



Universidad de Valladolid



**PROGRAMA DE DOCTORADO EN
INVESTIGACIÓN EN CIENCIAS DE LA SALUD**

TESIS DOCTORAL:

**IDENTIFICACIÓN DE PACIENTES CON UN MAYOR
RIESGO DE PEOR EVOLUCIÓN CLÍNICA TRAS LA
INFECCIÓN POR SARS-COV-2 MEDIANTE EL
DISEÑO DE ALGORITMOS BASADOS EN LA
VARIABILIDAD GENÉTICA**

Presentada por Dña. Sofía Jaurrieta Largo
para optar al Grado de
Doctora por la Universidad de Valladolid

Dirigida por:
Dr. Ricardo Usategui Martín
Dr. José Pablo Miramontes González

Para mi madre, Pilar, por ser mi ejemplo.

Para Javier, por no dudar y ser mi apoyo.

Para mi hijo, que aún sin estar presente me ha enseñado lo realmente importante.

Agradecimientos

Me gustaría agradecer a todos aquellos, que de forma directa o indirecta, me han ayudado a realizar esta tesis doctoral.

En concreto a mis directores y tutor, Ricardo, Pablo y José Luis, por su paciencia durante este tiempo y su confianza depositada, aún cuando ni yo misma confiaba en poder conseguirlo.

A mi familia, por enseñarme a lo largo de la vida que todo esfuerzo tiene su recompensa.

A Javier, por ser siempre un refugio y darme lo mejor de esta vida, que sin querer me ha acompañado durante los meses de escritura de esta tesis.

Y por último, pero no menos importante, a todos los pacientes que han formado parte de este estudio y hecho posible esta investigación.

Financiación

Este estudio ha sido financiado por la Gerencia Regional de Salud de Castilla y León, España (GRS 2255/A/20 y GRS COVID91/A/20). También ha recibido el apoyo del Instituto de Tecnología de la Universidad de Castilla-La Mancha, la Cátedra de Inteligencia Artificial (patrocinada por Bayer) y el Instituto de Investigación Sanitaria de Castilla-La Mancha (IDISCAM).

Índice

RESUMEN	1
INTRODUCCIÓN	3
2.1. SITUACIÓN DEL SARS-CoV2	4
2.1.1. Mecanismos de transmisión.....	4
2.1.2. Virología.....	5
2.2. RESPUESTA INMUNE	6
2.3. MANIFESTACIONES CLÍNICAS.....	7
2.3.1. Posibles complicaciones.....	9
2.3.2. Factores de riesgo.....	11
2.3.3. Mortalidad	12
2.4.1. Sistema renina-angiotensina-aldosterona o SRAA	13
2.4.2. Metabolismo de la vitamina D.....	14
2.4.3. Respuesta hiperinflamatoria.....	17
2.5. INFLUENCIA DE LA GENÉTICA EN LA ENFERMEDAD COVID-19	19
2.6. LA INTELIGENCIA ARTIFICIAL EN LA MEDICINA.....	21
HIPÓTESIS Y OBJETIVOS.....	25
MATERIAL Y MÉTODO.....	27
4.1. PACIENTES.....	28
4.2. AISLAMIENTO DE ADN DE ALTO PESO MOLECULAR.....	29
4.3. DISCRIMINACIÓN ALÉLICA MEDIANTE PCR CON SONDAS TAQMAN.....	30
4.4. ANÁLISIS MEDIANTE APRENDIZAJE AUTOMÁTICO	32
4.4.1. Selección de modelo de aprendizaje automático	32
4.4.2. Fase de entrenamiento.....	36
4.4.3. Métricas para analizar y comparar los algoritmos	38
4.4.4. Selección del modelo y validación	39
4.5. ASPECTOS ÉTICOS	40
RESULTADOS	41

5.1. RIESGO DE NEUMONÍA POR COVID-19	49
5.2. RIESGO DE MORTALIDAD	53
5.3. RIESGO DE REINGRESO DURANTE EL PRIMER AÑO TRAS COVID-19.....	56
5.4. RIESGO DE MORTALIDAD DURANTE EL REINGRESO EN EL PRIMER AÑO TRAS COVID-19	58
 DISCUSIÓN.....	62
 CONCLUSIONES	69
 BIBLIOGRAFÍA	71
 ANEXOS	84

Índice Figuras

Figura 1. Estructura general del virus SARS-CoV-2 mostrando las cuatro proteínas estructurales: proteína Spike (S), proteína de membrana (M), proteína de envoltura (E) y proteína de nucleocápside (N). Imagen adaptada de: *Maguiña Vargas et al. (2020)*(17)

Pág. 6

Figura 2. Corte axial de una tomografía computarizada del tórax en el que se observa una neumonía por SARS-CoV-2, imagen típica en “vidrio deslustrado” (indicada con asteriscos). Imagen adaptada de: Wu, Jiong MS *et al. (2020)*(30)

Pág. 10

Figura 3. Mecanismo de entrada del SARS-CoV-2. La proteína S del virus reconoce y se une al receptor del ECA2; de esta forma la proteína S es procesada por TMPRSS2 para producirse la fusión de ambas membranas y así entrar en la célula huésped. Imagen adaptada de: Pastrian-Soto Gabriel (2020) (48).

Pág. 13

Figura 4: vía de la vitamina D y sus efectos en la respuesta inmune innata y adquirida.

Pág. 16

Figura 5. Esquema en el que se detalle el mecanismo de entrada del virus SARS-CoV-2 en los neumocitos tipo II gracias al receptor de la ECA2. Tras la infección de la célula huésped se produce un aumento de citoquinas proinflamatorias (IL-1 β , IL-6, IL-8, factor estimulante de colonias de granulocitos y macrófagos o GM-CSF, TNF- α , IFN- γ). Posteriormente se produce un reclutamiento de monocitos inflamatorios, linfocitos T CD4⁺ y CD8⁺, neutrófilos y células NK en el intersticio pulmonar favoreciendo la respuesta inflamatoria. Este estado hiperinflamatorio consiste en una tormenta de citoquinas que se asocia al desarrollo de SDRA. Imagen adaptada de: Gustine JN *et al. (2021)* (59).

Pág. 18

Figura 6. Ejemplo de representación de los resultados de la discriminación alélica con sondas TaqMan. Cada nube de puntos representa uno de los genotipos posibles: los azules son homocigotos, los verdes son heterocigotos y los rojos son homocigotos mutados.

Pág. 31

Figura 7: Pasos del algoritmo de aprendizaje automático.

Pág. 37

Figura 8. Polimorfismos genéticos relacionados con el riesgo de neumonía COVID-19.

Pág. 50

Figura 9. Comparativa de las curvas de ROC de los diferentes modelos de aprendizaje automático evaluados en la predicción de riesgo de neumonía por COVID-19. Pág. 52

Figura 10. Diagrama de radar de la fase de entrenamiento y fase de prueba. Pág. 53

Figura 11. Polimorfismos genéticos relacionados con la mortalidad por COVID-19.

Pág. 54

Figura 12. Comparativa de las curvas de ROC de los diferentes modelos de aprendizaje automático evaluados en la predicción de riesgo de mortalidad. Pág. 55

Figura 13. Diagrama de radar de la fase de entrenamiento y fase de prueba. Pág. 55

Figura 14. Polimorfismos genéticos relacionados con el reingreso en el primer año tras COVID-19. Pág. 56

Figura 15. Comparativa de las curvas de ROC de los diferentes modelos de aprendizaje automático evaluados en la predicción de riesgo de reingreso tras un año de COVID-19.

Pág. 57

Figura 16. Diagrama de radar de la fase de entrenamiento y fase de prueba. Pág. 58

Figura 17. Polimorfismos genéticos relacionados con la mortalidad durante el reingreso en el primer año tras COVID-19. Pág. 59

Figura 18. Comparativa de las curvas de ROC de los diferentes modelos de aprendizaje automático evaluados en la predicción de riesgo de mortalidad durante el reingreso tras un año de COVID-19. Pág. 60

Figura 19. Diagrama de radar de la fase de entrenamiento y fase de prueba. Pág. 60

Índice Tablas

Tabla 1. Síntomas más frecuentes en el COVID-19.	Pág. 8
Tabla 2. Características generales de los pacientes del estudio.	Pág. 43
Tabla 3. Características clínicas de los pacientes del estudio.	Pág. 44
Tabla 4. Características analíticas de los pacientes del estudio.	Pág. 45
Tabla 5. Tratamientos recibidos por los pacientes del estudio.	Pág. 46
Tabla 6. Distribución genotípica según el riesgo de neumonía por COVID-19, mortalidad, reingreso y mortalidad asociada al reingreso.	Pág. 47
Tabla 6 continuación 1. Distribución genotípica según el riesgo de neumonía por COVID-19, mortalidad, reingreso y mortalidad asociada al reingreso.	Pág. 48
Tabla 6 continuación 2. Distribución genotípica según el riesgo de neumonía por COVID-19, mortalidad, reingreso y mortalidad asociada al reingreso.	Pág. 49
Tabla 7. Resultados de los diferentes métodos de aprendizaje automático utilizados para evaluar el riesgo de neumonía por COVID-19.	Pág. 51
Tabla 8. Resultados de los diferentes métodos de aprendizaje automático utilizados para evaluar el riesgo de mortalidad.	Pág. 54
Tabla 9. Resultados de los diferentes métodos de aprendizaje automático utilizados para evaluar el riesgo de reingreso en el primer año tras COVID-19.	Pág. 57
Tabla 10. Resultados de los diferentes métodos de aprendizaje automático utilizados para evaluar el riesgo de mortalidad durante el reingreso en el primer año tras COVID-19.	Pág. 59

RESUMEN

En la evolución clínica de la enfermedad por coronavirus (COVID-19), la genética constituye un determinante clave. El presente estudio tuvo como objetivo evaluar la incidencia de polimorfismos en genes asociados con el sistema renina-angiotensina-aldosterona (SRAA), la producción de citoquinas del sistema inflamatorio y la vía metabólica de la vitamina D, en relación con la gravedad de la infección por SARS-CoV-2. Para ello, se analizó una cohorte conformada por 338 pacientes con diagnóstico confirmado de COVID-19, empleando algoritmos de aprendizaje automático con el propósito de identificar las variantes genéticas con mayor impacto sobre el pronóstico clínico de la enfermedad.

Los hallazgos obtenidos revelan que los polimorfismos localizados en los genes *IL6*, *IL6R*, *IL1 α* , *IL1R*, *IFN- γ* , *TNF- α* , *CRP*, *VDR*, *VDBP* y *ECA2* representan los factores genéticos más relevantes en la modulación de la gravedad clínica de la COVID-19. Estas variantes genéticas se asociaron de manera significativa con un mayor riesgo de desarrollar neumonía por COVID-19, incremento de la mortalidad, reingreso hospitalario en el siguiente año y mortalidad relacionada con dicho reingreso. Los modelos de aprendizaje automático utilizados alcanzaron un área bajo la curva (AUC) de 0,86 para la predicción de neumonía, mortalidad y mortalidad asociada al reingreso hospitalario, así como un AUC de 0,85 para la predicción de reingreso durante el primer año postinfección. Estos resultados refuerzan la hipótesis de que el perfil genético individual desempeña un papel determinante en el pronóstico de la COVID-19 y permiten la estratificación de pacientes con mayor susceptibilidad clínica.

En conclusión, este estudio evidencia que los modelos predictivos basados en aprendizaje automático y dirigidos por información genética constituyen una herramienta eficaz para identificar individuos con riesgo aumentado de evolución adversa, siendo especialmente relevantes las variantes genéticas relacionadas con el gen *ECA2*, el sistema inflamatorio y el metabolismo de la vitamina D.

INTRODUCCIÓN

2.1. Situación del SARS-CoV-2

2.1.1. Mecanismos de transmisión

2.1.2. Virología

2.2. Respuesta inmune

2.2.1. Inmunidad celular.

2.2.1. Inmunidad humoral.

2.3. Manifestaciones clínicas

2.3.1. Posibles complicaciones.

2.3.2. Factores de riesgo.

2.3.3. Mortalidad

2.4. Vías metabólicas relacionadas con el desarrollo de COVID-19

2.4.1. Sistema renina-angiotensina-aldosterona o SRAA.

2.4.2. Vitamina D

2.4.3. Respuesta hiperinflamatoria

2.5. Influencia de la genética en la enfermedad COVID-19

2.6. Evolución de la inteligencia artificial en medicina.

2.1. Situación del SARS-CoV2

Los coronavirus son virus que afectan a los humanos de manera diferente; son responsables de hasta un tercio de las infecciones respiratorias de vía aérea superior en adultos, así como de enfermedades respiratorias infecciosas graves en niños y adultos. Dentro de esta familia se encuentra el coronavirus del síndrome respiratorio agudo severo 2 o SARS-CoV-2, que es un *Betacoronavirus* responsable de la enfermedad por coronavirus 2019 (COVID-19) (1). Esta enfermedad, iniciada en diciembre de 2019, es una infección respiratoria severa con una importante morbilidad y mortalidad que ha supuesto un cambio en cuanto al manejo y respuesta a nivel mundial frente a este tipo de infecciones (2). Según datos de la Organización Mundial de la Salud (OMS), la pandemia de COVID-19, declarada el 11 de marzo de 2020, resultó en 777 millones de casos y 7 millones de muertes en todo el mundo (3). El fin de la emergencia sanitaria fue establecido el 5 de mayo de 2023. En España el número de casos confirmados se contabilizó en casi 14 millones, con cerca de 121.760 fallecidos por esta causa. El inicio del programa de vacunación a finales de 2020 supuso un cambio en la evolución de los contagios e impacto en la mortalidad. Según datos de la OMS, entre el periodo de diciembre de 2020 y marzo de 2023 en Europa, el número de fallecidos por COVID-19 se redujo en un 59% (aproximadamente 1,6 millones de vidas), siendo mayores de 60 años el 96% de estos (4).

2.1.1. Mecanismos de transmisión

La principal vía de transmisión es por vía respiratoria, aunque no podemos descartar otros medios, y se debe tener en cuenta su elevada tasa de contagio debida a la alta capacidad de replicación en el aparato respiratorio, así como a la estabilidad que presenta en determinadas condiciones ambientales (5).

- Transmisión por vía respiratoria:

Se trata del principal mecanismo de transmisión del virus mediante aerosoles y gotas respiratorias, sobre todo entre individuos que se encuentran en espacios pequeños con poca ventilación. También son frecuentes los contagios mediante esta vía en ambientes con mucha afluencia y mal ventilados, como reuniones grupales o medios de transporte.

- Transmisión por superficies contaminadas:

Es una vía poco común, posible en situaciones en las que una persona entra en contacto con una superficie donde hay virus viables y posteriormente esa parte del cuerpo entra en contacto con diferentes mucosas. Hay ciertos estudios que sugieren la viabilidad del virus durante varios días (aproximadamente 9 días) en superficies (6,7).

- Transmisión por vía feco-oral:

Se ha demostrado la presencia de ARN viral en heces de pacientes con COVID-19 sin poder demostrar que sea causa de transmisión (8).

2.1.2. Virología

Los coronavirus forman parte de la familia Coronaviridae; dentro de esta se encuentra la subfamilia Betacoronavirus a la cual pertenece el SARS-CoV-2. Este *Betacoronavirus* está formado por una cadena sencilla de ARN, con una cápside como envoltura proteica, esférica y espiculada (9). La secuencia de ARN capaz de codificar proteínas, o marco de lectura (ORF1a/b), traduce las proteínas no estructurales (NSP) que formarán la replicasa-transcriptasa viral, responsable de la replicación y transcripción del genoma del virus (10). Además, el genoma expresa cuatro proteínas estructurales esenciales: la glicoproteína de superficie *Spike* (S), la proteína de la *envoltura* (E), la proteína matriz de la *membrana* (M) y la proteína de la *nucleocápside* (N), entre otra serie de proteínas capaces de interferir en el funcionamiento del sistema inmunitario del huésped (11,12).

- Proteína *Spike* (S): Forma las espigas características de la envoltura del virus y es clave para su unión con el receptor de la célula huésped y su fusión con la membrana. Es el principal antígeno que inicia la producción de anticuerpos y es un objetivo para los linfocitos citotóxicos o CD8⁺ (13).
- Proteína de la *envoltura* (E): es esencial para la liberación del virus en el huésped junto con las proteínas de *membrana* y *nucleocápside* (14).
- Proteína de *membrana* (M): se encarga de la unión del virus con la célula del huésped (15).

- Proteína de la *nucleocápside* (N): se encarga de la regulación de la síntesis del ARN viral uniéndose a este (15, 16).

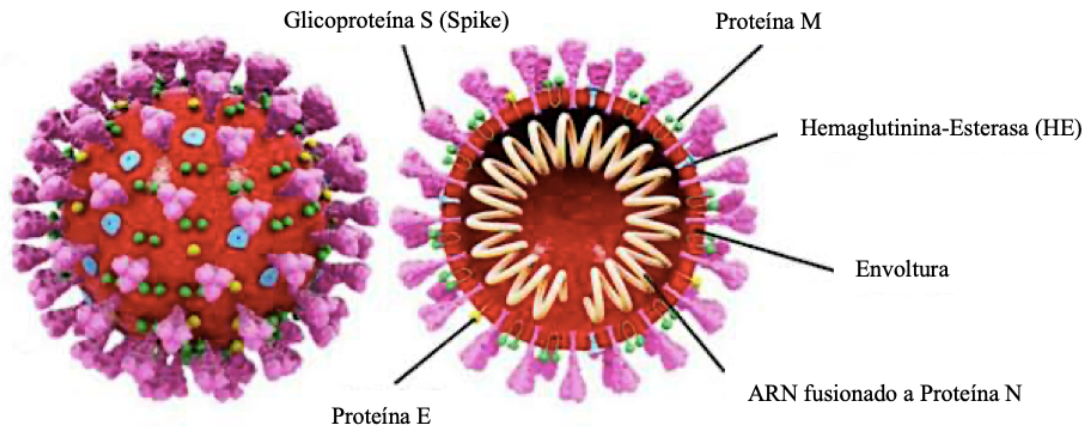


Figura 1. Estructura general del virus SARS-CoV-2 mostrando las cuatro proteínas estructurales: proteína Spike (S), proteína de membrana (M), proteína de envoltura (E) y proteína de nucleocápside (N). Imagen adaptada de: Maguiña Vargas *et al.* (2020)(17)

2.2. Respuesta inmune

Una vez que se produce la infección del huésped, se desencadena en este una respuesta inmunológica, tanto humoral como celular. Por un lado, se activan linfocitos T, $CD4^+$ y $CD8^+$, que favorecen la memoria inmunológica así como la eliminación del virus, limitando su replicación viral mediante la destrucción de células infectadas (18).

