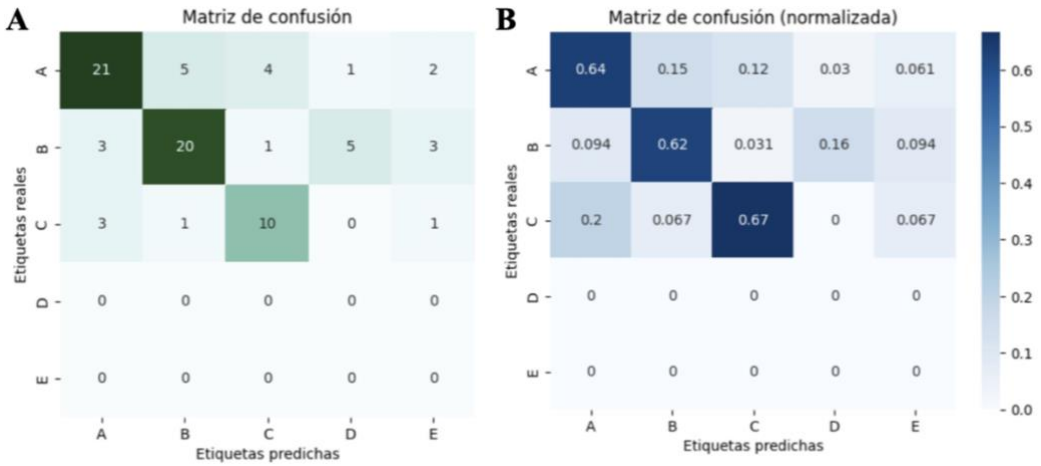

Figura 15: Matrices de confusión del modelo de *grading* embrionario sobre los resultados del conjunto de prueba final. A) Matriz de confusión original que representa el número de elementos. B) Matriz de confusión normalizada que contiene la distribución de los valores de acierto y error sobre 1.

En la matriz de confusión normalizada (Figura 15; B), se puede observar cómo en más de un 50% en todos los casos (A, B y C) la predicción del modelo coincide con la manual (cuadrados de azul oscuro). Los casos de fallo más grave serían cuando el modelo predice una calidad de C cuando era una A (y viceversa). Estos casos han ocurrido en un 12% y 20% respectivamente; que en número de imágenes según la matriz original (Figura 15; A) se tratarían de 7 en total de las 80 ilustraciones. Las razones del fallo en la predicción pueden ser debidas a la calidad de algunas de las fotografías que dificulta la detección de los componentes y a que el sistema visualice características en distinto plano focal que le lleven a conclusiones erróneas. Además, los porcentajes se acentúan porque el tamaño de muestra es reducido, aunque realmente solo se han dado los casos más graves en un bajo número de ilustraciones. Entonces, como estas son en una menor proporción, las matrices indican un rendimiento aceptable de la IA.

En adición se han obtenido las curvas ROC y los valores de las áreas bajo las curvas (AUC) para cada tipo de calidad (Figura 16).

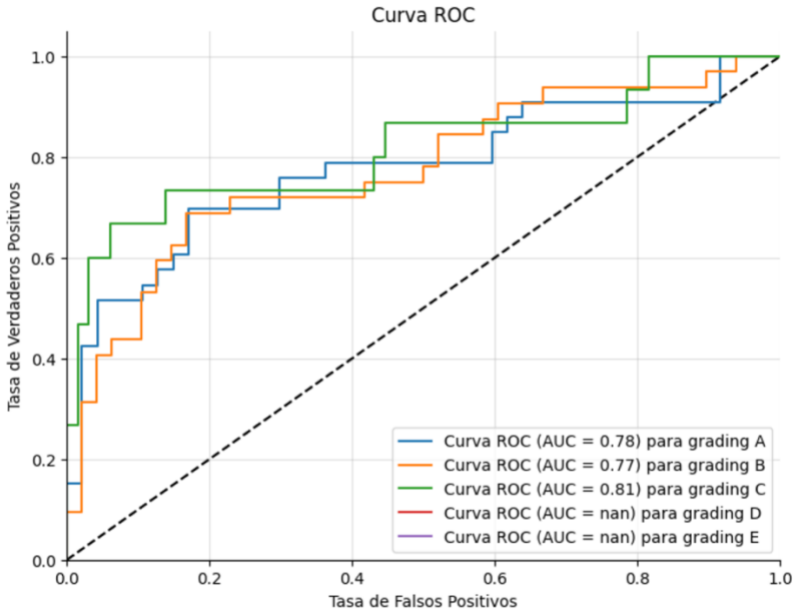

Figura 16: Gráfica representativa de las curvas ROC de cada clase de calidad de blastocisto: A, B y C. En ella no se representan curvas ROC para la calidad D y E pues no estaban presentes en la base de datos de test. Cada curva sigue la leyenda de color y esta contiene los valores del área bajo la curva.