Por otro lado, se inicia la formación de anticuerpos neutralizantes que bloquean la capacidad de infectar células del huésped mediante la unión a regiones del virus; esta acción de los anticuerpos puede mantenerse en el tiempo durante varios meses, favoreciendo la protección del huésped de forma prolongada.

- Inmunidad celular: se produce una activación de linfocitos T $CD4^+$ y $CD8^+$ frente al virus que, como se ha expresado previamente, favorece la eliminación del virus y la memoria inmunológica del huésped que podría ayudar frente a exposiciones posteriores al virus y sus variantes. Es importante tener en cuenta, de cara al desarrollo de enfermedad grave por

SARS-CoV-2, que esta respuesta puede estar comprometida en pacientes con alteraciones del sistema inmune celular (19, 20).

- Inmunidad humoral: al inicio de la infección por SARS-CoV-2 se producen anticuerpos frente al dominio de unión al receptor de la proteína S del virus con acción neutralizante frente al virus, estos niveles de anticuerpos descienden de forma progresiva con el tiempo. Hay que tener en cuenta que la respuesta humoral es inversamente proporcional a la gravedad de la enfermedad, es decir, en infecciones leves los niveles de anticuerpos son bajos (21-23). Por otro lado, durante la infección también aumentan las células B de memoria frente a la proteína S del virus con acción en los meses posteriores.

2.3. Manifestaciones clínicas

La enfermedad por COVID-19 presenta gran variabilidad clínica, puede expresarse desde formas asintomáticas hasta casos de desarrollo de enfermedad grave con afectación multisistémica. Desde el momento de la infección, los síntomas suelen aparecer en torno a los 3-5 días, si bien es cierto que el período de incubación puede variar según la persona e incluso llegar a los 14 días tras la exposición (24). Los síntomas más frecuentes son fiebre, tos, astenia, cefalea, dolor de garganta y alteraciones del gusto y el olfato, así como alteraciones gastrointestinales, siendo las más frecuentes la epigastralgia y la diarrea. Es cierto que no hay unos síntomas específicos que nos hagan distinguir la infección por SARS-CoV-2 de otras infecciones respiratorias causadas por otros virus; también hay que tener en cuenta que, conforme avanzó la pandemia y el virus presentó otras variantes, los síntomas más frecuentes variaron. En la Tabla 1 se muestran los síntomas más frecuentes recogidos en un metaanálisis realizado en el 2020 (25).

Tabla 1: Síntomas más frecuentes en la COVID-19.

Síntoma	%	IC 95%
Fiebre	81.2	77.9-84.4
Tos	58.5	54.2-62.8
Fatiga	38.5	30.6-45.3
Disnea	26.1	20.4-31.8
Espujo	25.8	21.1-30.4

A continuación se reflejan las posibles presentaciones de la enfermedad, pudiendo iniciarse en el mismo individuo en una situación clínica y progresar a otra conforme avanza el curso de la enfermedad.

- Casos asintomáticos: este tipo de presentación es el más frecuente en la población joven y sin comorbilidades. Un estudio realizado previo al inicio de la vacunación indicaba que hasta un 33 % de los infectados eran asintomáticos (26); también es cierto que aun no presentando síntomas se pueden presentar alteraciones analíticas o radiológicas en pruebas complementarias.
- Casos leves: en este grupo se incluye la mayoría de los pacientes con COVID-19, los cuales presentan síntomas comunes como tos, fiebre, malestar general, pero sin presentar compromiso respiratorio ni otras alteraciones sistémicas que requieran de medidas hospitalarias.
- Casos moderados: estos pacientes presentan una evolución más tórpida de sus síntomas; por lo general, son pacientes con edad más avanzada y/o que presentan otras enfermedades asociadas. En estos casos sí observamos un compromiso de la vía respiratoria, presentando hipoxemia ($SpO_2 < 94\%$ o $pO_2 < 80$ mmHg). En estos casos, la hospitalización es habitual para el inicio del tratamiento y el control estrecho.

- Casos graves: este grupo de pacientes desarrollan un compromiso respiratorio con insuficiencia respiratoria ($\text{spO}_2 < 90\%$ o $\text{pO}_2 < 60\text{ mmHg}$) y afectación en pruebas de imagen del parénquima pulmonar, así como afectación a otros niveles como compromiso cardiocirculatorio, renal, alteración de la función hepática o shock séptico condicionando un fallo multiorgánico que requiera de hospitalización en Unidades de Cuidados Intensivos (UCI) y medidas invasivas. En este grupo, la mortalidad al inicio de la pandemia en nuestro país llegó a ser del 30 %. Estas tasas de desarrollo de enfermedad grave y mortalidad elevada han disminuido progresivamente gracias al desarrollo de la vacunación y a la expansión de la inmunización de la población.

2.3.1. Posibles complicaciones

- **Respuesta inflamatoria:** en los casos más graves se observa una respuesta inflamatoria acentuada y mantenida en el tiempo con niveles elevados de citoquinas y mala evolución clínica.
- **Respiratorias:** son las más frecuentes y las que en un inicio nos van a guiar sobre la evolución de la enfermedad. El inicio es la neumonía vírica que se puede diagnosticar con pruebas de imagen como radiografía de tórax o tomografía computarizada, observando afectación intersticial en forma de vidrio deslustrado (Figura 2). Sobre estas neumonías víricas pueden producirse sobreinfecciones bacterianas más frecuentes en pacientes hospitalizados (20 %) (27); también se han descrito casos de infecciones fúngicas en pacientes con inmunosupresión previa a la infección (28,29). El síndrome de dificultad respiratoria aguda (SDRA) es otra de las principales complicaciones respiratorias que va a precisar de asistencia en unidades de cuidados intensivos (UCI) así como de ventilación mecánica invasiva (VMI), presentando una elevada mortalidad.

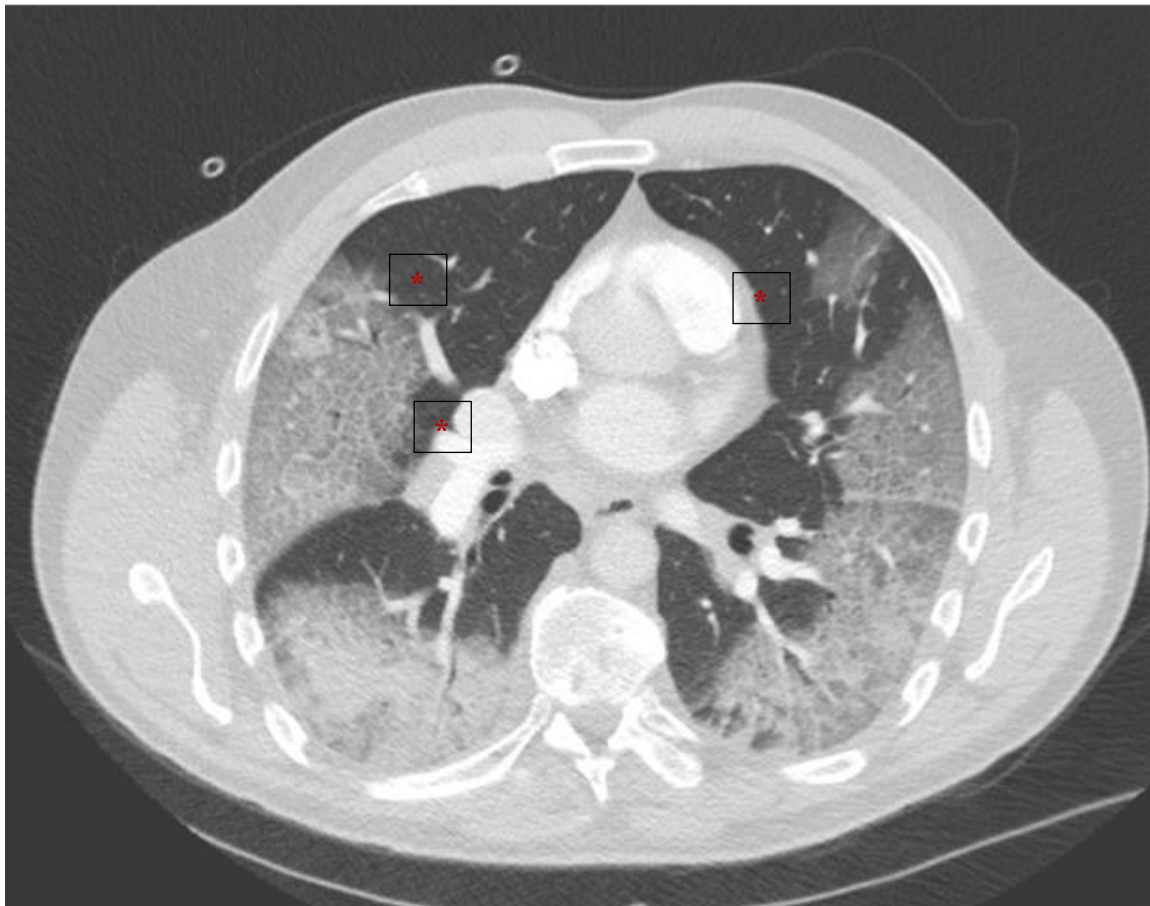


Figura 2. Corte axial de una tomografía computarizada del tórax en el que se observa una neumonía por SARS-CoV-2, imagen típica en “vidrio deslustrado” (indicada con asteriscos). Imagen adaptada de: Wu, Jiong MS *et al.* (2020)(30)

- **Enfermedad tromboembólica venosa:** la aparición de coágulos en pacientes graves es frecuente, siendo la incidencia en la UCI en estos pacientes de hasta 40% (31,32). Esta condición puede generar inestabilidad hemodinámica, así como empeoramiento de la situación respiratoria del paciente.

- **Alteraciones cardiovasculares:** también es frecuente la aparición de arritmias durante la infección, así como el desarrollo de estas de forma posterior durante el seguimiento, condicionando una mala tolerancia al ejercicio en estos pacientes. Otras alteraciones cardiovasculares han sido el desarrollo de insuficiencia cardíaca, más frecuente en pacientes mayores o con otras comorbilidades, y miocarditis, también frecuente en otro tipo de infecciones víricas (33,34).

- **Alteraciones neurológicas:** las más frecuentes son alteraciones de la movilidad, convulsiones y cefalea crónica. Al inicio de la pandemia y en casos graves, se pudo describir el desarrollo de encefalopatía vírica por SARS-CoV-2 en hasta 1/3 de estos pacientes (35). Otras de las complicaciones neurológicas son debidas al desarrollo de coágulos a nivel arterial, produciendo eventos cerebrovasculares isquémicos.
- **Miopatía del paciente crítico:** se trata de una complicación no específica de la infección por SARS-CoV-2, más bien de una complicación debida a ingresos prolongados de este tipo de pacientes, sobre todo en UCI. Precisan posteriormente de una recuperación lenta y progresiva a nivel motor debido a la pérdida de masa muscular secundaria al ingreso prolongado motivado por la gravedad de la infección vírica y a sus complicaciones asociadas. En ciertas ocasiones esta recuperación no puede ser completa, sobre todo en pacientes de edad avanzada o con otro tipo de patologías condicionantes.

2.3.2. Factores de riesgo

La evolución clínica de la COVID-19 es muy variable en cuanto a sintomatología y evoluciona de forma diferente en función de cada individuo debido a los diversos factores asociados así mismo a un peor pronóstico.

- Enfermedades previas han sido identificadas como factores de riesgo para el desarrollo de una forma más grave de COVID-19. Las más frecuentes son hipertensión arterial, obesidad, diabetes mellitus, enfermedades cardiovasculares y respiratorias. Estas patologías, además de asociar una mayor mortalidad, también favorecen la aparición de complicaciones durante la enfermedad (36,37).
- Edad avanzada se relaciona este factor con una mayor probabilidad de desarrollo de enfermedad grave, complicaciones y una mayor mortalidad (38). Hay que tener en cuenta que en muchas ocasiones este factor no es único, sino que va asociado a la presencia de otras enfermedades más frecuentes con la edad avanzada.

- Respuesta inflamatoria o tormenta de citoquinas. La producción de citoquinas inflamatorias y la respuesta inflamatoria exagerada que se producen en algunos individuos se han relacionado con la progresión en estos hacia formas más graves y cuadros críticos, como el SDRA o el fallo multiorgánico (39).
- Factores genéticos. La variabilidad en cuanto al desarrollo y evolución de la enfermedad ha sido justificada en diversos estudios por una determinada predisposición genética. La patogénesis de la enfermedad incluye diversas vías como el sistema renina-angiotensina y la vitamina D, además de la respuesta inflamatoria (40-42).

2.3.3. Mortalidad

Diversos factores influyen en la mortalidad, como las comorbilidades preexistentes (enfermedades cardiovasculares, enfermedades respiratorias, diabetes, alteraciones inmunológicas), la edad o el sexo.

Al inicio de la pandemia, primer semestre del 2020, en nuestro país observamos una tasa de mortalidad muy elevada entre los pacientes hospitalizados, siendo esta de aproximadamente el 16-22 %. Estos datos mejoraron durante el año 2021, reduciéndose la mortalidad en pacientes hospitalizados a un 6-12,5 %, probablemente relacionado con el inicio de la vacunación a finales del año 2020 y durante la primera mitad del año 2021 (43).

2.4. Vías metabólicas relacionadas con el desarrollo y evolución de COVID-19

En el desarrollo de la enfermedad, existen diferentes vías patogénicas que van a condicionar la evolución y el desarrollo de la enfermedad. Así mismo, diversos estudios en los últimos años han ido identificando dentro de estas vías alteraciones genéticas que condicionan una evolución desfavorable de la enfermedad (40,41,44). Las vías más reconocidas en la patogénesis de la enfermedad y en las que hemos centrado este estudio son el sistema renina-angiotensina-aldosterona (SRAA), la vitamina D y la respuesta

hiperinflamatoria.

2.4.1. Sistema renina-angiotensina-aldosterona o SRAA

El SRAA desempeña un papel crucial en la patogénesis; es el sistema hormonal encargado de regular la presión arterial, el ajuste del volumen extracelular, así como el equilibrio electrolítico. La relación del SRAA, en concreto de la enzima convertidora de angiotensina tipo 2 (ECA2) que forma parte de este sistema, y el SARS-CoV-2 se debe a dos cuestiones: por un lado, el mecanismo de infección por el virus; y por otro, el papel regulador de esta enzima durante la activación del SRAA (45).

El SARS-CoV-2 es un virus con gran afinidad por la unión con la ECA2 que se expresa en las células epiteliales de los alvéolos pulmonares, en los endotelios y en los macrófagos. La glicoproteína de *superficie* (S) o *Spike* se une a la mucosa nasofaríngea y a los neumocitos alveolares (más frecuentemente los neumocitos tipo 2) que expresan los receptores ECA2 en su superficie (46,47). Este receptor presenta en una de sus subunidades un dominio susceptible de ser escindido por proteasas como la furina o la TMPRSS2 situadas en la membrana de la célula huésped, provocando la escisión de diferentes formas de la proteína S, motivo por el cual este virus es más infeccioso que otros coronavirus (12). También sabemos que esta unión de la proteína S al receptor ECA2 reduce su expresión y, por lo tanto, favorece la insuficiencia respiratoria grave aguda. Este mecanismo de unión se ve reflejado en la Figura 3.

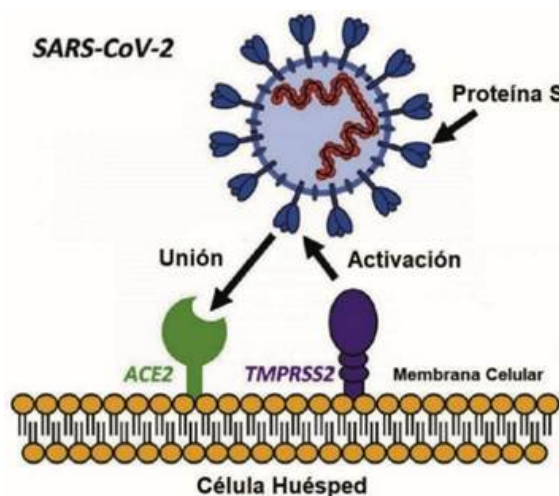


Figura 3. Mecanismo de entrada del SARS-CoV-2. La proteína S del virus reconoce y se une al receptor del ECA2; de esta forma, la proteína S es procesada por TMPRSS2 para producir la fusión de ambas membranas y así entrar en la célula huésped. Imagen adaptada de: Pastian-Soto Gabriel (2020) (48).

El SDRA producido por el SARS-CoV-2 se caracteriza por una lesión pulmonar aguda en la que hay un aumento de la permeabilidad vascular, produciendo edema pulmonar y acúmulo de células inflamatorias en el parénquima pulmonar, resultando en una hipoxia severa. La ECA2, la angiotensina II (Ang-II) y la angiotensina 1-7 tienen funciones clave en la regulación de la fibrosis, la inflamación y la trombosis, modificando así el edema, la permeabilidad y el daño pulmonar durante el SDRA.

La enzima convertidora de angiotensina o ECA provoca la transformación de angiotensina I en angiotensina II, que activa los receptores tipo 1 de la angiotensina II (AT1) y los receptores tipo 2 de la angiotensina II (AT2), favoreciendo la patogénesis de la enfermedad y asociándose a una peor evolución en pacientes con SDRA. La activación del receptor AT1 favorece la vasoconstricción, la inflamación y la fibrosis, disminuyendo la biodisponibilidad del óxido nítrico; frente a la activación del receptor AT2, que causa vasodilatación (49-51). Este último es menos frecuente por lo que la ECA2 debe convertir la angiotensina II, que favorece el aumento de presión hidrostática aumentando el edema y la permeabilidad vascular pulmonar, en angiotensina 1-7 que tiene efectos protectores para el pulmón y así compensar esa menor expresión del receptor AT2 y de esta forma contrarrestar los efectos de la ECA buscando un equilibrio.