Las curvas ROC se pueden emplear como resumen del rendimiento del modelo. Cuanto más a la izquierda se encuentre la curva mejor clasificador será, debido a que menor tasa de falsos positivos habrá y mayor de verdaderos positivos. Las curvas para las tres calidades son parecidas con un ligero desplazamiento hacia la izquierda de la curva ROC del *grading* C que indicaría que es la etiqueta mejor clasificada. Esto también se ve reflejado en los valores del área bajo la curva (AUC), ya que el *grading* C posee el valor más alto. No obstante, los tres valores son parecidos y su promedio es de 0,787. Este valor permite considerar al modelo como un buen clasificador.

5.DISCUSIÓN.

El empleo de la IA de procesado de imágenes en el ámbito de las TRA proporciona diversas ventajas. Da sentido a datos complejos, puede detectar características que los embriólogos no son capaces de visualizar; estandarizar y automatizar procesos como el de la selección de embriones, reducir la variabilidad de dicha elección y, por tanto, proporcionar una decisión más objetiva.

El desarrollo de la inteligencia artificial en este ámbito se encuentra todavía en fases tempranas. Un artículo muy crítico con ello es el de Afnan y su equipo. En este se plantean los riesgos que podrían conllevar el empleo de la IA que toma decisiones en base a datos de imágenes en medicina reproductiva y califica a los modelos como "*black-boxes*", lo que significa que en ocasiones se desconoce la razón de sus decisiones o recomendaciones. Por tanto, a pesar de todo el avance que se está dando en el campo, siempre habrá que tener en cuenta que estas tecnologías únicamente son una ayuda complementaria a la función que realiza el embriólogo; puesto que el empleo de únicamente la decisión de la inteligencia artificial, genera una brecha de responsabilidad en la toma de decisiones (Afnan et al., 2021) .

Generalmente, este tipo de inteligencias artificiales se evalúan de dos maneras: mediante el resultado de embarazo o por comparación con las decisiones del embriólogo que ha tomado de forma manual. En este caso se trataría de la segunda, pues se desconocían los datos de nacidos vivos o no de los embriones empleados. Además, el modelo se ha generado como una prueba de concepto para determinar las calidades de los blastocistos en estadio *hatching*. La calidad y el éxito de la transferencia se encuentran estrechamente relacionados, pero existen una gran cantidad de factores que influyen en la gestación, como podrían ser las condiciones fisiológicas de la receptora, que no son predecibles por la calidad del blastocisto. En nuestro caso, debido a la obtención de un resultado de precisión de 0,75 ; sensibilidad de 0,64 y puntuación F1 de 0,69; se podría decir que se ha desarrollado un modelo capaz de segmentar y clasificar con una eficiencia considerable en comparación al procesado por parte humano.

Un modelo desarrollado con imágenes estáticas de 1231 embriones que se podría considerar eficiente es ERICA (Chavez-Badiola, Flores-Saiffe-Farías, et al., 2020). En él obtienen resultados de *accuracy* de 0,7; a diferencia de los resultados del presente trabajo de 0,64. La diferencia puede ser debida a que emplearon mayor número de embriones, 1231 frente a los 891. Además, Chavez-Badiola y su equipo emplearon condiciones concretas para el cultivo de los embriones y obtención de las fotos, como son: que todos los ovocitos fueron fecundados con ICSI, que se aplicaron ciertos filtros a las imágenes, las cuales fueron tomadas con los mismos microscopios, etc. Esto les permite obtener mejores resultados pues todas las características están controladas y establecidas; condiciones que este tipo de métrica que se ve influenciada por la prevalencia de las categorías, necesita para proporcionar un resultado que no sea únicamente intuitivo. Sin embargo, ello hace que su modelo sea menos reproducible entre las clínicas ya que deberían seguir sus restricciones para poder emplearlo.

Este hecho se denomina (Afnan et al., 2021) como "*buying-into*", en el que las compañías de IA en cierta medida obligarían a las clínicas a comprar los productos sobre los que fue entrenado exactamente el modelo para no afectar a los resultados. Entonces, se recomienda el desarrollo de modelos más robustos como en cierta medida se podría considerar al trabajo presentado. En este caso, el *set* de imágenes proviene de diferentes clínicas, es totalmente randomizado y no se ha filtrado ninguna fotografía. Por ende, el contar con una base de datos más limitada y la no estandarización de las imágenes son causas que contribuyen a la obtención de peores resultados en comparación. No obstante, se podrían considerar más realistas para la práctica clínica diaria.

El DL generado por el grupo de Chen (Chen et al., 2019) es destacable por los resultados obtenidos de las curvas ROC. Su estudio se basa únicamente en tres características: TF, MCI y expansión del blastocisto. Las áreas bajo la curva para cada una se encuentran entre el intervalo de valores de 0,89 a 1. Ello sería indicativo de un rendimiento de clasificación del modelo excelente a diferencia de nuestro AUC obtenido de 0,787. La razón de su éxito puede ser debida a que únicamente analizan esas tres características y no otras, que en ocasiones en nuestro modelo penalizan los resultados. Además, de que su *dataset* contenía 171.239 imágenes de 16.201 embriones, lo que demuestra que un incremento de ejemplos en la base de datos mejora los resultados obtenidos. No obstante, a su vez, se destapa uno de los inconvenientes de las IA, como es la necesidad de una base de datos masiva.

Otro ejemplo es el de Sawada *et al.* (2021), en el que también preestablecen las condiciones de la base de datos y pese a que emplean menor número de embriones (470), poseen mayor número de imágenes (144.444) ya que provienen de *Time Lapse*. El empleo de incubadores que toman fotografías a tiempo real proporciona la ventaja de que se conoce el momento y el modo de división del embrión, lo que afecta al resultado de la implantación como se está demostrando en diversos artículos como (Armstrong et al., 2019; Serrano-Novillo et al., 2023). Sin embargo, para el desarrollo del *Machine Learning* únicamente se emplean las imágenes estáticas. Por tanto, un posible desarrollo futuro podría ser la inclusión de los vídeos generados por estos incubadores que aportan más información morfocinética. Aunque pueda generar cierta controversia puesto que no todos los laboratorios pueden permitírselos. Por ello, la flexibilidad de adaptación a diferentes instrumentos y tipo de imágenes que aporta el modelo desarrollado en este trabajo es un buen avance favorable para que los sistemas de IA se puedan establecer en todas las clínicas.

El modelo presentado podría considerarse bueno y la ventaja indudable es que su uso en la selección de embriones es un proceso no invasivo con todos los beneficios que ello conlleva. Sin embargo, será necesario un estudio a mayor escala en el que se aumente el tamaño de muestra, se incluyan más clínicas participantes y se aumenten los datos conocidos de cada embrión, entre otros aspectos.

6.CONCLUSIÓN.

En conclusión, nuestro modelo de DL posee resultados que pueden ser considerados favorables en términos de métricas de rendimiento. Además, el uso de imágenes de embriones sin manipulación brinda un enfoque más auténtico y fiel a las condiciones reales. En el presente estudio de 891 embriones, se ha conseguido segmentar y calificar sus componentes con un grado de certeza considerable. Este hecho es ventajoso puesto que el modelo se está desarrollando frente a los desafíos y complejidades que se encontrarán en la práctica clínica diaria y no frente a situaciones ideales. Aunque se requieran mejoras en el rendimiento, en especial de la segmentación, se ha llevado a cabo un progreso importante por el desarrollo de un modelo robusto y confiable para la tarea del análisis y elección del embrión en estadio *hatching* en diferentes centros de medicina reproductiva. Futuras mejoras son planteadas para incrementar la precisión de este sistema de IA, como el aumento de la base de datos, la inclusión de datos de *Time Lapse*, el empleo de estudios clínicos prospectivos, la incorporación de más clínicas participantes y/o la prueba de un modelo diferente, ya que no hay un solo método de IA correcto o uno que se ajuste perfectamente en todos los casos.

En adición, pocos modelos se han desarrollado para el estudio del momento de eclosión del blastocisto humano que es un proceso considerable a la hora del éxito de la transferencia, lo que aporta a este estudio un carácter novedoso en comparación con los trabajos previos publicados.

Con los rápidos avances que están sucediendo en el campo de la inteligencia artificial, es importante comprender la amplia gama de posibilidades que puede brindar tanto para las TRA como para cualquier otro ámbito. Lo que permite no solo ayudar en el alcance de los objetivos de la OMS planteados en el trabajo (Anexo I), sino también en los 15 restantes.

7. REFERENCIAS BIBLIOGRÁFICAS.

7.1. GENERAL.

Adamson, G. D., de Mouzon, J., Chambers, G. M., Zegers-Hochschild, F., Mansour, R., Ishihara, O., Banker, M., & Dyer, S. (2018). International Committee for Monitoring Assisted Reproductive Technology: world report on assisted reproductive technology, 2011. *Fertility and Sterility*, *110*(6), 1067–1080. https://doi.org/10.1016/j.fertnstert.2018.06.039

Afnan, M. A. M., Liu, Y., Conitzer, V., Rudin, C., Mishra, A., Savulescu, J., & Afnan, M. (2021). Interpretable, not black-box, artificial intelligence should be used for embryo selection. *Human Reproduction Open*, *2021*(4). https://doi.org/10.1093/hropen/hoab040

Ahmed Fawzy Gad. (2020). *Evaluating Object Detection Models Using Mean Average Precision (mAP)*. https://blog.paperspace.com/mean-average-precision/

Armstrong, S., Bhide, P., Jordan, V., Pacey, A., & Farquhar, C. (2019). Time-lapse systems for embryo incubation and assessment in assisted reproduction. *Cochrane Database of Systematic Reviews*.