Tras definir cómo funciona el SRAA y cómo se produce la unión del SARS-CoV-2 al receptor de la ECA2, disminuyendo su actividad y aumentando la angiotensina II, entendemos cómo se produce el daño pulmonar mediante los mecanismos de inflamación, vasoconstricción y fibrosis, contribuyendo al SDRA.

2.4.2. Metabolismo de la vitamina D

Otro de los mecanismos que influyen en la patogénesis de la COVID-19 es el metabolismo de la vitamina D debido a que actúa como un inmunomodulador. La vitamina D se produce en la piel tras la exposición a rayos ultravioleta B (UVB) provenientes de la luz solar; esta producción es la principal fuente, aunque hay otras formas de adquirirla, como la dieta (productos lácteos o pescado). La melanina influye en la penetración de los rayos UVB en la piel; por eso es conocido que ciertos grupos étnicos que tienen una menor expresión de melanina tienen a su vez una menor síntesis de vitamina D a pesar de una adecuada exposición a los rayos UVB. El déficit de vitamina

D (niveles inferiores a 50 nmol/L) está presente en el 30-60 % y el déficit severo (niveles inferiores a 30 nmol/L) está presente en el 10% de la población de Europa (52,53).

La vitamina D, dentro de la célula, se une al receptor nuclear de la vitamina D (VDR) que va a modular unos determinados genes que influyen tanto en la respuesta inmune innata como en la adquirida.

Entre sus funciones en la regulación de la inmunidad innata se encuentra la expresión de receptores tipo Toll cuya función es reconocer las proteínas patógenas; el aumento de sustancias, como la catelicidina, que favorecen la activación de neutrófilos y macrófagos que degradan la membrana del virus y disminuyen su replicación; las beta-defensinas que destruyen la membrana del virus; y la liberación del óxido nítrico (NO) que contribuye a combatir la infección.

En cuanto a la regulación de la inmunidad adquirida, la vitamina D provoca una supresión de esta respuesta en las células epiteliales respiratorias durante las infecciones virales mediante el descenso de la proliferación de células T y su cambio de células Th1 a Th2. La vitamina D también influye en el desarrollo de los linfocitos Th17, disminuyendo estos a favor de los linfocitos T reguladores (T-reg) con acción antiinflamatoria. Al producirse un descenso de células Th1, esto conlleva unos niveles inferiores de citoquinas proinflamatorias (IL-12,IL-6,IL-8), además de, a su vez, favorecer la expresión de citoquinas antiinflamatorias (IL-4,IL-5,IL-10). Otra vía por la que la vitamina D disminuye la producción de citoquinas proinflamatorias es mediante la inhibición del factor nuclear kappa B (NF- κ B) (Figura 4).

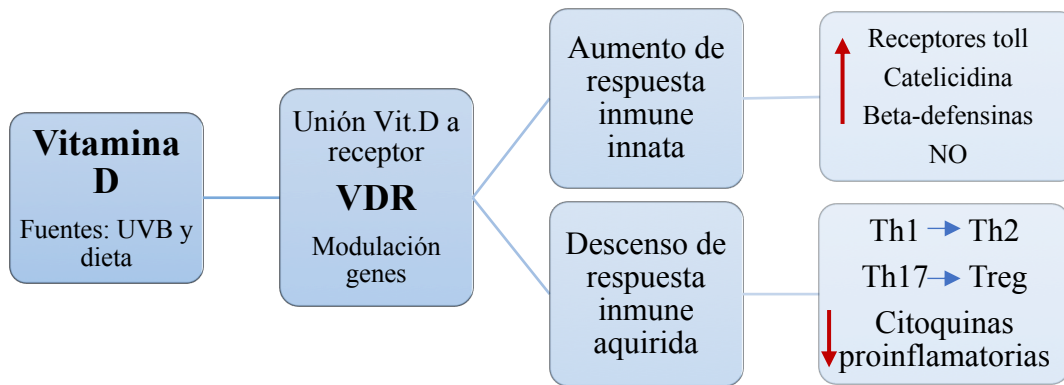


Figura 4. Vía de la vitamina D y sus efectos en la respuesta inmune innata y adquirida.

Durante la infección por SARS-CoV-2, en la respuesta inmune también participa el sistema del complemento, que es un mecanismo de defensa innato y adaptativo cuyas funciones son potenciar la respuesta inflamatoria para así favorecer la destrucción de patógenos. En este caso, tiene una acción dual: favoreciendo la respuesta inflamatoria produciendo anafilotoxinas (C3a y C5a) que aumentan la permeabilidad vascular atrayendo células inmunitarias; y, por otro lado, favoreciendo una mayor expresión del VDR, lo cual aumenta la activación de esta vitamina favoreciendo su acción antiinflamatoria (52).

Se han registrado en diversos estudios niveles de vitamina D en pacientes con COVID-19, determinando que los niveles disminuidos de esta vitamina se relacionan con una peor evolución de la enfermedad y mayor gravedad. Esto se debe a una mayor respuesta inflamatoria bien por déficit de vitamina D o por disfunción de VDR inducida por fallo en el sistema del complemento, lo cual implica una menor producción de citoquinas antiinflamatorias y un aumento de las proinflamatorias (54-56).

Otros estudios apoyan que la vitamina D podría reducir la expresión de receptores ECA2. Como ya se ha revisado previamente, esto es un punto clave durante la infección por SARS-CoV-2, ya que al disminuir estos receptores, el virus tendría menos posibilidades de unirse a la célula huésped, ya que este es su mecanismo de entrada.

Por lo tanto, niveles adecuados de vitamina D reducen esta respuesta inflamatoria y, de esta forma, la gravedad de la enfermedad. Esto ha llevado a diversos grupos de

investigación a considerar, sobre todo en pacientes con comorbilidades como hipertensión arterial o diabetes mellitus debido a su peor pronóstico, la suplementación con vitamina D en pacientes con déficit como parte del tratamiento de la COVID-19 (52,53,57).

2.4.3. Respuesta hiperinflamatoria

Durante la evolución de la COVID-19 se produce una alteración del sistema inmune del huésped desencadenando un aumento de la respuesta inflamatoria produciendo una tormenta de citoquinas que se relaciona con la gravedad de la enfermedad; esta respuesta se caracteriza por un aumento de interleucinas (IL-2, IL-7), interferón- γ , factor de necrosis tumoral- α (TNF- α), y factor estimulante de colonias de granulocitos entre otros. Esta situación hiperinflamatoria se asemeja al síndrome de activación macrofágica que se produce por la linfohistiocitosis hemofagocítica secundaria (sHLH), se trata de un síndrome hiperinflamatorio ocasionado con frecuencia por infecciones virales, y entre el 3,7 % y el 4,3 % de los casos de sepsis, que cursa, entre otros aspectos, con fiebre elevada, citopenia, aumento de ferritina y afectación pulmonar severa como el SDRA. En un estudio realizado por Bryce *et al.*, en el cual estudiaron las autopsias de 67 pacientes fallecidos por COVID-19, se identificaron macrófagos hemofagocíticos en bazo, ganglios linfáticos, médula ósea y corazón que, junto con el aumento de citoquinas, apoyan la teoría del desarrollo de linfocitosis hemofagocítica en pacientes con COVID-19 grave (58).

Este aumento de citoquinas inflamatorias recluta macrófagos activados y células T a la zona infectada, estudios han demostrado que en el parénquima pulmonar se observa durante la infección por SARS-CoV-2, un aumento del infiltrado inflamatorio mononuclear consistente en linfocitos CD4⁺ y CD8⁺ así como macrófagos, estableciéndose un ciclo de aumento progresivo de la inflamación y por consiguiente del daño tisular.

Otro factor que apoya esta respuesta inflamatoria es la afinidad del SARS-CoV-2 por la ECA2, como se ha descrito previamente, que se presenta principalmente en los neumocitos tipo 2, los cuales desencadenan mayor producción de citoquinas. Otros

estudios han revelado que las células *natural killer* (NK) se encuentran reducidas en los pacientes que tienen COVID-19 grave, no contribuyendo a la tormenta de citoquinas; aunque en contraposición hay teorías que respaldan el reclutamiento de estas células hacia el pulmón infectado por el virus, reduciendo su número a nivel periférico y favoreciendo el daño pulmonar (59).

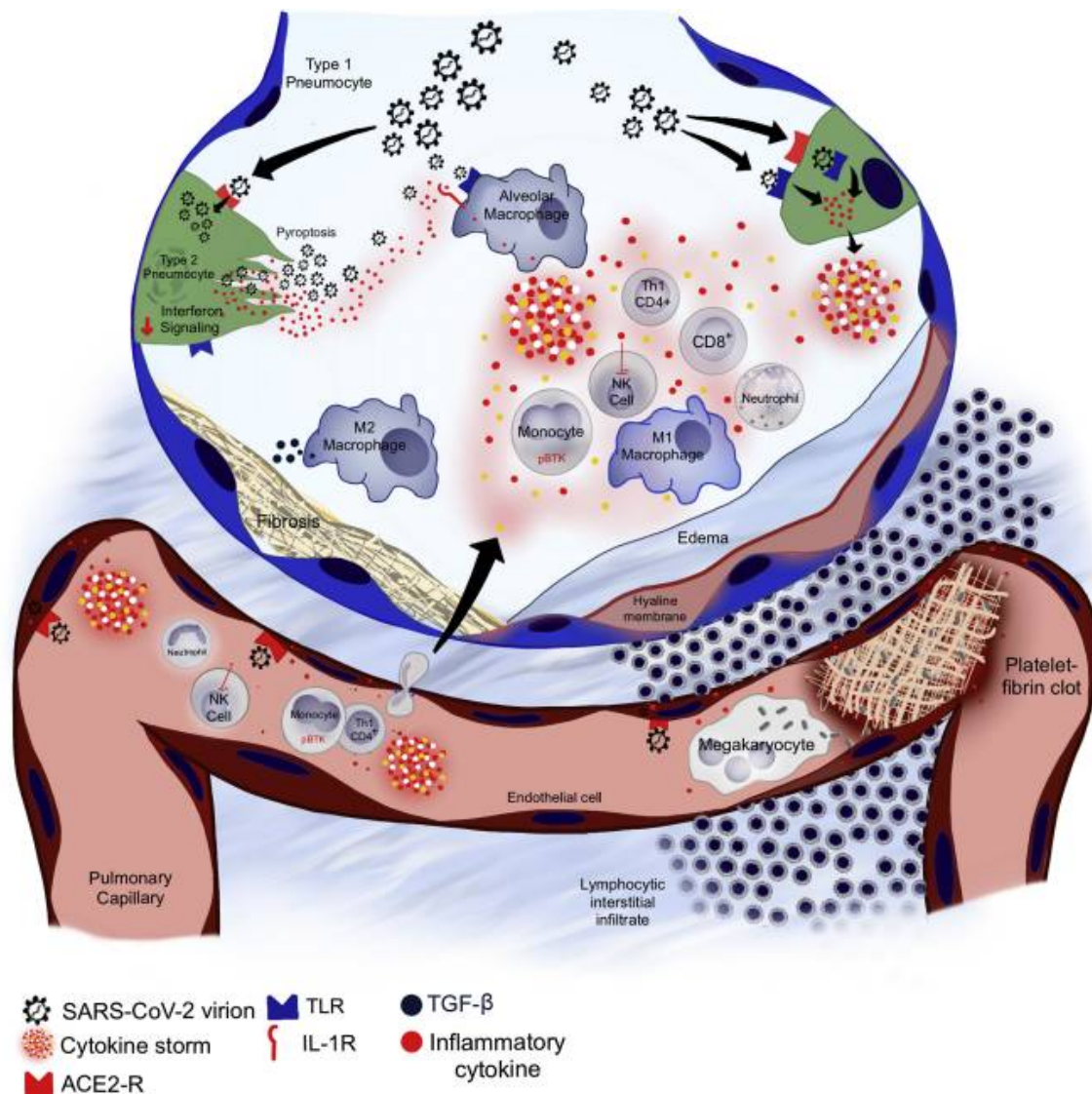


Figura 5. Esquema en el que se detalla el mecanismo de entrada del virus SARS-CoV-2 en los neumocitos tipo II gracias al receptor de la ECA2. Tras la infección de la célula huésped se produce un aumento de citoquinas proinflamatorias (IL-1 β , IL-6, IL-8, factor estimulante de colonias de granulocitos y macrófagos o GM-CSF, TNF- α , IFN- γ). Posteriormente se produce un reclutamiento de monocitos inflamatorios, linfocitos T CD4⁺ y CD8⁺, neutrófilos y células NK en el intersticio pulmonar favoreciendo la respuesta inflamatoria. Este estado hiperinflamatorio consiste en una tormenta de citoquinas que se asocia al desarrollo de SDRA. Imagen adaptada de: Gustine JN *et al.* (2021) (59).

Por otro lado, al producirse la infección por SARS-CoV-2, se produce un bloqueo selectivo de la vía de señalización del interferón tipo 1 (IFN tipo 1) de la célula huésped que ayuda a evadir la respuesta del sistema inmune; las causas de esta inhibición aún no están claramente establecidas pero entre las teorías se apoya la existencia de una supresión del reconocimiento del ARN viral y una interferencia en la señalización del IFN tipo 1 que favorece la replicación del virus y apoya la tormenta de citoquinas (60).

Esta respuesta inmune alterada, como ya se ha expuesto, produce un aumento de citoquinas proinflamatorias, favoreciendo la activación de las vías de coagulación y aumentando la permeabilidad vascular, otra de las causas del daño tisular observado. La trombina favorece la formación de coágulos gracias a la agregación plaquetaria y a la formación de fibrina; durante la inflamación, los mecanismos anticoagulantes fisiológicos (antitrombina III, inhibidor del factor tisular y sistema de proteína C) se ven alterados al presentar un desequilibrio entre la producción y el consumo. Esta alteración en el sistema de coagulación se traduce analíticamente en concentraciones elevadas de Dímero D que resultan como factor pronóstico negativo, ya que se asocia con formación de microtrombos en diversos órganos (pulmón, cerebro, hígado), coagulación intravascular diseminada (CID) y fallo multiorgánico en la enfermedad grave COVID-19 (61,62).

El conocer la alteración del sistema inmunitario innato del huésped durante la infección por SARS-CoV-2 ha impulsado diferentes vías de investigación que estudian tratamientos inmunomoduladores, así como el control del sistema de coagulación para prevenir o modificar el daño pulmonar durante la infección.

2.5. Influencia de la genética en la enfermedad COVID-19

La variabilidad clínica entre individuos con COVID-19 descrita previamente, desde formas asintomáticas hasta casos graves, ha justificado el estudio de la genómica como un factor clave en esta enfermedad. La mayoría de los estudios genéticos han centrado su objetivo en identificar variaciones en el genoma que estén en relación con la susceptibilidad a la infección, la progresión y la gravedad de la enfermedad y el desarrollo

de síntomas (41,42). Diversos estudios de asociación genómica (como el Genome-Wide Association Study o GWAS), meta-análisis y revisiones sistemáticas han puesto en evidencia que ciertos polimorfismos genéticos como los polimorfismos de nucleótido único o SNPs, sobre todo aquellos relacionados con el sistema inmunológico, el SRAA y el sistema de vitamina D, pueden modular significativamente el riesgo de infección y la gravedad de la enfermedad. Estos polimorfismos de nucleótido único o SNPs son las variaciones genéticas más frecuentes que se encuentran en al menos el 1 % de la población general.

En una revisión en la que se estudiaron polimorfismos relacionados con genes inflamatorios (como por ejemplo IL6, IL1B o el TNF- α) encontraron asociaciones relevantes entre esas variaciones y las formas graves de COVID-19 en las que se observaba una respuesta hiperinflamatoria con mayor riesgo de hospitalización, mayor daño pulmonar y sistémico y mayor mortalidad (63-65). El estudio CARGENCORS identificó variantes en genes implicados en la respuesta inmune innata y adaptativa, apoyando la hipótesis de que una de las causas de progresión más agresiva de la enfermedad es una disfunción en la respuesta antiviral de este sistema (40).

Otros estudios han valorado los polimorfismos en el SRAA, ya que la ECA2 presenta un papel muy importante en la fisiopatología de la enfermedad, evaluando estos polimorfismos y su relación con la severidad de la enfermedad. Entre los hallazgos de algunos estudios destaca el polimorfismo rs2285666 en el gen *ECA2* que se relaciona con la susceptibilidad en la enfermedad y las formas más graves de esta debido a la alteración de la entrada del virus en la célula del huésped y el equilibrio entre la angiotensina II y la angiotensina 1-7, resultando en una respuesta proinflamatoria y vasoconstrictora a nivel respiratorio (66). Han sido estudiados otros polimorfismos que no solo influyen en la susceptibilidad, si no que también predicen la evolución clínica, algunos de ellos también relacionados con el SRAA y con el sistema de vitamina D (67).

Otro de los puntos de la investigación se ha centrado en el análisis de la influencia de los polimorfismos en el sistema de la vitamina D y su receptor (VDR), debido a su función inmunomoduladora, que se relacionan con una menor expresión de este receptor y por lo

tanto con una disminución de los efectos antiinflamatorios de la vitamina D aumentando las complicaciones severas (68).

Estos estudios han puesto en evidencia que la variabilidad interindividual y la severidad de la enfermedad se ven influenciadas por diversos factores, entre ellos factores genéticos, y que estos tienen unas implicaciones en la estratificación del riesgo, el pronóstico y el abordaje terapéutico personalizado. La validación de estos biomarcadores genéticos junto con datos clínicos, aunque los resultados aún no han sido validados para el manejo clínico de la enfermedad, representa un paso crucial en el manejo futuro de esta enfermedad.

2.6. La inteligencia artificial en la medicina

El inicio de la inteligencia artificial en la medicina se remonta a 1956, cuando John McCarthy estableció el término refiriéndose a una nueva disciplina que establecía sus bases en la reproducción mediante máquinas de las funciones cognitivas humanas (69).

Durante los años 1970 y 1980, la inteligencia artificial en la medicina se centró en el diseño de programas que simulaban el razonamiento de especialistas a partir de bases de conocimiento y motores de inferencia, es decir, desarrollando sistemas expertos en el ámbito. Algunos de estos estudios son el INTERNIST-1, que es un sistema de diagnóstico en medicina interna que se desarrolló en la Universidad de Pittsburgh, cuyo objetivo es ayudar a expertos en medicina interna en el diagnóstico diferencial, incluyendo cientos de patologías y diversas presentaciones clínicas (70). Se trataba de un sistema en el que se incluía una amplia base integrando enfermedades y síntomas, así como datos probabilísticos cuyo objetivo era estimar la probabilidad relativa de los diagnósticos. A pesar de limitaciones propias de la aplicabilidad en la práctica clínica, su desarrollo marcó un hito sirviendo de referente para estudios posteriores como el estudio CADUCEUS que utilizaba el razonamiento abductivo y ampliaba el alcance a cientos de enfermedades (71). Otro de los ejemplos que tuvieron más importancia fue el estudio MYCIN que se desarrolló en la Universidad de Stanford, con el objetivo de diagnosticar infecciones bacterianas y recomendar tratamientos antibióticos (72). Estos sistemas iniciaron la aplicación práctica de la inteligencia artificial en la medicina y, a su vez, pusieron en evidencia retos como la integración de la opinión y el conocimiento del experto. En este

mismo marco, surgió otro modelo denominado Pathfinder que se centraba en el diagnóstico de enfermedades linfáticas basándose en la histopatología. Este sistema incorporó redes bayesianas, un tipo de técnica probabilística cuya particularidad es que integra la incertidumbre y los datos empíricos, de esta forma, estando más actualizados. Estas gráficas probabilísticas permiten representar de forma compacta las relaciones entre síntomas, signos y diagnósticos (73,74).

En los años posteriores y de forma progresiva, la inteligencia artificial en medicina empezó a centrarse en el aprendizaje automático, ya que el sistema de expertos basado en reglas presentaba limitaciones como la dificultad para obtener el conocimiento de expertos, mantener las bases de datos actualizadas y la falta de flexibilidad de estas reglas en determinados aspectos clínicos. De esta forma, el aprendizaje automático comenzó a evolucionar dada la facilidad de obtención de patrones de los datos disponibles, disminuyendo la codificación manual, y la posibilidad de adaptación a cada contexto clínico. A partir de este momento comenzaron a usarse algoritmos que incluían *redes neuronales* artificiales, máquinas de soporte vectorial y árboles de decisión que demostraron mayor capacidad de manejo de datos heterogéneos como son variables clínicas, datos de imagen y datos de laboratorio (75 - 77). Este cambio en la metodología coincidió con la digitalización del sector sanitario y la incorporación de registros médicos electrónicos, factores que facilitaron el desarrollo de modelos más avanzados y representativos. Así, se consolidó una distinción entre la inteligencia artificial simbólica, basada en reglas, y la inteligencia artificial estadística, basada en datos, siendo esta última la que comenzó a progresar en su desarrollo.

Estos algoritmos de aprendizaje automático de forma progresiva comenzaron a demostrar mejores resultados y más precisos comparados con los algoritmos de especialistas humanos, sobre todo en áreas de predicción de mortalidad, estratificación de riesgo o valoración de imágenes radiológicas. Otro de los usos de estos métodos es la investigación biomédica, en la cual sería de ayuda para identificar variables genéticas o biomarcadores que tienen mayor influencia en la evolución, gravedad y pronóstico de la enfermedad y así identificar de forma temprana a los pacientes de mayor riesgo. Dentro de sus beneficios se encuentran la accesibilidad a profesionales sin gran experiencia en manejo de datos y la eficacia o el ahorro de tiempo en el análisis de grandes volúmenes de datos. Otro cambio relevante en esta etapa es el avance en las técnicas de

almacenamiento de datos, procesamiento de estos y manejo de valores perdidos, dado que son fundamentales para garantizar una calidad en el entrenamiento de modelos. Por otro lado, el estándar usado para valorar de forma fiable el rendimiento de estos algoritmos se estableció con la validación cruzada; a su vez, se comenzaron a integrar en estos estudios los valores de sensibilidad, especificidad y área bajo la curva ROC.

De esta forma, el aprendizaje automático comenzó a formar parte de la medicina, entendiendo la importancia en este campo de la necesidad de integrar algoritmos detallados y la generalización de modelos. Aunque también presentaba ciertas limitaciones como la simplicidad de estos modelos y su escasa capacidad para considerar datos no estructurados (datos no numéricos o subjetivos).

A partir de 2010, el desarrollo del aprendizaje profundo se vio impulsado debido a la mejor capacidad de procesamiento gráfico, la mejora de las *redes neuronales* y el mejor manejo de grandes volúmenes de datos. Gracias a estas mejoras, sobre todo a las *redes neuronales*, el rendimiento en la valoración y estudio de datos clínicos en ocasiones supera el criterio de los expertos. Por otro lado, se ha implementado el manejo y procesamiento de datos secuenciales y series temporales obtenidos durante la monitorización de pacientes en unidades de cuidados intensivos; de esta forma, las redes neuronales son capaces de identificar de forma temprana eventos críticos, sobre todo hemodinámicos. Otro progreso durante esta etapa ha sido el desarrollo de modelos que han sido capaces de integrar datos de diferentes características como información clínica, datos de laboratorio, datos genéticos y resultados de imagen; con el objetivo de mejorar la precisión diagnóstica, así como predecir la respuesta terapéutica y la evolución de la patología (78).

Esta evolución de la inteligencia artificial hacia el aprendizaje profundo ha demostrado grandes avances en la medicina, mejorando sus capacidades, diversificando sus aplicaciones y teniendo como objetivo una asistencia médica más precisa y personalizada. Bien es cierto que presenta ciertas limitaciones, como, por ejemplo, la dificultad para explicar el funcionamiento de determinadas herramientas a profesionales, el manejo de gran cantidad de datos que puede poner en riesgo ciertos aspectos éticos o legales, las diversas formas de adquirir estos datos según el centro, entre otros. Pese a estas limitaciones, debemos ser conscientes de que estos avances en el aprendizaje profundo

progresarán y cada vez se integrarán de forma más activa en la práctica clínica, sirviendo como una herramienta avanzada para el médico, ayudándole en su proceso diagnóstico y terapéutico.

Actualmente, la inteligencia artificial ha pasado de ser una idea utópica a formar parte de las herramientas reales en el sistema sanitario actual. Es por eso que agencias, como por ejemplo la Agencia Europea del Medicamento (EMA), han comenzado a apoyar el desarrollo de estos sistemas y se están estableciendo marcos legales para garantizar la seguridad de estas herramientas. El desarrollo de modelos multimodales que integren datos de diferentes características es uno de los avances más prometedores, permitiendo el desarrollo de la medicina personalizada, optimizando aspectos de prevención, diagnóstico y tratamiento. Los retos futuros en estos sistemas son la validación externa en diferentes poblaciones para asegurar una igualdad en su aplicación y ser capaces de ofrecer una transparencia adecuada y facilidad de uso a los profesionales sanitarios para que puedan integrar estos métodos en su actividad diaria.

HIPÓTESIS Y OBJETIVOS

La COVID-19 es una enfermedad infecciosa causada por el virus SARS-CoV-2 que presenta una gran variabilidad clínica, desde casos asintomáticos hasta casos que presentan afectación multisistémica, siendo la presencia de SDRA la más condicionante. Esta enfermedad fue la causante de una pandemia en 2020, suponiendo en ese momento un reto a nivel sanitario debido a la complejidad del manejo de estos pacientes que a día de hoy sigue estando presente. La variabilidad clínica de la enfermedad se ve influenciada por factores genéticos que influyen en el sistema renina-angiotensina, en el metabolismo de la vitamina D y en la respuesta inflamatoria. En este contexto, la integración de técnicas de aprendizaje automático mediante inteligencia artificial podría ayudarnos a identificar qué pacientes tendrán un mayor riesgo de desarrollo de complicaciones durante el transcurso de la enfermedad y así optimizar el diagnóstico y el manejo de estos.

Por ello, el objetivo principal de nuestro trabajo fue el desarrollo de un algoritmo mediante técnicas de inteligencia artificial que nos permitiera identificar las variables genéticas que más influyen en la evaluación de los pacientes con COVID-19. Los objetivos específicos fueron:

1. Identificar polimorfismos genéticos que influyen en el riesgo de neumonía por COVID-19.
2. Identificar polimorfismos genéticos que influyen en el riesgo de mortalidad.
3. Identificar polimorfismos genéticos que influyen en el reingreso en el primer año de seguimiento tras la infección por SARS-CoV-2.
4. Identificar polimorfismos genéticos que influyen en la mortalidad durante ese reingreso.

MATERIAL Y MÉTODO

- 4.1. Pacientes.
- 4.2. Aislamiento de ADN de alto peso molecular.
- 4.3. Discriminación alélica mediante PCR con sondas TaqMan.
- 4.4. Análisis mediante aprendizaje automático.
- 4.5. Aspectos éticos.

4.1. Pacientes

Se trata de un estudio observacional con una cohorte prospectiva en el que se incluyeron pacientes con COVID-19 que precisaron ingreso y desarrollaron SDRA. Los pacientes fueron diagnosticados y hospitalizados en el Hospital Clínico Universitario Río Hortega (Valladolid, España) entre marzo y noviembre de 2020.

Los criterios de inclusión fueron pacientes mayores de 18 años con diagnóstico microbiológico de COVID-19 confirmado mediante test PCR de SARS-CoV-2 con resultado positivo, síntomas de afectación pulmonar, imagen radiológica al ingreso compatible con el diagnóstico de neumonía por COVID-19, y variables de laboratorio y clínicas al ingreso compatibles con la enfermedad. En el estudio, la mayoría de los pacientes tenían una presentación clínica que correspondía a la descripción realizada en el apartado de manifestaciones clínicas compatibles con casos moderados y casos graves. Otro de los criterios de inclusión fue la firma del consentimiento informado.

Las imágenes radiológicas de afectación pulmonar se clasificaron en tres grupos según Litmanovich *et al.* (79). Presentación típica de neumonía por COVID-19; hallazgos indeterminados o menos típicos de neumonía por COVID-19, que pueden presentarse en diversos procesos infecciosos o no infecciosos; y hallazgos atípicos o poco comunes de neumonía por COVID-19, que hacen necesario considerar diagnósticos alternativos.

En cuanto a los criterios de exclusión, estos fueron edad inferior a los 18 años, enfermedad tumoral activa u otras enfermedades con una expectativa de vida inferior a 6 meses, hallazgos radiológicos atípicos o poco comunes en el COVID-19 y el hecho de no haber firmado el consentimiento informado.

Se recopilaron variables demográficas como sexo, edad, nivel de dependencia; antecedentes médicos como enfermedades cardiovasculares, antecedentes de tabaquismo, enfermedades hepáticas, enfermedades respiratorias y autoinmunes, así como antecedentes neoplásicos o inmunodeficiencias adquiridas; se incluyeron tratamientos previos farmacológicos y no farmacológicos. En cuanto a las variables recopiladas durante el ingreso, se incluyeron características clínicas al inicio de la infección, complicaciones y evolución durante la hospitalización, así como parámetros de

laboratorio, incluyendo hemograma completo, bioquímica, perfil de coagulación, situación gasométrica y marcadores inflamatorios y afectación radiológica; también se incluyeron tratamientos farmacológicos y no farmacológicos recibidos durante el ingreso. Se realizó seguimiento clínico de los participantes del estudio durante un año, valorando el reingreso hospitalario por el mismo motivo. Además, se recogieron muestras de sangre venosa en tubos con EDTA durante la hospitalización. El estudio empleó un diseño doble ciego para minimizar el sesgo.

4.2. Aislamiento de ADN de alto peso molecular

A partir de las muestras de sangre periférica extraídas, se extrajo el ADN genómico de alto peso molecular. Posteriormente, las células nucleadas se aislaron mediante centrifugación repetida y lisis eritrocitaria con solución hipotónica (centrifugación de la sangre total en 50 mL de H₂O durante 30 minutos a 1500 rpm a 4 °C). Tras ese procedimiento se procedió a la recuperación de la interfase creada y a la lisis de glóbulos rojos con agua destilada; se realizó el lavado de células mononucleadas en tampón Fornace (0.25 M sacarosa, 50 mM Tris-HCl pH=7.5, 25 mM KCl y 5 mM MgCl₂) para proporcionar la osmolaridad necesaria y no dañar los leucocitos. Posteriormente, las células fueron centrifugadas a 1500 rpm durante 10 minutos para su precipitación.

Las células nucleadas fueron suspendidas de nuevo en tampón Fornace a una concentración estimada de 5×10^6 células/mL. A continuación, se añadió EDTA 0.5 M, pH 8.0, hasta alcanzar una concentración final de 10 mM, con un fin de quelante de iones divalentes para inactivar las nucleases; también se incorporó SDS hasta una concentración final del 1 % con el objetivo de destruir las membranas celulares; y proteinasa K para degradar las proteínas. La mezcla fue incubada a 55 °C durante un periodo de 8 a 16 horas.

Tras este periodo de incubación de la mezcla se llevaron a cabo la extracción y purificación del ADN. Fue tratado como una mezcla de fenol taponado con un pH de 8, para evitar que el ADN se mantuviese en una fase intermedia, entre la fase acuosa y la orgánica, y con CIIAA (una composición de cloroformo y alcohol isoamílico con una proporción de 24:1). De esta forma el ADN permaneció en la fase acuosa, pudiéndose

eliminar los residuos, y posteriormente centrifugándose de nuevo para separar las dos fases: una en la que se encuentra el ADN y otra en la que se encuentran los detritus celulares. Una vez finalizado este paso, se añadió etanol frío a la fase acuosa, consiguiendo la precipitación del ADN, que posteriormente fue precipitado y lavado con etanol al 70 % y suspendido en agua estéril.

A continuación, para calcular la concentración de ADN y el grado de contenido proteico se debe medir en un espectrofotómetro automático de ADN/ARN (GeneQuant, Pharmacia) la absorbancia a 260 y 280 nm, respectivamente, con la siguiente fórmula:

$$\mu\text{g de DNA /ml} = (\text{DO } 260) \times (\text{factor de dilución}) \times 50$$

Donde 50 es un factor de corrección, ya que la unidad de densidad óptica con una luz incidente de 260 nm es un valor de absorbancia que tiene 50 μg de DNA/mL.

Para determinar el grado de contaminación proteica se utiliza el cociente DO260/DO280, considerando valores adecuados un resultado entre 1.65 y 2. Los valores inferiores a estos indican contaminación por proteínas o solventes orgánicos, por lo que en estos casos se volvió a realizar una nueva purificación del ADN; en los casos con valores superiores indican exceso de ARN, en esta situación se trató la solución de ADN con ARNasa y se realizó un nuevo proceso de purificación.

4.3. Discriminación alélica mediante PCR con sondas TaqMan

En la discriminación alélica mediante reacción en cadena de la polimerasa (PCR) con sondas TaqMan, la amplificación y la detección del genotipo se realizaron simultáneamente. Las sondas TaqMan son oligonucleótidos que hibridan específicamente con cada alelo en la región donde se localiza el polimorfismo a estudiar (80). Estas sondas estaban marcadas con un fluorocromo donador en el extremo 5' que emitía fluorescencia al ser estimulado; y un aceptor ("quencher") en el extremo 3', que absorbía la fluorescencia liberada por el donador cuando la sonda estaba intacta y ambos estaban próximos. Las sondas utilizadas en nuestro estudio estaban marcadas con fluorocromos VIC y FAM para cada alelo. Para llevar a cabo la técnica PCR se requirieron otros componentes como Taq polimerasa, dNTPs, cloruro de magnesio y cebadores sentido y

antisentido. Esta reacción se llevó a cabo en termocicladores que integran un lector de fluorescencia; este modelo es el termociclador *Step One Plus Real Time PCR* (Applied Biosystems). El proceso se llevó a cabo durante la amplificación del ADN diana, la sonda híbrida con su cadena complementaria y la ADN polimerasa, que tiene actividad 5'-3' exonucleasa, al desplazarse por la cadena de ADN, hidrolizó el extremo libre 5' de la sonda, produciéndose la liberación del fluorocromo donador. Al permanecer separados, la fluorescencia emitida por el fluorocromo fue captada por el lector (Figura 6). Además, la emisión de fluorescencia producida durante la reacción es proporcional a la cantidad de ADN formado, permitiendo en todo momento conocer y registrar la cinética de la amplificación.

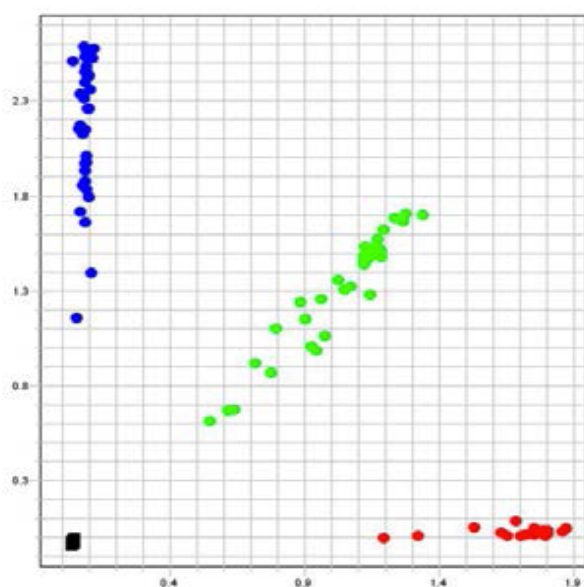


Figura 6. Ejemplo de representación de los resultados de la discriminación alélica con sondas TaqMan. Cada nube de puntos representa uno de los genotipos posibles: los azules son homocigotos, los verdes son heterocigotos y los rojos son homocigotos mutados.