Balaban, B., Brison, D., Calderon, G., Catt, J., Conaghan, J., Cowan, L., Ebner, T., Gardner, D., Hardarson, T., Lundin, K., Cristina Magli, M., Mortimer, D., Mortimer, S., Munne, S., Royere, D., Scott, L., Smitz, J., Thornhill, A., van Blerkom, J., & Van den Abbeel, E. (2011). The Istanbul consensus workshop on embryo assessment: proceedings of an expert meeting. *Human Reproduction*, *26*(6), 1270–1283. https://doi.org/10.1093/humrep/der037

Baştanlar, Y., & Özuysal, M. (2014). *Introduction to Machine Learning* (pp. 105–128). https://doi.org/10.1007/978-1-62703-748-8_7

Carlson, B. M. (2014). *Embriología humana y Biología del desarrollo* (Elsevier, Ed.; 5th ed.).

Chavez-Badiola, A., Flores-Saiffe Farias, A., Mendizabal-Ruiz, G., Garcia-Sanchez, R., Drakeley, A. J., & Garcia-Sandoval, J. P. (2020). Predicting pregnancy test results after embryo transfer by image feature extraction and analysis using machine learning. *Scientific Reports*, *10*(1), 4394. https://doi.org/10.1038/s41598-020-61357-9

Chavez-Badiola, A., Flores-Saiffe-Farías, A., Mendizabal-Ruiz, G., Drakeley, A. J., & Cohen, J. (2020). Embryo Ranking Intelligent Classification Algorithm (ERICA): artificial intelligence clinical assistant predicting embryo ploidy and implantation. *Reproductive BioMedicine Online*, *41*(4), 585–593. https://doi.org/10.1016/j.rbmo.2020.07.003

Chen, T.-J., Zheng, W.-L., Liu, C.-H., Huang, I., Lai, H.-H., & Liu, M. (2019). Using Deep Learning with Large Dataset of Microscope Images to Develop an Automated Embryo Grading System. *Fertility & Reproduction*, *01*(01), 51–56. https://doi.org/10.1142/S2661318219500051

Cuevas Saiz, I., Carme Pons Gatell, M., Vargas, M. C., Delgado Mendive, A., Rives Enedáguila, N., Moragas Solanes, M., Carrasco Canal, B., Teruel López, J., Busquets Bonet, A., & Hurtado de Mendoza Acosta, M. V. (2018). The Embryology Interest Group: updating ASEBIR's morphological scoring system for early embryos, morulae and blastocysts. *Medicina Reproductiva y Embriología Clínica*, *5*(1), 42–54. https://doi.org/10.1016/j.medre.2017.11.002

De Geyter, C. (2019). Assisted reproductive technology: Impact on society and need for surveillance. *Best Practice & Research Clinical Endocrinology & Metabolism*, *33*(1), 3–8. https://doi.org/10.1016/j.beem.2019.01.004

Eastick, J., Venetis, C., Cooke, S., & Chapman, M. (2021). The presence of cytoplasmic strings in human blastocysts is associated with the probability of clinical pregnancy with fetal heart. *Journal of Assisted Reproduction and Genetics*, *38*(8), 2139–2149. https://doi.org/10.1007/s10815-021-02213-1

Edwards, R. G., Purdy, J. M., Steptoe, P. C., & Walters, D. E. (1981). The growth of human preimplantation embryos in vitro. *American Journal of Obstetrics and Gynecology*, *141*(4), 408–416. https://doi.org/10.1016/0002-9378(81)90603-7

Farias, A. F.-S., Chavez-Badiola, A., Mendizabal-Ruiz, G., Valencia-Murillo, R., Drakeley, A., Cohen, J., & Cardenas-Esparza, E. (2023). Automated identification of blastocyst regions at different development stages. *Scientific Reports*, *13*(1), 15. https://doi.org/10.1038/s41598-022-26386-6

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Fernandez, E. I., Ferreira, A. S., Cecílio, M. H. M., Chéles, D. S., de Souza, R. C. M., Nogueira, M. F. G., & Rocha, J. C. (2020). Artificial intelligence in the IVF laboratory: overview through the application of different types of algorithms for the classification of reproductive data. *Journal of Assisted Reproduction and Genetics*, *37*(10), 2359–2376. https://doi.org/10.1007/s10815-020-01881-9

Fong, C.-Y., Bongso, A., Sathananthan, H., Ho, J., & Ng, S.-C. (2001). Ultrastructural observations of enzymatically treated human blastocysts: zona-free blastocyst transfer and rescue of blastocysts with hatching difficulties. *Human Reproduction*, *16*(3), 540–546. https://doi.org/10.1093/humrep/16.3.540

Gardner, D. K., Lane, M., Stevens, J., Schlenker, T., & Schoolcraft, W. B. (2000). Blastocyst score affects implantation and pregnancy outcome: towards a single blastocyst transfer. *Fertility and Sterility*, *73*(6), 1155–1158. https://doi.org/10.1016/S0015-0282(00)00518-5

Gerri, C., Menchero, S., Mahadevaiah, S. K., Turner, J. M. A., & Niakan, K. K. (2020). Human Embryogenesis: A Comparative Perspective. *Annual Review of Cell and Developmental Biology*, *36*(1), 411–440. https://doi.org/10.1146/annurev-cellbio-022020-024900

Glujovsky, D., Quinteiro Retamar, A. M., Alvarez Sedo, C. R., Ciapponi, A., Cornelisse, S., & Blake, D. (2022). Cleavage-stage versus blastocyst-stage embryo transfer in assisted reproductive technology. *Cochrane Database of Systematic Reviews*, *2022*(6). https://doi.org/10.1002/14651858.CD002118.pub6

Greener, J. G., Kandathil, S. M., Moffat, L., & Jones, D. T. (2022). A guide to machine learning for biologists. *Nature Reviews Molecular Cell Biology*, *23*(1), 40–55. https://doi.org/10.1038/s41580-021-00407-0

Haddad, M., Stewart, J., Xie, P., Cheung, S., Trout, A., Keating, D., Parrella, A., Lawrence, S., Rosenwaks, Z., & Palermo, G. D. (2021). Thoughts on the popularity of ICSI. *Journal of Assisted Reproduction and Genetics*, *38*(1), 101–123. https://doi.org/10.1007/s10815-020-01987-0

Hammadeh, M. E., Fischer-Hammadeh, C., & Ali, K. R. (2011). Assisted hatching in assisted reproduction: a state of the art. *Journal of Assisted Reproduction and Genetics*, *28*(2), 119–128. https://doi.org/10.1007/s10815-010-9495-3

Harun, M. Y., Huang, T., & Ohta, A. T. (2019). Inner Cell Mass and Trophectoderm Segmentation in Human Blastocyst Images using Deep Neural Network. *2019 IEEE 13th International Conference*

on Nano/Molecular Medicine & Engineering (NANOMED), 214–219. https://doi.org/10.1109/NANOMED49242.2019.9130618

He, K., Gkioxari, G., Dollar, P., & Girshick, R. (2017). Mask R-CNN. *2017 IEEE International Conference on Computer Vision (ICCV)*, 2980–2988. https://doi.org/10.1109/ICCV.2017.322

Isa, I. S., Yusof, U. K., & Mohd Zain, M. (2023). Image Processing Approach for Grading IVF Blastocyst: A State-of-the-Art Review and Future Perspective of Deep Learning-Based Models. *Applied Sciences*, *13*(2), 1195. https://doi.org/10.3390/app13021195

Kragh, M. F., & Karstoft, H. (2021). Embryo selection with artificial intelligence: how to evaluate and compare methods? *Journal of Assisted Reproduction and Genetics*, *38*(7), 1675–1689. https://doi.org/10.1007/s10815-021-02254-6

Mark Everingham, & John Winn. (2012, May 21). *The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Development Kit*. http://host.robots.ox.ac.uk/pascal/VOC/voc2012/htmldoc/devkit_doc.html

McQuin, C., Goodman, A., Chernyshev, V., Kamentsky, L., Cimini, B. A., Karhohs, K. W., Doan, M., Ding, L., Rafelski, S. M., Thirstrup, D., Wiegraebe, W., Singh, S., Becker, T., Caicedo, J. C., & Carpenter, A. E. (2018). CellProfiler 3.0: Next-generation image processing for biology. *PLOS Biology*, *16*(7), e2005970. https://doi.org/10.1371/journal.pbio.2005970

Naciones Unidas. (2015, September 25). *Transformar nuestro mundo: la Agenda 2030 para el Desarrollo Sostenible*. https://unctad.org/system/files/official-document/ares70d1_es.pdf

Ombelet, W., Deblaere, K., Bosmans, E., Cox, A., Jacobs, P., Janssen, M., & Nijs, M. (2003). Semen quality and intrauterine insemination. *Reproductive BioMedicine Online*, *7*(4), 485–492. https://doi.org/10.1016/S1472-6483(10)61894-9

P. Saeedi, D. Yee, J. Au, & J. Havelock. (2017). Automatic Identification of Human Blastocyst Components via Texture. *IEEE Transactions on Biomedical Engineering*, *64*(12), 2968–2978. https://doi.org/10.1109/TBME.2017.2759665

PALERMO, G. (1992). Pregnancies after intracytoplasmic injection of single spermatozoon into an oocyte. *The Lancet*, *340*(8810), 17–18. https://doi.org/10.1016/0140-6736(92)92425-F

Rad, R. M., Saeedi, P., Au, J., & Havelock, J. (2019). BLAST-NET: Semantic Segmentation of Human Blastocyst Components via Cascaded Atrous Pyramid and Dense Progressive Upsampling. *2019 IEEE International Conference on Image Processing (ICIP)*, 1865–1869. https://doi.org/10.1109/ICIP.2019.8803139