De esta forma, se analizaron diversos polimorfismos en genes del SRAA, del sistema inflamatorio y del metabolismo de la vitamina D que se encuentran implicados en el riesgo de neumonía por COVID-19, mortalidad y reingreso por esta causa y mortalidad durante el reingreso. Los polimorfismos genéticos fueron seleccionados según los siguientes criterios: (1) funcionalidad (efecto descrito previamente o potencial) y (2) distribución a lo largo del gen, con preferencia por aquellos ubicados en exones o regiones contiguas. Se seleccionaron SNPs con frecuencia del alelo menor (MAF) > 5 %. Se utilizaron las bases de datos Genecard (www.genecards.org) y NCBI (www.ncbi.nlm.nih.gov/snp) para identificar los genes y las vías asociadas a cada polimorfismo incluido.

4.4. Análisis mediante aprendizaje automático

La inteligencia artificial (IA) es la disciplina que integra técnicas que permiten a máquinas analizar grandes volúmenes de datos, reconocer patrones y realizar predicciones. Dentro de esta disciplina, el aprendizaje automático es la forma actualmente más desarrollada en el ámbito de la medicina. Este aprendizaje integra modelos supervisados, aquellos con datos etiquetados, y modelos no supervisados que son prácticos para descubrir agrupaciones de pacientes. Estos sistemas requieren de procesos de validación para garantizar la exactitud en los diferentes ámbitos de aplicación, pero son una herramienta útil para mejorar la calidad de la atención sanitaria, ya que permiten tomar decisiones personalizadas.

En este estudio, la aplicación del aprendizaje automático es uno de los puntos clave, ya que nos permitió analizar las asociaciones entre la distribución genética de los polimorfismos, la neumonía por COVID-19 (diagnosticada mediante patrones radiográficos compatibles) y el riesgo de muerte durante la hospitalización inicial. Además, los pacientes fueron seguidos clínicamente durante un año y se realizó seguimiento en cuanto a reingresos hospitalarios y riesgo de muerte relacionada con COVID-19.

Para elaborar un sistema de aprendizaje automático se siguieron las siguientes etapas:

- Identificar las variables clave que van a influir en el objetivo del estudio.
- Diseño e implementación de modelos: se estudian diferentes métodos de aprendizaje.
- Comparativa y selección de algoritmos: se estudian los diferentes métodos para determinar aquellos que tengan una adecuada precisión y velocidad de entrenamiento.

4.4.1. Selección de modelo de aprendizaje automático

Se realizó una evaluación de los diferentes algoritmos para elegir el modelo de aprendizaje más adecuado para el estudio. Existen dos tipos de técnicas de aprendizaje:

- Algoritmos supervisados: conjuntos de datos etiquetados, donde cada ejemplo incluye características de entrada y el resultado deseado. El modelo aprende a

relacionar los patrones de entrada con el resultado esperado y, una vez entrenado, generaliza para predecir etiquetas de nuevos casos. Algunos ejemplos son la regresión logística, las máquinas de soporte vectorial (SVM) y redes neuronales supervisadas.

- Algoritmos no supervisados: trabajan con datos sin etiquetas y su objetivo es descubrir el patrón. Entre sus aplicaciones destaca el clustering o agrupamiento de pacientes con características similares mediante K-means.

De esta forma se escogió el más adecuado en cuanto a precisión y capacidad de generalización. Para ello se realizó un proceso de validación cruzada mediante el rendimiento y se ajustaron los parámetros mediante estrategias de búsqueda sistémica o probabilística de forma que se evitó tanto el subajuste como el sobreajuste. Una vez se seleccionó la configuración adecuada, el modelo se reentrenó y se evaluó sobre un bloque de datos reservados, obteniéndose métricas finales que confirmaron su robustez. De esta forma, la decisión tiene en cuenta la métrica principal, pero también otros criterios como la aplicación clínica o el tiempo y la facilidad de integración.

Los algoritmos evaluados durante este estudio fueron los siguientes:

a) Support Vector Machine (SVM):

Se trata de un método supervisado utilizado para la clasificación; su objetivo es separar de forma adecuada diferentes datos etiquetados. Este método busca un hiperplano que mejor separe dos clases de puntos en un espacio de características.

Este modelo tiene 3 fases:

- Preprocesado: normalización de características e imputación de valores perdidos.
- Optimización: se resuelve el problema para encontrar los multiplicadores asociados a las muestras de entrenamiento.
- Validación de hiperparámetros: se utiliza validación cruzada para ajustar el parámetro de penalización evaluando métricas como la sensibilidad o el área bajo la curva (AUC).

Una vez entrenado, el modelo clasifica nuevas observaciones aplicando la combinación de vectores de soporte y el umbral obtenido (81).

b) Decision Tree (DT):

Es un método de aprendizaje supervisado que toma decisiones mediante una estructura de un árbol en la que cada nodo corresponde una característica y cada rama representa una regla de decisión (82).

El método tiene 3 fases:

- Preprocesado: conversión de variables categóricas y manejo de valores faltantes.
- Construcción del árbol: selección recursiva de la mejor estructura. Para cada nodo se evalúa cada característica, calculando el criterio de pureza resultante y eligiendo el que mejor separa las clases. Se realiza una división continua hasta alcanzar el tamaño mínimo de muestra.
- Poda y validación de hiperparámetros: se realiza una poda, eliminando ramas que no aportan mejor en la validación, para evitar el sobreajuste. Ajuste de parámetros por validación cruzada.

c) Gaussian Naïve Bayes (GNB):

Se trata de un método supervisado que asume independencia condicional entre las características y modela cada una con una distribución normal. El modelo calcula la probabilidad de pertenencia a cada clase aplicando el teorema de Bayes, multiplicando las densidades gaussianas de cada variable condicionadas a la clase por su probabilidad; posteriormente, se asigna la de mayor probabilidad (83).

Para su desarrollo tiene 3 fases:

- Preprocesado: verificar que las variables continuas cumplan las características de normalidad y tratar valores faltantes.
- Estimación de parámetros: cálculo de parámetros de cada característica sobre el conjunto de entrenamiento; cómputo de las probabilidades de cada clase y almacenaje de estos parámetros.

- Validación de hiperparámetros: ajuste de parámetros y uso de validación cruzada, eligiendo la configuración que mejor se ajuste a las probabilidades predichas.

Una vez entrenado, el método clasifica nuevas observaciones según los parámetros aprendidos y los combina con probabilidades previas para obtener probabilidades posteriores de cada clase.

d) K-Nearest Neighbors (KNN):

Es un método que asigna a cada nueva observación la clase más frecuente entre sus k o número de vecinos medidos a través de una métrica de distancia (84).

El modelo tiene 3 fases:

- Preprocesado: normalización de las características para que la escala de cada variable no distorsione las distancias, codificación de variables categóricas y eliminación de variables incompletas.
- Almacenamiento de instancias: se produce un almacenaje de todas las observaciones de entrenamiento, sin ajustar parámetros.
- Validación de hiperparámetros: selección de k y elección de la métrica de distancia, usando la validación cruzada que maximice la métrica clínica de interés.

e) eXtreme Gradient Boosting (XGB):

Este método se basa en árboles de decisión y la técnica de “boosting”, que es aquella en la que se combinan diferentes modelos más débiles para formar un modelo predictivo más preciso. De forma progresiva se va ajustando un nuevo árbol de decisión para corregir el modelo anterior y mejorar los términos de regularización (85).

Este método comprende 3 fases:

- Preprocesado: codificación de variables categóricas y corrección de valores faltantes.

- Construcción y optimización de árboles: se construye un nuevo árbol y se aplican técnicas para reducir la varianza y acelerar el entrenamiento.

Validación de hiperparámetros: equilibrar la velocidad de aprendizaje y precisión y monitorización de la métrica de evaluación.

4.4.2. Fase de entrenamiento

Para una correcta implementación del método en este estudio se empleó la validación cruzada anidada junto con técnicas de optimización bayesiana para ajustar de manera efectiva y fiable los parámetros. La validación cruzada tenía como objetivo evaluar los métodos de aprendizaje automático, realizando una división adecuada de datos para evitar la pérdida de información y garantizar la validez estadística de los resultados. En este caso se utilizó el proceso de validación cruzada anidada, en el que el ciclo externo evaluaba el rendimiento general del modelo, mientras que el ciclo interno se centraba en optimizar los parámetros. Las técnicas de optimización bayesiana se utilizaron en el ciclo interno para explorar de manera eficiente el espacio de parámetros críticos, como la profundidad máxima del árbol (`max_depth`), el número de estimadores (`n_estimators`), la tasa de aprendizaje (`learning_rate`), y los términos de regularización (`lambda` y `alpha`). La optimización bayesiana empleó un modelo probabilístico basado en un proceso gaussiano para identificar combinaciones óptimas de parámetros. Este enfoque aprovechó información de repeticiones anteriores, reduciendo la necesidad de evaluaciones exhaustivas y concentrándose en las áreas más prometedoras (86,87). Este método mitigó el riesgo de sobreajuste al mantener los datos de prueba del ciclo externo separados del proceso de optimización. Mejoró la estabilidad del modelo mediante evaluaciones uniformes en diversas particiones de los datos. La sinergia de estas estrategias produjo modelos con mejor rendimiento y generalización robusta. A partir de un análisis sistemático de la relevancia de las características, se implementó un enfoque híbrido para identificar las variables más impactantes. Primero se evaluó la importancia de las características mediante un modelo preliminar de XGB, que generó puntuaciones basadas en ganancia, cobertura o peso. Esto fue seguido por métodos interactivos de selección de características, incluyendo la Eliminación Recursiva de Características (RFE) utilizando XGBoost como estimador base, con el objetivo de reducir el conjunto de variables manteniendo un alto rendimiento predictivo. Finalmente, se evaluó el efecto de eliminar

características específicas mediante validación cruzada, asegurando que solo se conservaran aquellas que mejoraran significativamente el rendimiento del modelo.

La Figura 7 resume los pasos seguidos para implementar los algoritmos de aprendizaje automático.

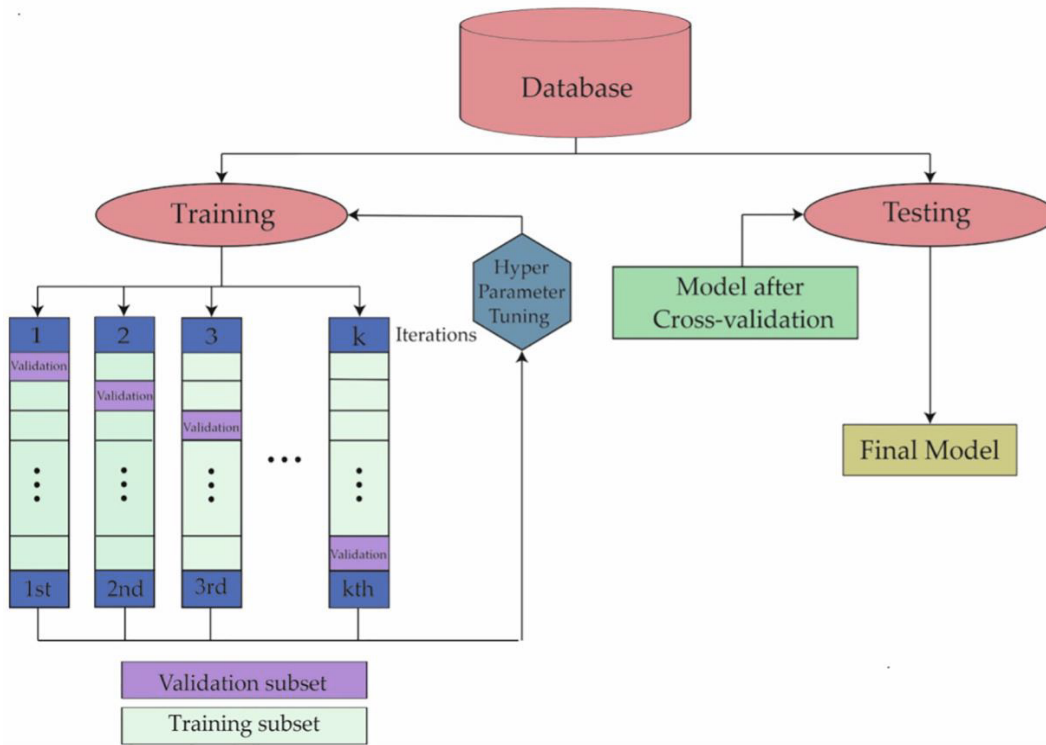


Figura 7. Pasos del algoritmo de aprendizaje automático

Para minimizar el sobreajuste en XGBoost, se implementaron varias técnicas. Estas incluyeron el uso de regularización explícita mediante los parámetros λ y α , el control del tamaño máximo del árbol (max_depth), la reducción de la tasa de aprendizaje (learning_rate) y la aplicación de parada temprana (early stopping) para finalizar el entrenamiento si las métricas de validación no mejoraban tras varias repeticiones consecutivas. Se realizó una validación creando subconjuntos de datos mediante muestreo para evaluar la incertidumbre en las métricas de rendimiento y garantizar la consistencia. Además, el modelo fue validado en una cohorte externa independiente para evaluar su capacidad de generalización en entornos clínicos diversos. Los datos se dividieron aleatoriamente en conjuntos de entrenamiento (70 %) y prueba (30 %) para garantizar una representación equilibrada de clases y características clave. Las simulaciones se ejecutaron rigurosamente durante 100 repeticiones, considerando cuidadosamente los

valores medios y de desviación estándar, reduciendo así el impacto potencial del ruido y asegurando la obtención de conclusiones estadísticamente válidas (88).

4.4.3. Métricas para analizar y comparar los algoritmos

Para valorar y cuantificar el comportamiento que los algoritmos del aprendizaje automático tienen con el conjunto de datos utilizamos las siguientes métricas:

- Balance Accuracy (BA) (%): mide el rendimiento de modelos de clasificación; es el promedio de la sensibilidad para cada clase.
- Sensibilidad (Recall o TPR): Proporción de verdaderos positivos correctamente identificados sobre el total de casos positivos reales. Indica la capacidad del modelo para detectar los casos de interés (p. ej., pacientes enfermos).
- Precisión o Valor Predictivo Positivo (VPP): Fracción de predicciones positivas que son verdaderamente positivas. Evalúa la fiabilidad de las predicciones de enfermedad:
- Área bajo la curva ROC (AUC): Integral de la curva ROC (Recall vs. FPR) desde FPR=0 hasta FPR=1. Un valor de 0,5 equivale a azar; cuanto más cercano a 1, mejor capacidad discriminativa global del modelo.
- F1 Score: Media armónica entre precisión y sensibilidad. Equilibra ambos aspectos en un solo valor, muy útil si las clases están desbalanceadas:
- Coeficiente de correlación de Matthews (MCC): Correlación entre las clases predichas y las reales, que incorpora TP, TN, FP y FN en una sola fórmula. Varía de -1 (predicción inversa) a +1 (perfecto), y 0 indica azar:

- Índice de Youden (DYI): resume la habilidad diagnóstica combinando sensibilidad y especificidad:
- Kappa de Cohen (k): Medida de concordancia entre las predicciones y las etiquetas reales, corregida por el acuerdo esperado por azar. Va de -1 (peor que azar) a 1 (concordancia perfecta), siendo 0 el nivel de azar.

4.4.4. Selección del modelo y validación

Tras realizar el entrenamiento del modelo, se comprueba su capacidad de generalización, es decir, evaluando nuevos datos. Esta fase es la principal en todo método que busque resultados reales y aplicables en la actividad habitual. Para ello han de seguirse unas etapas:

- Validación externa: una vez ajustados los parámetros del modelo, se procede a evaluar su rendimiento con un conjunto de datos nuevo, es decir, una cohorte distinta a la del entrenamiento, lo que nos va a permitir evaluar la capacidad de generalización en escenarios realistas.
- Generación de predicciones y cálculo de métricas: tras realizar la validación externa, se generan predicciones que son comparadas con los verdaderos valores para determinar un conjunto de métricas. Es fundamental en esta fase comparar con los resultados obtenidos durante la validación cruzada, ya que podemos descubrir alteraciones en el rendimiento que indiquen que el modelo ha memorizado los patrones de entrenamiento.
- Ajuste de umbrales de decisión: tras realizar un análisis sistemático de las métricas, se construyen unas curvas de decisión que permiten seleccionar los puntos de corte que más utilidad expresan en el estudio.
- Análisis de errores: de esta forma se evalúa el alcance del modelo y sus posibles sesgos; de esta forma se entienden los patrones en los que el modelo falla.
- Informe de robustez: refleja el rendimiento del modelo en todos los conjuntos de datos, la variabilidad de las métricas en los diferentes subgrupos, las

limitaciones detectadas y las condiciones más y menos óptimas en las que el modelo puede trabajar.

Esta fase de validación del modelo nos permite asegurar que este no ha sido sobreentrenado y que puede aportar valores clínicos en los nuevos escenarios, tras haber superado una evaluación externa adecuada, así como un análisis de métricas y un ajuste de errores. De esta forma podemos obtener una herramienta efectiva y adecuada.

En este estudio se compararon los diferentes modelos de aprendizaje automático ya expuestos previamente. La evaluación de estos modelos se realizó utilizando la Statistics and Machine Learning Toolbox de MATLAB (89).

4.5. Aspectos Éticos

Este estudio que involucró pacientes fue realizado de acuerdo con los principios de la Declaración de Helsinki (2008) y recibió la aprobación del Comité de Ética del Hospital Universitario Río Hortega de Valladolid (PI216-20). El estudio cumplió completamente con las normas éticas de la Asociación Médica Mundial, así como con las leyes españolas de protección de datos (LO 15/1999) y regulaciones relacionadas (RD 1720/2007). Todos los pacientes que aceptaron participar proporcionaron su consentimiento informado por escrito y firmado.

RESULTADOS

- 5.1. Riesgo de neumonía por COVID-19.
- 5.2. Riesgo de mortalidad.
- 5.3. Riesgo de reingreso durante el primer año tras COVID-19.
- 5.4. Riesgo de mortalidad durante el reingreso en el primer año tras COVID-19.

Los resultados obtenidos en este estudio incluyen una cohorte de 338 pacientes con diagnóstico de COVID-19 que precisaron ingreso en el HURH entre los meses de marzo y noviembre de 2020. El conjunto de características de la muestra en cuanto a datos demográficos, antecedentes y tratamiento farmacológico previo se recoge en la Tabla 2.