Radke, K. L., Kors, M., Müller-Lutz, A., Frenken, M., Wilms, L. M., Baraliakos, X., Wittsack, H.-J., Distler, J. H. W., Abrar, D. B., Antoch, G., & Sewerin, P. (2022). Adaptive IoU Thresholding for Improving Small Object Detection: A Proof-of-Concept Study of Hand Erosions Classification of Patients with Rheumatic Arthritis on X-ray Images. *Diagnostics*, *13*(1), 104. https://doi.org/10.3390/diagnostics13010104

Redacción KeepCoding. (2023, May 3). *Learning rate en Deep Learning*. https://keepcoding.io/blog/learning-rate-en-deep-learning/

Rocha, J. C., Passalia, F. J., Matos, F. D., Takahashi, M. B., Ciniciato, D. de S., Maserati, M. P., Alves, M. F., de Almeida, T. G., Cardoso, B. L., Basso, A. C., & Nogueira, M. F. G. (2017). A Method Based on Artificial Intelligence To Fully Automatize The Evaluation of Bovine Blastocyst Images. *Scientific Reports*, *7*(1), 7659. https://doi.org/10.1038/s41598-017-08104-9

Rochon, M. (1986). [Sterility and infertility: two concepts]. *Cahiers Quebecois de Demographie*, *15*(1), 27–56.

Sawada, Y., Sato, T., Nagaya, M., Saito, C., Yoshihara, H., Banno, C., Matsumoto, Y., Matsuda, Y., Yoshikai, K., Sawada, T., Ukita, N., & Sugiura-Ogasawara, M. (2021). Evaluation of artificial intelligence using time-lapse images of IVF embryos to predict live birth. *Reproductive BioMedicine Online*, *43*(5), 843–852. https://doi.org/10.1016/j.rbmo.2021.05.002

Serrano-Novillo, C., Uroz, L., & Márquez, C. (2023). Novel Time-Lapse Parameters Correlate with Embryo Ploidy and Suggest an Improvement in Non-Invasive Embryo Selection. *Journal of Clinical Medicine*, *12*(8), 2983. https://doi.org/10.3390/jcm12082983

Seshagiri, P. B., Vani, V., & Madhulika, P. (2016). Cytokines and Blastocyst Hatching. *American Journal of Reproductive Immunology*, *75*(3), 208–217. https://doi.org/10.1111/aji.12464

Shahbazi, M. N. (2020). Mechanisms of human embryo development: from cell fate to tissue shape and back. *Development*, *147*(14). https://doi.org/10.1242/dev.190629

Storr, A., Venetis, C. A., Cooke, S., Kilani, S., & Ledger, W. (2017). Inter-observer and intra-observer agreement between embryologists during selection of a single Day 5 embryo for transfer: a multicenter study. *Human Reproduction*, *32*(2), 307–314. https://doi.org/10.1093/humrep/dew330

Swain, J., VerMilyea, M. T., Meseguer, M., Ezcurra, D., Ezcurra, D., Letterie, G., Sánchez, P., Trew, G., Swain, J., Meseguer, M., Nayot, D., Campbell, A., Huangv, I., Choma, J., Loewke, K., Piqueras, M. P., Nader, P., Schindler, M., Lippolis, E., … Abshagen, D. (2020). AI in the treatment of fertility: key considerations. *Journal of Assisted Reproduction and Genetics*, *37*(11), 2817–2824. https://doi.org/10.1007/s10815-020-01950-z

Teh, W.-T., McBain, J., & Rogers, P. (2016). What is the contribution of embryo-endometrial asynchrony to implantation failure? *Journal of Assisted Reproduction and Genetics*, *33*(11), 1419–1430. https://doi.org/10.1007/s10815-016-0773-6

The Vienna consensus: report of an expert meeting on the development of ART laboratory performance indicators. (2017). *Reproductive BioMedicine Online*, *35*(5), 494–510. https://doi.org/10.1016/j.rbmo.2017.06.015

Vander Borght, M., & Wyns, C. (2018). Fertility and infertility: Definition and epidemiology. *Clinical Biochemistry*, *62*, 2–10. https://doi.org/10.1016/j.clinbiochem.2018.03.012

Wang, R., Pan, W., Jin, L., Li, Y., Geng, Y., Gao, C., Chen, G., Wang, H., Ma, D., & Liao, S. (2019). Artificial intelligence in reproductive medicine. *Reproduction*, *158*(4), R139–R154. https://doi.org/10.1530/REP-18-0523

7.2. REFERENCIAS DE LAS FIGURAS.

Figura 1: Cuevas Saiz, I., Carme Pons Gatell, M., Vargas, M. C., Delgado Mendive, A., Rives Enedáguila, N., Moragas Solanes, M., Carrasco Canal, B., Teruel López, J., Busquets Bonet, A., & Hurtado de Mendoza Acosta, M. V. (2018). The Embryology Interest Group: updating ASEBIR's morphological scoring system for early embryos, morulae and blastocysts. *Medicina Reproductiva y Embriología* Clínica, 5(1), 42–54. https://doi.org/10.1016/j.medre.2017.11.002