La edad media de los pacientes era de 73,26 años, siendo predominante el sexo masculino con 182 pacientes (53,84 %). Del total de pacientes, 248 (73,3 %) presentaron una neumonía asociada a COVID-19 confirmada por patrones radiográficos compatibles. La media de días desde el inicio de los síntomas hasta su ingreso hospitalario fue de 5,77 (DS 5,45), permaneciendo ingresados una media de 18,21 días (DS 22,75). En el seguimiento realizado durante el siguiente año, 77 pacientes (29,3 %) requirieron hospitalización de nuevo por la misma causa y de estos se produjo fallecimiento en 25 pacientes (9,5 %). En cuanto a las comorbilidades, 196 pacientes (58,16 %) presentaban hipertensión arterial, 133 (39,58 %) dislipemia y 68 pacientes (20,24 %) diabetes mellitus. En cuanto a los hábitos de tabaquismo, 57 pacientes (16,91 %) eran exfumadores y 12 (3,56 %) fumadores activos.

Algunos de los tratamientos domiciliarios, entre otros, incluían inhibidores de la enzima convertidora de la angiotensina (IECAs) en 82 pacientes (24,40 %); bloqueadores de los receptores de la angiotensina II (ARA II) en 77 pacientes (22,91 %); estatinas en 81 pacientes (24,11 %); y tratamientos inmunosupresores en 16 pacientes (4,76 %) y corticoides sistémicos en 10 pacientes (2,97 %).

Tabla 2. Características generales de los pacientes del estudio.

Características generales	Pacientes
Edad, media (DS)(años)	73,26 (13,09)
Sexo (hombre; mujer), n(%)	182 (53,84);156 (46,15)
Días desde inicio de síntomas hasta ingreso, media (DS)	5,77 (5,45)
Neumonía por COVID-19, n(%)	248 (73,3)
Días de ingreso, media (DS)	18,21(22,75)
Fallecimiento debido a COVID-19, n(%)	76 (22,48)
Dependencia (sí, moderada, leve, independiente), n (%)	42 (12,42); 47 (13,90); 169 (50); 80 (83,67)
Rehospitalización en el primer año, n(%)	77 (29,3)
Tabaquismo (activo; exfumador; nunca fumador), n (%)	12 (3,56); 57 (16,91); 268 (79,52)
Demencia, n(%)	41 (12,20)
Hipertensión, n(%)	196 (58,16)
Dislipemia, n(%)	133 (39,58)
Infarto de miocardio, n(%)	15 (4,45)
Fallo cardíaco, n(%)	20 (5,97)
Ictus cerebral, n(%)	18 (5,36)
Diabetes mellitus, n(%)	68 (20,24)
EPOC, n(%)	8 (2,43)
Asma, n(%)	32 (9,46)
Apnea obstructiva del sueño, n(%)	20 (5,93)
Enf. renal crónica (4-5), n(%)	26 (7,15)
Neoplasia sin metástasis, n(%)	39 (11,57)
Neoplasia con metástasis, n(%)	5 (1,48)
IECAs, n (%)	82 (24,40)
ARA II, n (%)	77 (22,91)
Estatinas, n (%)	81 (24,11)
Metformina, n (%)	36 (10,71)
Inhibidores DPP-4, n (%)	30 (8,92)
Insulina, n (%)	20 (5,95)
Corticoides inhalados, n (%)	33 (9,82)
Corticoides, n (%)	10 (2,97)
Inmunosupresores/inmunomoduladores, n(%)	16 (4,76)
Confirmación test COVID-19, n (%)	338 (100)

DS: desviación estándar. EPOC: Enfermedad Pulmonar obstructiva Crónica. IECAs: inhibidores de la enzima convertidora de la angiotensina. ARA II: bloqueadores de los receptores de la angiotensina II. DPP-4: dipeptidil peptidasa-4.

Las variables clínicas recogidas en nuestra muestra se presentan en la Tabla 3. Destaca la presencia de disnea en 180 pacientes (53,57 %), febrícula y fiebre en 262 (77,97 %) y 209 (62,20 %) pacientes, respectivamente, tos en 175 pacientes (51,93 %) y astenia en 111 pacientes (33,04 %). Durante la evolución clínica en el ingreso, 61 pacientes (18,94 %) desarrollaron síndrome de distrés respiratorio agudo y 15 pacientes (4,66 %) shock séptico. Del total de la muestra, mostraron datos analíticos de hiperinflamación 126 pacientes (39,13 %).

Tabla 3. Características clínicas de los pacientes del estudio.

Variables clínicas	Pacientes
Disnea, n(%)	180 (53,57)
Fiebre, n(%)	209 (62,20)
Febrícula, n(%)	262 (77,97)
Tos, n(%)	175 (51,93)
Diarrea, n(%)	74 (21,96)
Náuseas, n(%)	45 (13,35)
Anorexia, n(%)	45 (13,35)
Dolor abdominal, n(%)	23 (6,85)
Astenia, n(%)	111 (33,04)
Artralgias, n(%)	33 (9,82)
Anosmia, n(%)	20 (5,93)
Ageusia, n(%)	22 (6,55)
Faringitis, n(%)	14 (4,17)
Cefalea, n(%)	28 (8,31)
SDRA, n(%)	61 (18,94)
Datos analíticos de hiperinflamación, n(%)	126 (39,13)
TEP/TVP, n(%)	5 (1,55)
Shock séptico, n(%)	15 (4,66)

TEP: tromboembolismo pulmonar. TVP: trombosis venosa profunda.

Durante el seguimiento de los pacientes en el ingreso se registraron diferentes variables analíticas que nos servían de apoyo para valorar la evolución de estos (Tabla 4). Debido a que la COVID-19 en sus formas más graves implicaba un deterioro respiratorio, una de estas variables registradas fue el índice de oxigenación PaO₂/FiO₂ que indica la eficacia de intercambio de oxígeno en los pulmones, siendo normal por encima de 300 mmHg. En nuestra muestra, la media fue de 282,39 (DS 87,77), lo cual indica que existe un daño pulmonar agudo que impide un adecuado intercambio gaseoso a nivel respiratorio. Por otro lado, se registraron variables que apoyaban el diagnóstico de una situación

inflamatoria en estos pacientes, como son la proteína C reactiva (PCR) cuya media en nuestro estudio fue de 103,05 (DS 88,73) mg/l siendo sus valores normales entre 0 y 10 mg/l; la LDH o lactato deshidrogenasa cuya media fue de 342,09 (DS 147,53) UI/l siendo sus valores normales entre 110 y 210; y el Dímero D cuya media fue de 1889,25 (DS 4801,7) µg/l siendo límite normal 500 µg/l.

Tabla 4. Características analíticas de los pacientes del estudio.

Variables analíticas	Pacientes	Valores normales
PaO ₂ , media (DS)(mmHg)	73.98(36.4)	83-108
Leucocitos, media(DS)(x10 ⁶ /l)	6.843(3100)	4-10.5
Linfocitos, media(DS)(x10 ⁶ /l)	1.007(581)	0.9-5.2
Plaquetas, media(DS)(x10 ⁹ /l)	201.523(93.746)	150.000-400.000
Hemoglobina, media(DS)(g/dl)	12.74(2.11)	12-18
Dímero D, media(DS)(µg/l)	1.889,25(4.801,7)	0-500
PCR, media(DS)(mgr/l)	103.05(88.73)	0-10
Ferritina, media(DS)(µg/l)	763.40(801.43)	20-300
LDH, media(DS)(UI/l)	342.09(147.53)	110-210
Creatinina, media(DS)(mg/dl)	1.88(11.18)	0.8-1.3
Sodio, media(DS)(mEq/l)	136.52(5.33)	136-146
GPT, media(DS)(UI/l)	35.81(29.57)	1-50
GOT, media(DS)(UI/l)	44.81(25.91)	0-50
IL-6, media(DS)(pg/ml)	31.05(17.28)	<7
SAFI(SpO ₂ /FiO ₂)%, media(DS)	424.22(68.51)	>400
PAFI(PaO ₂ /FiO ₂)%, media(DS)	282.39(87.77)	>300

PaO₂: presión parcial de oxígeno en sangre arterial. PCR: proteína C reactiva. LDH: lactato deshidrogenasa. GPT: Glutamato-Pirúvica Transaminasa. GOT: glutámico-oxalacético transaminasa. IL-6: interleucina 6. SAFI: saturación de oxígeno/fracción inspirada de oxígeno. PAFI: presión parcial de oxígeno/fracción inspirada de oxígeno.

En cuanto a los tratamientos de nuestros pacientes (Tabla 5) , se debe tener en cuenta el periodo en el que se recogieron los datos del estudio, entre marzo y noviembre de 2020, donde algunas de las indicaciones farmacológicas que posteriormente fueron retiradas tras resultados no positivos estaban aún vigentes. Este caso, por ejemplo, es el uso de hidroxicloroquina, en nuestra muestra, administrada a 268 pacientes (84,27 %) del total, cuya justificación se debía a su acción inmunomoduladora. Otro de los fármacos que no mostró evidencia posterior fue el antirretroviral lopinavir/ritonavir, administrado en nuestra muestra en 189 pacientes (59,43 %), cuyo objetivo era inhibir la enzima proteasa del virus disminuyendo su multiplicación viral. Destacar el uso en nuestra población de

tratamiento anticoagulante, bien a dosis profilácticas o a dosis terapéuticas, en 229 pacientes (71,56 %).

Otro de los puntos a señalar en el tratamiento de los pacientes con COVID-19 es la necesidad de oxigenoterapia debido al daño pulmonar que se produce en situación aguda, que es el principal condicionante de la evolución de la enfermedad. En cuanto a los sistemas de aporte de oxígeno de bajo flujo utilizados en nuestra muestra, las gafas nasales se usaron en 186 pacientes (57,76 %) y la mascarilla con reservorio en 68 pacientes (21,12 %). Por otro lado, sistemas de alto flujo como la mascarilla Ventimask fueron utilizados en 68 pacientes (21,12 %) y las gafas nasales de oxígeno a alto flujo en 24 pacientes (7,47 %). En cuanto a ventilación mecánica no invasiva (VMNI), fue necesaria en 38 pacientes (11,84 %) de la muestra, y la necesidad de intubación orotraqueal y ventilación mecánica invasiva en unidades de cuidados intensivos se dio en 43 pacientes (13,4 %).

Tabla 5. Tratamientos recibidos por los pacientes del estudio.

Tratamientos	Pacientes
Antibiótico, n(%)	314 (97,82)
Hidroxiclороquina, n(%)	268 (84,27)
Lopinavir/ritonavir, n(%)	189 (59,43)
Interferon, n(%)	0 (0)
Remdesivir, n(%)	0 (0)
Anticoagulación, n(%)	229 (71,56)
Tocilizumab, n(%)	14 (4,42)
Glucocorticoides, n(%)	98 (30,62)
Gafas nasales/Ventimask/reservorio n(%)	186 (57,76); 68 (21,12); 68 (21,12)
Gafas nasales de alto flujo, n(%)	24 (7,47)
Ventilación mecánica no invasiva, n(%)	38 (11,84)
Ventilación mecánica invasiva, n(%)	43 (13,4)
Oxigenación por membrana extracorpórea, (ECMO) n(%)	7(2,17)

En la Tabla 6 reflejamos la distribución genotípica de los SNPs que se han analizado en el estudio en relación con el riesgo de neumonía por COVID-19, mortalidad por esta causa, reingreso y mortalidad asociada a ese reingreso.

Tabla 6. Distribución genotípica según el riesgo de neumonía por COVID-19, mortalidad, reingreso y mortalidad asociada al reingreso.

SNP	Distribución genotípica, n (%)				
	Todos los pacientes COVID-19, n (%)	Neumonía COVID-19, n (%)	Mortalidad, n (%)	Reingreso, n (%)	Mortalidad relacionada con ese reingreso, n (%)
rs1544410 (CC/CT/TT)	131(38.8) / 151(44.7) / 56(16.6)	88(35.5) / 117(47.2) / 43(17.3)	29(38.2) / 29(38.2) / 18 (23.7)	36(46.8%) / 38(49.4) / 3(3.9)	12 (48.0) / 11 (44.0) / 2(8%)
rs1800795 (CC/CG/GG)	34(10.1) / 156(46.2) / 148(43.8)	26(10.5) / 119(48.0) / 103(41.5)	4(5.3) / 43(56.6) / 29(38.2)	12(15.6) / 32(41.6%) / 33(42.9)	6(24.0) / 10(40.0) / 9(36.0)
rs2069827 (GG/GT/TT)	300(88.8) / 18(5.3) / 20(5.9)	218(87.9) / 29(11.7) / 1(0.4)	72(94.7%) / 4(5.3) / (-)	68(88.3) / 8(10.4) / 1(1.3)	22(88.0) / 3(12.0) / (-)
rs2074192 (CC/CT/TT)	163(48.2) / 78(23.1) / 97(28.7)	117(47.2) / 54(21.8) / 77(31.0)	36(47.4) / 17(22.4) / 23(30.3)	41(53.2) / 15(19.5) / 21(27.3)	10(40.0) / 6(24.0) / 9(36.0)
rs2228570 (GG/GA/AA)	153(45.3) / 147(43.5) 36 (10.7)	113(45.9) / 106(43.1) / 27(11.0)	35(46.7) / 34(45.3) / 6(8.0)	33(42.9) / 34(44.2) / 10(13.0)	10(40.0) / 14(56.0) / 1(4.0)
rs2285666 (TT/TC/CC)	44(13.0) / 54(16.0) / 240(71.0)	30(12.1) / 35(14.1) / 183(73.8)	9(11.8) / 12(15.8%) / 55(72.4)	12(15.6) / 11(14.3) / 54(70.1)	4(16.0) / 5(20.0) / 16(64.0)
rs731236 (GG/GA/AA)	59(175) / 145(42.9) / 133(39.3)	45(18.2) / 116(47.0) / 86(34.8)	20(26.3) / 26(34.2) / 30(39.5)	4(5.2) / 36(46.8) / 37(48.1)	2(8.0) / 11(44.0) / 12(48.0)
rs7975232 (CC/CA/AA)	71(21.0) / 166(49.1) / 101(29.9)	45(18.1) / 128(51.6) / 75(30.2)	17(22.4) / 34(44.7) / 25(32.9)	20(26.0) / 42(54.5) / 15(19.5)	5(20.0) / 15(60.0) / 5(20.0)
rs879922 (CC/CG/GG)	103(30.5) / 60(17.8) / 174(51.5)	77(31.2) / 38(15.4) / 132(53.4)	24(31.6) / 11(14.5) / 41(53.9)	26(34.2) / 13(17.1) / 37(48.7)	6(24.0) / 3(12.0) / 16(64.0)
rs1059047 (TT/TC/CC)	295(87.3) / 20 (5.9) / 21(6.8)	214(87.0) / 32(13.0) / (-)	65(85.5) / 11(14.5) / (-)	69(89.6) / 7(9.1) / 1(1.3)	22(88.0) / 2(8.0) / 1(4.0)
rs111521887 (CC/CG/GG)	246(72.8) / 79(23.4) / 8(2.4)	187(76.6) / 53(21.7) / (-)	49(66.2) / 23(31.1) / 2(2.7)	56(73.7) / 19(25.0) / 1(1.3)	18(72.0) / 5(20.0) / 2(8.0)
rs1143634 (GG/GA/AA)	210(62.1) / 112(33.1) / 14(4.1)	151(61.4) / 84(34.1) / 11(4.5)	45(59.2) / 27(35.5) / 4(5.3)	44(57.1) / 25(32.5) / 8(10.4)	18(72.0) / 6(24.0) 1(4)
rs1205 (TT/TC/CC)	37(10.9) / 148 (43.8) / 151(44.7)	26(10.6) / 117(47.6) / 103(41.9)	7(9.2) / 37(48.7) / 32(42.1)	10(13.0) / 29(37.7) / 38(49.4)	1(4.0) / 11(44.0) / 13(52.0)
rs1544410 (TT/TC/CC)	55(16.3) / 150(44.4) / 131(38.8)	42(17.1) / 116(47.2) / 88(35.8)	18(23.7) / 29(38.2) / 29(38.2)	3(3.9) / 38(49.4) / 36(46.8)	2(8.0) / 11(44.0) / 12(48.0)
rs16944 (GG/GA/AA)	164(48.5) / 143(42.3) / 29(8.6)	120(48.8) / 104(42.3) / 22(8.9)	37(48.7) / 33(43.4) / 6(7.9)	39(50.6) / 32(41.6) / 6(7.8)	12(48.0) / 9(36.0) / 4(16.0)
rs17561 (CC/CA/AA)	190(56.2) / 121(35.8) / 25(7.4)	136(55.3) / 91(37.0) / 19(7.7)	42(55.3) / 28(36.8) / 6(7.9)	42(54.5) / 25(32.5) / 10(13.0)	19(76.0) / 4(16.0) / 2(8.0)

Tabla 6 continuación 1. Distribución genotípica según el riesgo de neumonía por COVID-19, mortalidad, reingreso y mortalidad asociada al reingreso.