Figura 4: Wang, R., Pan, W., Jin, L., Li, Y., Geng, Y., Gao, C., Chen, G., Wang, H., Ma, D., & Liao, S. (2019). Artificial intelligence in reproductive medicine. *Reproduction*, *158*(4), R139–R154. https://doi.org/10.1530/REP-18-0523

Figura 8: Redacción KeepCoding. (2023, May 3). *Learning rate en Deep Learning*. https://keepcoding.io/blog/learning-rate-en-deep-learning/

Figura 9: Kragh, M. F., & Karstoft, H. (2021). Embryo selection with artificial intelligence: how to evaluate and compare methods? *Journal of Assisted Reproduction and Genetics*, *38*(7), 1675–1689. https://doi.org/10.1007/s10815-021-02254-6

Figura 10: Ahmed Fawzy Gad. (2020). *Evaluating Object Detection Models Using Mean Average Precision (mAP)*. https://blog.paperspace.com/mean-average-precision/

ANEXO.

ANEXO I: RELACIÓN DEL TRABAJO CON LOS OBJETIVOS DE DESARROLLO SOSTENIBLE DE LA AGENDA 2030.

Tabla I.1: Grado de relación del trabajo con los Objetivos de Desarrollo Sostenible (ODS).

Objetivos de Desarrollo Sostenibles	Alto	Medio	Bajo	No procede
ODS 1. Fin de la pobreza.				X
ODS 2. Hambre cero.				X
ODS 3. Salud y bienestar.	X			
ODS 4. Educación de calidad.				X
ODS 5. Igualdad de género.				X
ODS 6. Agua limpia y saneamiento.				X
ODS 7. Energía asequible y no contaminante.				X
ODS 8. Trabajo decente y crecimiento económico.			X	
ODS 9. Industria, innovación e infraestructuras.			X	
ODS 10. Reducción de las desigualdades.		X		
ODS 11. Ciudades y comunidades sostenibles.				X
ODS 12. Producción y consumo responsables.				X
ODS 13. Acción por el clima.				X
ODS 14. Vida submarina.				X
ODS 15. Vida de ecosistemas terrestres.				X
ODS 16. Paz, justicia e instituciones sólidas.				X
ODS 17. Alianzas para lograr objetivos.				X

El presente estudio contribuye al alcance de alguno de los objetivos y metas de desarrollo sostenible de la Organización Mundial de la Salud (OMS) (Naciones Unidas, 2015). El ejemplo más claro sería el objetivo de Salud y Bienestar (ODS 3); el cual pretende "garantizar una vida sana y promover el bienestar en todas las edades". La razón de ello reside en que modelos de IA como el que se ha desarrollado, cooperan en la selección del mejor embrión que conlleve el éxito de la técnica de reproducción asistida; lo que permitirá la consecución de un bienestar físico y mental en las personas

que desean concebir un hijo y requieren de estas técnicas para llevarlo a cabo. En adición, la IA secunda una personalización en los tratamientos, ya que los algoritmos de *Machine Learning* pueden adaptarse y aprender de los datos específicos de cada paciente y tratamiento, lo que permite una individualización y optimización de los protocolos de fecundación *in vitro*.

Otro ejemplo podría ser la Reducción de las Desigualdades (ODS 10), porque si gracias a la selección de embriones mediante una IA se pueden aumentar las tasas de éxito de embarazos, se da la posibilidad a ciertos colectivos a cumplir su deseo de ser padres. Dentro de estos colectivos se pueden encontrar personas estériles o infértiles, madres solteras con cualquier orientación sexual, etc.

ANEXO II: MÉTRICAS DEL MODELO DE SEGMENTACIÓN PARA CADA UNA DE LAS CARACTERÍSTICAS SEGMENTADAS EN LOS EMBRIONES.

Tabla II.1: Resultados de las métricas del test del modelo de segmentación únicamente para la característica de las células excluidas. A la izquierda se recoge en esta tabla la cuantificación de los verdaderos positivos (TP), falsos positivos (FP) y falsos negativos (FN) de todo el conjunto. Mientras que a la derecha se encuentran los resultados de las métricas del modelo de segmentación expresados sobre 1.

CÉLULAS EXCLUIDAS			
Métrica	**Cuantificación**	**Métrica**	**Resultado sobre 1**
TP	0	*Precision*	0,000
FP	0	*Recall*	0,000
FN	9	*F1 – score*	0,000
		IoU	0,000
		Model Confidence	0,000

Tabla II.2: Resultados de las métricas del test del modelo de segmentación únicamente para la característica del blastocele. A la izquierda se recoge en esta tabla la cuantificación de los verdaderos positivos (TP), falsos positivos (FP) y falsos negativos (FN) de todo el conjunto. Mientras que a la derecha se encuentran los resultados de las métricas del modelo de segmentación expresados sobre 1.