SNP	Distribución genotípica, n (%)				
	Todos los pacientes COVID-19, n (%)	Neumonía COVID-19, n (%)	Mortalidad, n (%)	Reingreso, n (%)	Mortalidad relacionada con ese reingreso, n (%)
rs17690703 (CC/CT/TT)	166(49.1) / 139(41.1) / 31(9.2)	131(53.3) / 95(38.6) / 20(8.1)	37(48.7) / 32(42.1) / 7(9.2)	38(49.4) / 30(39.0) / 9(11.7)	9(36.0) / 15(60.0) / 1(4.0)
rs1799724 (TT/TC/CC)	5(1.5) / 68(20.1) / 263(77.8)	3(1.2) / 50(20.3) / 193(78.5)	2(2.6) / 13(17.1) / 61(80.3)	2(2.6) / 15(19.5) / 60(77.9)	1(4.0) / 5(20) / 19(76.0)
rs1800587 (AA/AG/GG)	25(7.4) / 121(35.8) / 190(56.2)	19(7.7) / 91(37.0) / 136(55.3)	6(7.9) / 28(36.8) / 42(55.3)	10(13.0) / 25(32.5) / 42(54.5)	2(8.0) / 4(16.0) / 19(76.0)
rs1800629 (AA/AG/GG)	9(2.7) / 60(17.8) / 267(79.0)	6(2.4) / 46(18.7) / 194(78.9)	3(3.9) / 17(22.4) / 56(73.7)	3(3.9) / 13(16.9) / 61(79.2)	(-) / 4(16.0) / 21(84.0)
rs1800796 (CC/CG/GG)	8(2.4) / 43(12.7) / 285(84.3)	6(2.4) / 34(13.8) / 206(83.7)	1(1.3) / 10(12.2) / 65(85.5)	1(1.3) / 10(13.0) / 66(85.7)	2(8.0) / 3(12.0) / 20(80.0)
rs1800797 (AA/AG/GG)	34(10.1) / 156(46.2) / 146(43.2)	25(10.2) / 120(48.8) / 101(41.1)	5(6.6) / 43(56.6) / 28(36.8)	12(15.6) / 33(42.9%) / 32(41.6)	6(24.0) / 9(36.0) / 9(36.0)
rs1800872 (GG/GT/TT)	173(51.2) / 132(39.1) / 31(9.2)	130(52.8) / 97(39.4) / 19(7.7)	41(53.9) / 28(36.8) / 7(9.2)	46(59.7) / 22(28.6) / 9(11.7)	10(40.0) / 9(36.0) / 6(24.0)
rs1800896 (CC/CT/TT)	51(15.1) / 154(45.6) / 131(38.8)	40(16.3) / 114(46.3) / 92(37.4)	14(18.4) / 30(39.5) / 32(42.1)	13(16.9) / 35(45.5) / 29(37.7)	(-) / 13(52.0) / 12(48.0)
rs1800947 (CC/CG/GG)	301(89.1) / 34(10.1) / 1(0.3)	223(90.7) / 22(8.9) / 1(0.4)	67(88.2) / 9(11.8) / (-)	68(88.3) / 8(10.4) / 1(1.3)	23(92.0) / 2(8.0) / (-)
rs1884082 (GG/GT/TT)	102(30.2) / 157(46.4) / 76(22.5)	77(31.4) / 111(45.3) / 57(23.3)	21(27.6) / 35(46.1) / 20(26.3)	24(31.2) / 37(48.1) / 16(20.8)	5(20.0) / 14(56.0) / 6(24.0)
rs199422297 (GG/GA/AA)	336(100) / (-) / (-)	246(100) / (-) / (-)	76(100) / (-) / (-)	77(100) / (-) / (-)	25(100) / (-) / (-)
rs2069705 (GG/GA/AA)	31(9.2) / 141(41.7) / 163(48.2)	23(9.4) / 102(41.6) / 120(49.0)	6(7.9) / 33(43.4) / 37(48.7)	6(7.8) / 31(40.3) / 40(51.9)	2(8.0) / 9(36.0) / 14(56.0)
rs2227306 (TT/TC/CC)	52(15.4) / 162(47.9) / 122(36.1)	34(13.8) / 121(49.2) / 91(37.0)	11(14.5) / 39(51.3) / 26(34.2)	14(18.2) / 35(45.5) / 28(36.4)	4(16.0) / 8(32.0) / 13(52.0)
rs2227564 (CC/CT/TT)	226(66.9) / 101(29.9) / 9(2.7)	170(69.1) / 70(28.5) / 6(2.4)	51(67.1) / 23(30.3) / 2(2.6)	51(66.2) / 25(32.5) / 1(1.3)	17(68.0) / 7(28.0) / 1(4.0)
rs2228145 (CC/CA/AA)	52(15.4) / 168(49.7) / 116(34.3)	34(13.8) / 119(48.4) / 93(37.8)	10(13.2) / 39(51.3) / 27(35.5)	12(15.6) / 44(57.1) / 21(27.3)	8(32.0) / 15(60) / 2(8.0)
rs2228570 (AA/AG/GG)	36(10.7) / 147(43.5) / 152(45.0)	27(11.0) / 106(43.4) / 112(45.7)	6(7.9) / 34(45.3) / 35(46.7)	10(13.0) / 34(44.2) / 33(42.9)	1(4.0) / 14(56.0) / 10(40.0)
rs2282679 (GG/GT/TT)	28(8.3) / 133(39.3) / 175(51.8)	22(8.9) / 103(41.9) / 121(49.2)	7(9.2) / 29(38.2) / 40(52.6)	9(11.7) / 30(39.0) / 38(49.4)	1(4.8) / 8(32.0) / 16(64.0)
rs2430561 (TT/TA/AA)	91(26.9) / 161(47.6) / 84(24.9)	65(26.5) / 123(50.0) / 58(23.6)	20(26.3) / 37(48.7) / 19(25.0)	20(26.0) / 32(41.6) / 25(32.5)	5(20.0) / 12(48.0) / 8(32.0)
rs2735940 (GG/GA/AA)	74(21.9) / 169(50.0) / 92(27.2)	55(22.4) / 121(49.4) / 69(28.2)	16(21.1) / 41(53.9) / 19(25.0)	14(18.2) / 40(51.9) / 23(29.9)	5(20.0) / 15(60.0) / 5(20.0)
rs2794521 (TT/TC/CC)	181(53.6) / 136(40.2) / 19(5.6)	131(53.3) / 104(42.3) / 11(4.5)	47(61.8) / 24(31.6) / 5(6.6)	43(55.8) / 29(37.7) / 5(6.5)	13(52.0) / 10(40.0) / 2(8.0)
rs35697037 (GG/GA/AA)	178(52.7) / 77(22.8) / 81(24.0)	128(52.0) / 53(21.5) / 65(26.4)	41(53.9) / 18(23.7) / 17(22.4)	42(54.5) / 16(20.8) / 19(24.7)	10(40.0) / 7(28.0) / 8(32.0)

Tabla 6 continuación 2. Distribución genotípica según el riesgo de neumonía por COVID-19, mortalidad, reingreso y mortalidad asociada al reingreso.

SNP	Distribución genotípica, n (%)				
	Todos los pacientes COVID-19, n (%)	Neumonía COVID-19, n (%)	Mortalidad, n (%)	Reingreso, n (%)	Mortalidad relacionada con ese reingreso, n (%)
rs35705950 (GG/GT/TT)	265(78.4) / 68(20.1) / 3(0.9)	191(77.6) / 53(21.5) / 2(0.8)	59(77.6) / 16(21.1) / 1(1.3)	59(76.6) / 18(23.4) / (-)	19(76.0) / 4(16.0) / 2(8.0)
rs373740199 (CC/CT/TT)	336(100) / (-) / (-)	246 (100) / (-) / (-)	76(100) / (-) / (-)	77(100) / (-) / (-)	25(100) / (-) / (-)
rs386713 (CC/CT/TT)	40(11.8) / 134(39.6) / 160(47.3)	28(11.5) / 94(38.5) / 122(50.0)	11(14.5) / 26(35.7) / 38(50.7)	8(10.4) / 34(44.2) / 35(45.5)	3(12.0) / 11(44.0) / 11(44.0)
rs398123017 (CC/CT/TT)	336(100) / (-) / (-)	246 (100) / (-) / (-)	76(100) / (-) / (-)	77(100) / (-) / (-)	25(100) / (-) / (-)
rs408265 (CC/CT/TT)	155(45.9) / 136(40.2) / 45(13.3)	117(47.2) / 95(38.6) / 34(13.8)	37(48.7) / 29(38.2) / 10(13.2)	33(42.9) / 36(46.8) / 8(10.4)	11(44.0) / 8(32.0) / 6(24.0)
rs419598 (TT/TC/CC)	175(51.8) / 141(41.7) / 20(5.9)	127(51.6) / 104(42.3) / 15(6.1)	39(51.3) / 28(36.8) / 9(11.8)	42(54.5) / 33(42.9) / 2(2.6)	15(60.0) / 9(36.0) / 1(4.0)
rs4588 (GG/GT/TT)	173(51.2) / 132(39.1) / 28(8.3)	120(49.2) / 102(41.8) / 22(9.0)	40(52.6) / 29(38.2) / 7(9.2)	37(48.7) / 30(39.5) / 9(11.8)	16(64.0) / 8(32.0) / 1(4.0)
rs5743890 (CC/CT/TT)	10(3.0) / 94 (27.8) / 232 (68.6)	7(2.8) / 68(27.6) / 171(69.5)	4(5.3) / 21(27.6) / 51(67.1)	1(1.3) / 17(22.1) / 59(76.6)	(-) / 7(28.0) / 18(72.0)
rs5743894 (TT/TC/CC)	245(72.5) / 83(24.6) / 8(2.4)	186(75.6) / 56(22.8) / 4(1.6)	49(66.2) / 25(32.9) / 2(2.6)	55(71.4) / 21(27.3) / 1(1.3)	18(72.0) / 5(20.0) / 2(8.0)
rs7041 (CC/CA/AA)	114(33.7) / 172(50.9) / 50(14.8)	72(29.3) / 134(54.5) / 40(16.3)	27(35.5) / 35(46.1) / 14(18.4)	22(28.6) / 45(58.4) / 10(13.0)	11(44.0) / 12(48.0) / 2(8.0)
rs727503468 (CC/CT/TT)	336(100) / (-) / (-)	246 (100) / (-) / (-)	76(100) / (-) / (-)	77(100) / (-) / (-)	25(100) / (-) / (-)
rs760506977 (CC/CT/TT)	336(100) / (-) / (-)	246 (100) / (-) / (-)	76(100) / (-) / (-)	77(100) / (-) / (-)	25(100) / (-) / (-)
rs876661305 (GG/GA/AA)	336(100) / (-) / (-)	246 (100) / (-) / (-)	76(100) / (-) / (-)	77(100) / (-) / (-)	25(100) / (-) / (-)

A continuación se presentan los diferentes algoritmos basados en los SNPs analizados para predecir el riesgo de neumonía por COVID-19, mortalidad, reingreso en el siguiente año y mortalidad en dicho reingreso.

5.1. Riesgo de neumonía por COVID-19

En la Figura 8, se reflejan SNPs más influyentes en el riesgo de sufrir neumonía, están ordenados por orden creciente de influencia. Dentro de ellos destacan polimorfismos en el receptor IL-6, en la proteína de la unión de vitamina D, en el gen *PCR*, en IL-6, en el

receptor de vitamina D y en el gen *IFN-γ*. Hemos realizado la agrupación en diferentes colores en función de su vía de acción: el rojo para la vía de la inflamación, el verde para el metabolismo de la vitamina D y el azul para el metabolismo de la ECA2. El eje horizontal (eje X) indica la importancia de esos polimorfismos, sin condicionar el efecto positivo o negativo que tienen. Los más relevantes fueron el rs2228145 del gen *IL6R*, particularmente el genotipo AA, seguido por el genotipo AA del rs7041 del gen *VDBP*; y, a continuación, otros como el rs1205 en el gen *PCR*, el rs1800795 y rs1800797 en *IL6*, el rs731236 en *VDR*, el rs2282679 en *VDBP* y el rs2430661 en *IFN-γ*.

En este caso podemos decir que los polimorfismos que más influyen en el desarrollo de neumonía por COVID-19 son aquellos relacionados con el sistema inflamatorio y el sistema de la vitamina D.

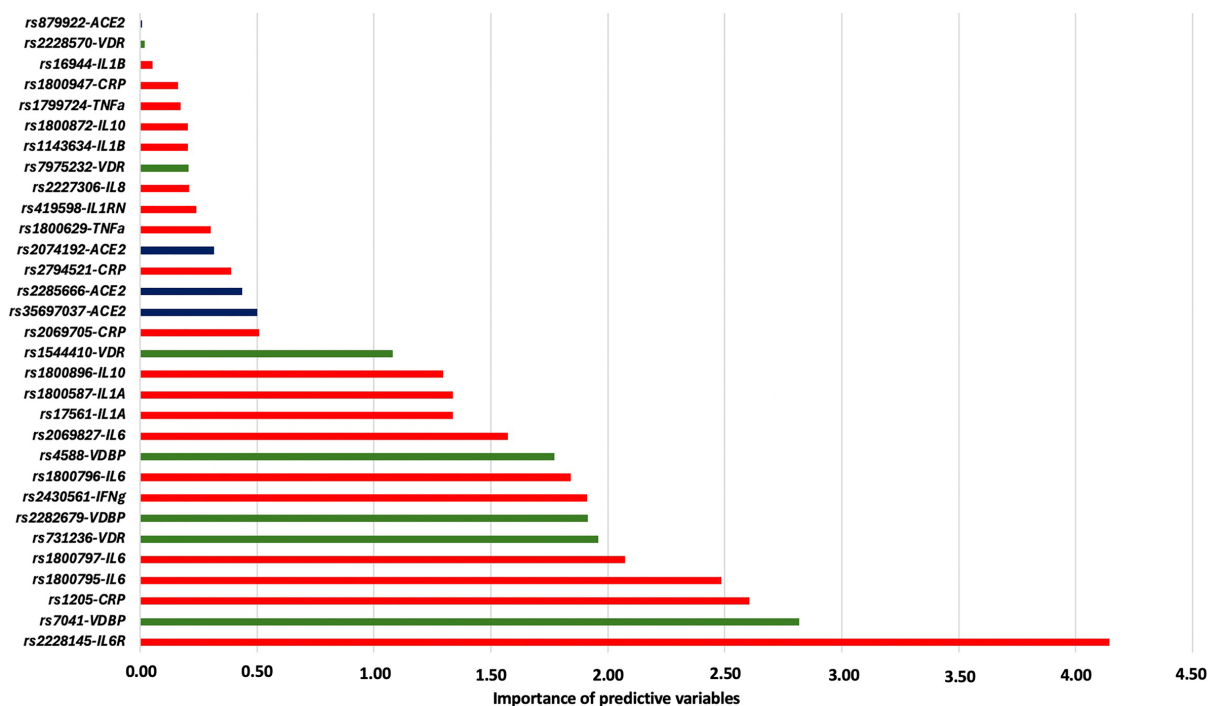


Figura 8. Polimorfismos genéticos relacionados con el riesgo de neumonía por COVID-19. Color rojo respuesta inflamatoria, color verde metabolismo vitamina D y color azul SRAA.

Se utilizaron diferentes modelos de aprendizaje automático para evaluar el riesgo de neumonía por COVID-19 con base en los polimorfismos genéticos estudiados en nuestro estudio (Tabla 7). Nuestros resultados mostraron que el mejor modelo en la predicción de neumonía es el XGB; fue con el modelo con las mejores métricas de rendimiento con las

que se obtuvo, siendo el área bajo la curva (AUC) de 0,86, seguido del modelo KNN con un AUC de 0,82.

Tabla 7. Resultados de los diferentes métodos de aprendizaje automático utilizados para evaluar el riesgo de neumonía por COVID-19.

	Método	BA (%)	Recall	Precision	AUC	F1score	MCC	DYI	Kappa
Neumonía por COVID-19	SVM	79,09	79,18	78,53	0,79	78,85	70,18	79,09	70,41
	DT	77,23	77,32	76,68	0,76	77,00	68,53	77,23	68,75
	GNB	71,16	71,25	70,65	0,71	70,95	63,14	71,16	63,35
	KNN	82,04	82,13	81,45	0,82	81,79	72,79	82,04	73,04
	XGB	86,10	86,20	85,48	0,86	85,84	76,40	86,10	76,65

BA: Balanced Accuracy. AUC: Área Bajo la Curva. MCC: Coeficiente de correlación de Mathews. DYI: Índice de Youden. SVM: Support Vector Machine. DT: Decision Tree. GNB: Gaussian Naïve Bayes. KNN: K-Nearest Neighbors. XGB: Extreme Gradient Boosting.

En la representación de las curvas ROC o *receiver operating characteristic curve* de la comparativa de los diferentes modelos de aprendizaje (Figura 9). Cada curva representa la relación entre la sensibilidad o verdaderos positivos (eje vertical) y la tasa de falsos positivos o 1-especificidad (eje horizontal). Es el modelo XGB el que representa el mayor área bajo la curva (AUC), en este caso, de 0,86, lo cual se traduce en que para un mismo valor de especificidad alcanza valores de sensibilidad superiores a los de otros modelos. Cuando en un modelo el AUC se aproxima a 1, significa que presenta un equilibrio adecuado entre sensibilidad y especificidad, lo cual refleja una alta capacidad discriminativa en la predicción del riesgo de neumonía por COVID-19.

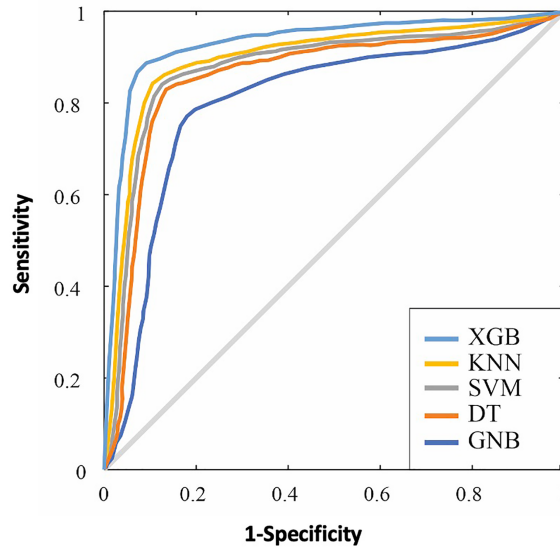


Figura 9. Comparativa de las curvas de ROC de los diferentes modelos de aprendizaje automático evaluados en la predicción de riesgo de neumonía por COVID-19. SVM: Support Vector Machine. DT: Decision Tree. GNB: Gaussian Naïve Bayes. KNN: K-Nearest Neighbors. XGB: Extreme Gradient Boosting.