BLASTOCELE			
Métrica	**Cuantificación**	**Métrica**	**Resultado sobre 1**
TP	66	*Precision*	0,795
FP	17	*Recall*	0,971
FN	2	*F1 – score*	0,874
		IoU	0,703
		Model Confidence	0,975

Tabla II.3: Resultados de las métricas del test del modelo de segmentación únicamente para la característica de los blastómeros de pésima calidad. A la izquierda se recoge en esta tabla la cuantificación de los verdaderos positivos (TP), falsos positivos (FP) y falsos negativos (FN) de todo el conjunto. Mientras que a la derecha se encuentran los resultados de las métricas del modelo de segmentación expresados sobre 1.

BLASTÓMERO PÉSIMO			
Métrica	**Cuantificación**	**Métrica**	**Resultado sobre 1**
TP	0	*Precision*	0,000
FP	5	*Recall*	0,000
FN	73	*F1 – score*	0,000
		IoU	0,235
		Model Confidence	0,823

Tabla II.4: Resultados de las métricas del test del modelo de segmentación únicamente para la característica de los blastómeros de buena calidad. A la izquierda se recoge en esta tabla la cuantificación de los verdaderos positivos (TP), falsos positivos (FP) y falsos negativos (FN) de todo el conjunto. Mientras que a la derecha se encuentran los resultados de las métricas del modelo de segmentación expresados sobre 1.

BLASTÓMERO BUENO			
Métrica	**Cuantificación**	**Métrica**	**Resultado sobre 1**
TP	242	*Precision*	0,219
FP	863	*Recall*	0,166
FN	1212	*F1 – score*	0,189
		IoU	0,352
		Model Confidence	0,960

Tabla II.5: Resultados de las métricas del test del modelo de segmentación únicamente para la característica de zonas degeneradas en el embrión. A la izquierda se recoge en esta tabla la cuantificación de los verdaderos positivos (TP), falsos positivos (FP) y falsos negativos (FN) de todo el conjunto. Mientras que a la derecha se encuentran los resultados de las métricas del modelo de segmentación expresados sobre 1.

ZONA DEGENERADA			
Métrica	**Cuantificación**	**Métrica**	**Resultado sobre 1**
TP	0	*Precision*	0,000
FP	0	*Recall*	0,000
FN	1	*F1 – score*	0,000
		IoU	0,000
		Model Confidence	0,000

Tabla II.6: Resultados de las métricas del test del modelo de segmentación únicamente para la característica de la MCI de pésima calidad. A la izquierda se recoge en esta tabla la cuantificación de los verdaderos positivos (TP), falsos positivos (FP) y falsos negativos (FN) de todo el conjunto. Mientras que a la derecha se encuentran los resultados de las métricas del modelo de segmentación expresados sobre 1.

MCI PÉSIMA			
Métrica	**Cuantificación**	**Métrica**	**Resultado sobre 1**
TP	0	*Precision*	0,000
FP	0	*Recall*	0,000
FN	8	*F1 – score*	0,000
		IoU	0,000
		Model Confidence	0,000

Tabla II.7: Resultados de las métricas del test del modelo de segmentación únicamente para la característica de la MCI de buena calidad. A la izquierda se recoge en esta tabla la cuantificación de los verdaderos positivos (TP), falsos positivos (FP) y falsos negativos (FN) de todo el conjunto. Mientras que a la derecha se encuentran los resultados de las métricas del modelo de segmentación expresados sobre 1.

MCI BUENA			
Métrica	**Cuantificación**	**Métrica**	**Resultado sobre 1**
TP	13	*Precision*	0,325
FP	27	*Recall*	0,271
FN	35	*F1 – score*	0,295
		IoU	0,447
		Model Confidence	0,870

Tabla II.8: Resultados de las métricas del test del modelo de segmentación únicamente para la característica de las *string*. A la izquierda se recoge en esta tabla la cuantificación de los verdaderos positivos (TP), falsos positivos (FP) y falsos negativos (FN) de todo el conjunto. Mientras que a la derecha se encuentran los resultados de las métricas del modelo de segmentación expresados sobre 1.

STRING			
Métrica	**Cuantificación**	**Métrica**	**Resultado sobre 1**
TP	0	*Precision*	0,000
FP	0	*Recall*	0,000
FN	12	*F1 – score*	0,000
		IoU	0,000
		Model Confidence	0,000

Tabla II.9: Resultados de las métricas del test del modelo de segmentación únicamente para la característica de la zona pelúcida. A la izquierda se recoge en esta tabla la cuantificación de los verdaderos positivos (TP), falsos positivos (FP) y falsos negativos (FN) de todo el conjunto. Mientras que a la derecha se encuentran los resultados de las métricas del modelo de segmentación expresados sobre 1.

ZONA PELÚCIDA			
Métrica	**Cuantificación**	**Métrica**	**Resultado sobre 1**
TP	31	*Precision*	0,326
FP	64	*Recall*	0,352
FN	57	*F1 – score*	0,339
		IoU	0,398
		Model Confidence	0,978