Para evaluar los diferentes modelos de aprendizaje automático, también se agruparon las métricas de rendimiento obtenidas tanto en la fase de entrenamiento como en la fase de prueba (Figura 10) y las representamos en un diagrama de radar. Comparando ambos diagramas, observamos que las métricas de la fase de prueba son superponibles a las métricas de la fase de entrenamiento, es decir, no existe sobreajuste en la comparativa. En ambos casos observamos cómo el modelo eXtreme Gradient Boosting o XGB es el que mejor equilibrio refleja de todas las métricas en comparación con el resto de modelos, seguido del modelo K-Nearest Neighbors o KNN. Por otro lado, el modelo Gaussian Naïve Bayes o GNB es el que menor capacidad predictiva presenta en este caso.

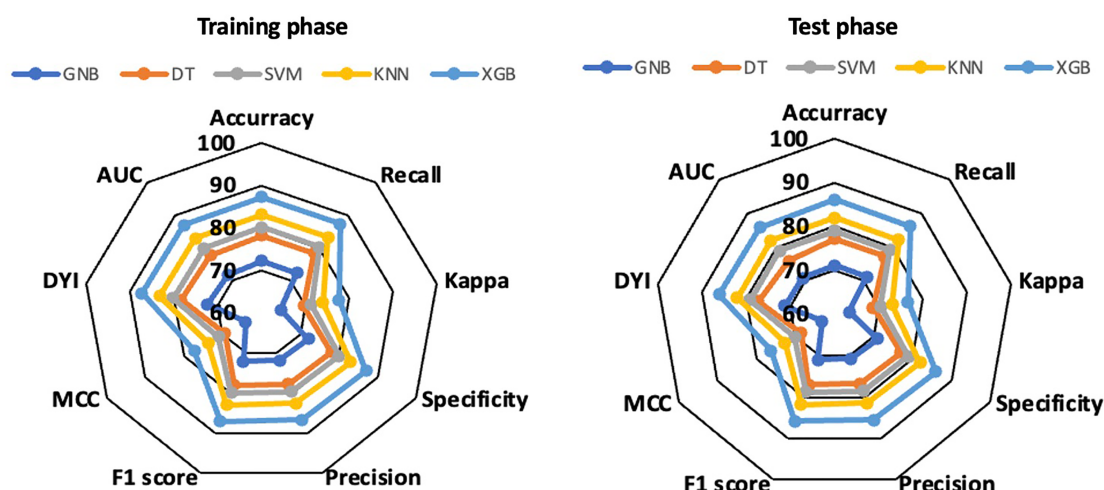


Figura 10. Diagrama de radar de la fase de entrenamiento y fase de prueba. AUC: Área Bajo la Curva. MCC: Coeficiente de correlación de Mathews. DYI: Índice de Youden. SVM: Support Vector Machine. DT: Decision Tree. GNB: Gaussian Naïve Bayes. KNN: K-Nearest Neighbors. XGB: Extreme Gradient Boosting.

5.2. Riesgo de mortalidad

En la Figura 11 se muestran los SNPs que más se han relacionado con la mortalidad por COVID-19. El polimorfismo más significativo fue el rs2069827 del gen *IL6*, específicamente el genotipo *TT*, seguido del genotipo *CC* del polimorfismo rs2794521 del gen *PCR*. Otras variantes que destacaron fueron los polimorfismos rs2074192, rs35697037 y rs2285666 en el gen *ECA2*; el rs1800872 del gen *IL10*; el rs2228570 del gen *VDR*; el rs1800587 y el rs17561 del gen *IL1A*; así como el rs1800796 y el rs1800797 del gen *IL6* y el rs7041 del gen *VDBP*. Teniendo en cuenta eso, podemos decir que los SNPs que más se han relacionado con la mortalidad son aquellos que involucran al sistema inflamatorio y el metabolismo de la *ECA2*.

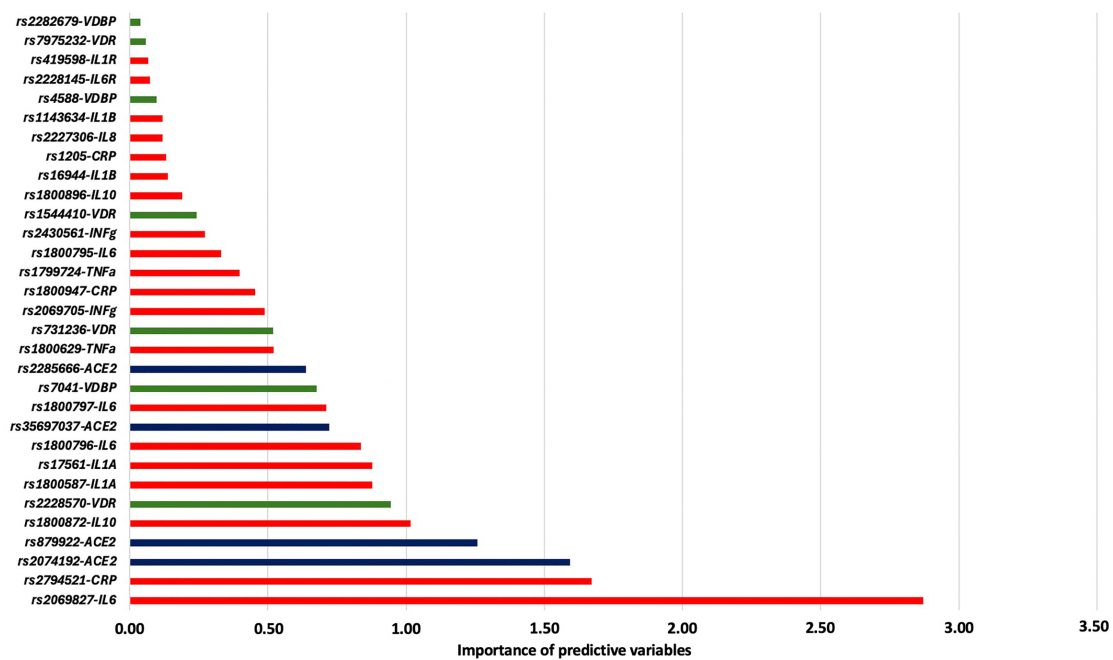


Figura 11. Polimorfismos genéticos relacionados con la mortalidad por COVID-19. Color rojo respuesta inflamatoria, color verde metabolismo vitamina D y color azul SRAA.

De nuevo se evaluaron los diferentes modelos de aprendizaje automático para evaluar la relación de los polimorfismos de nuestro estudio con la mortalidad (Tabla 8). Se observa en los resultados según el AUC que el modelo de aprendizaje que mejor valor predictivo tiene es el XGB, con un resultado de 0,86, seguido del modelo KNN con un AUC de 0,81.

Tabla 8. Resultados de los diferentes métodos de aprendizaje automático utilizados para evaluar el riesgo de mortalidad.

	Método	BA (%)	Recall	Precision	AUC	F1score	MCC	DYI	Kappa
Mortalidad	SVM	79,20	79,29	78,64	0,79	78,96	70,28	79,20	70,51
	DT	77,13	77,22	76,58	0,76	76,90	68,44	77,13	68,67
	GNB	70,18	70,26	69,68	0,70	69,97	62,27	70,18	62,48
	KNN	81,17	81,27	80,59	0,81	80,93	72,02	81,17	72,26
	XGB	86,00	86,10	85,39	0,86	85,74	76,31	86,00	76,56

BA: Balanced Accuracy. AUC: Área Bajo la Curva. MCC: Coeficiente de correlación de Mathews. DYI: Índice de Youden. SVM: Support Vector Machine. DT: Decision Tree. GNB: Gaussian Naïve Bayes. KNN: K-Nearest Neighbors. XGB: Extreme Gradient Boosting.

Al representar las curvas de ROC comparando cada modelo de aprendizaje, observamos que el modelo XGB presenta un área bajo la curva de 0,86, siendo el mejor valor respecto a los otros modelos, indicando mayor sensibilidad respecto a otros modelos (Figura 12).

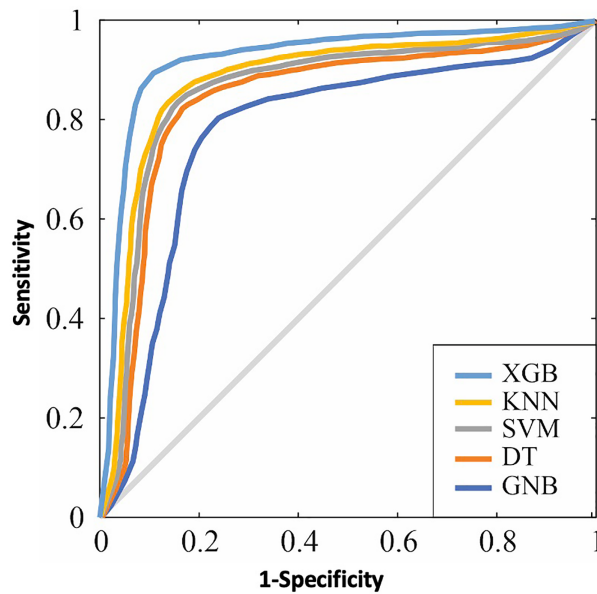


Figura 12. Comparativa de las curvas de ROC de los diferentes modelos de aprendizaje automático evaluados en la predicción de riesgo de mortalidad. SVM: Support Vector Machine. DT: Decision Tree. GNB: Gaussian Naïve Bayes. KNN: K-Nearest Neighbors. XGB: Extreme Gradient Boosting.

Comparamos de nuevo las métricas de los métodos de aprendizaje mediante diagramas de radar, en la fase de entrenamiento y en la fase de prueba (Figura 13), mostrando valores similares en ambos casos y siendo el modelo XGB el que mejores resultados obtiene.

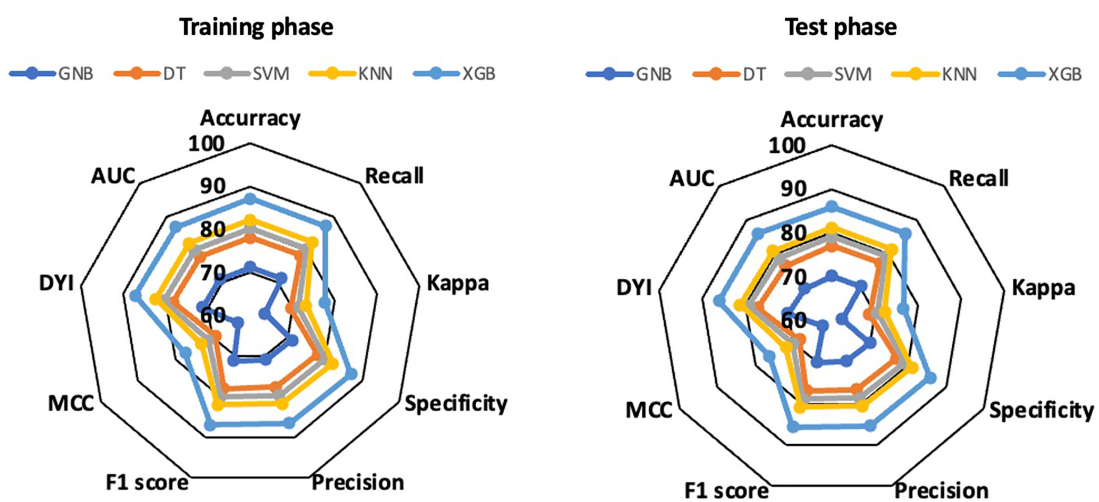


Figura 13. Diagrama de radar de la fase de entrenamiento y fase de prueba. AUC: Área Bajo la Curva. MCC: Coeficiente de correlación de Mathews. DYI: Índice de Youden. SVM: Support Vector Machine. DT: Decision Tree. GNB: Gaussian Naïve Bayes. KNN: K-Nearest Neighbors. XGB: Extreme Gradient Boosting.

5.3. Riesgo de reingreso durante el primer año tras COVID-19

Tal y como hemos recogido en la primera parte de los resultados, durante el primer año de seguimiento de nuestra muestra, el 29,3 % de los pacientes requirieron de nuevo ingreso por COVID-19. En este punto, los polimorfismos que se identificaron con mayor influencia en el riesgo de reingreso han sido el rs1544410, el rs731236 y el rs7975232 del gen *VDR*, siendo los genotipos TT, AA, y AA, respectivamente, los que se asocian a este desenlace. Estos y otros polimorfismos se recogen en la Figura 14, en la que observamos que la mayoría de los polimorfismos que influyen en este factor se relacionan con la vía inflamatoria (particularmente en los genes *IL6*, *IL1B*, *PCR*), aunque el más influyente es el gen *VDR*, del metabolismo de la vitamina D.

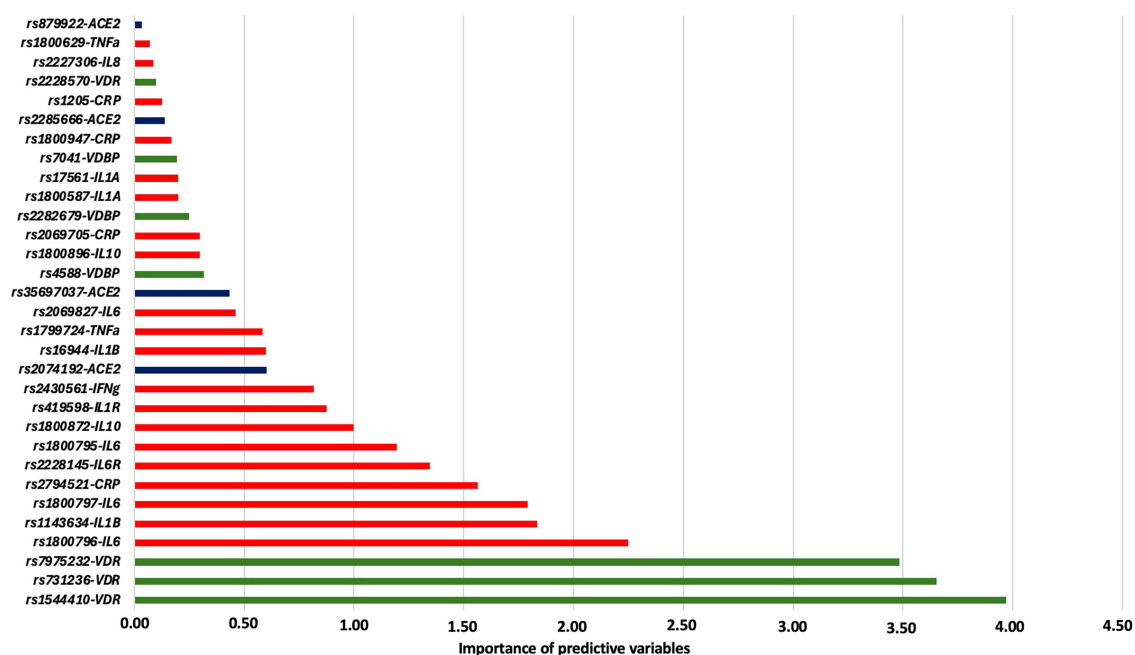


Figura 14. Polimorfismos genéticos relacionados con el reingreso en el primer año tras COVID-19. Color rojo respuesta inflamatoria, color verde metabolismo vitamina D y color azul SRAA.

En la Tabla 9 reflejamos los resultados de los modelos de aprendizaje utilizados para evaluar el riesgo de reingreso en el primer año tras la infección por SARS-CoV-2.

Tabla 9. Resultados de los diferentes métodos de aprendizaje automático utilizados para evaluar el riesgo de reingreso en el primer año tras COVID-19.

	Método	BA (%)	Recall	Precision	AUC	F1score	MCC	DYI	Kappa
Reingreso durante el primer año	SVM	79,26	79,35	78,69	0,78	79,02	70,33	79,26	70,56
	DT	77,48	77,58	76,93	0,77	77,25	68,75	77,48	68,98
	GNB	72,78	72,86	72,25	0,72	72,56	64,46	72,78	64,68
	KNN	81,94	82,04	81,36	0,81	81,69	72,71	81,94	72,95
	XGB	85,86	85,96	85,25	0,85	85,61	76,19	85,86	76,44

BA: Balanced Accuracy. AUC: Área Bajo la Curva. MCC: Coeficiente de correlación de Mathews. DYI: Índice de Youden. SVM: Support Vector Machine. DT: Decision Tree. GNB: Gaussian Naïve Bayes. KNN: K-Nearest Neighbors. XGB: Extreme Gradient Boosting.

En la representación de las curvas ROC de cada modelo de aprendizaje, en este caso, obtenemos un valor de AUC de 0,85 en el modelo XGB. En la Figura 15 comparamos todos los modelos estudiados.

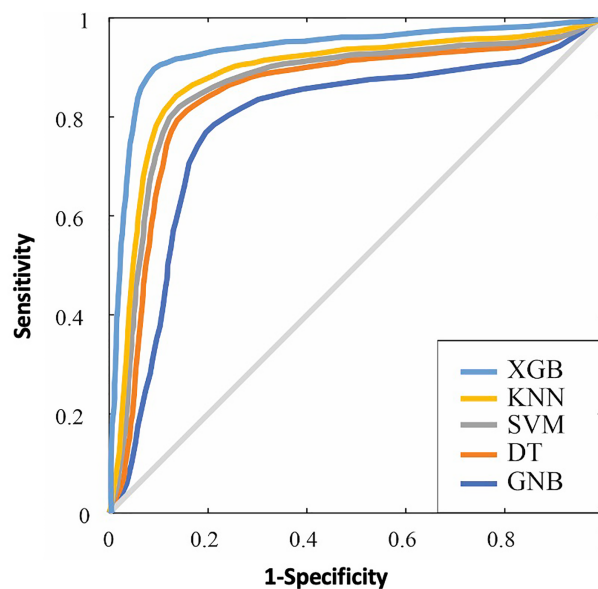


Figura 15. Comparativa de las curvas de ROC de los diferentes modelos de aprendizaje automático evaluados en la predicción de riesgo de reingreso tras un año de COVID-19.

SVM: Support Vector Machine. DT: Decision Tree. GNB: Gaussian Naïve Bayes. KNN: K-Nearest Neighbors. XGB: Extreme Gradient Boosting.

En cuanto a la comparación de la fase de entrenamiento y de la fase de prueba (Figura 16), donde se evalúan los diferentes valores que reflejan la capacidad predictiva de cada

modelo, observamos que en ambos casos es el modelo XGB el que presenta mayor precisión frente al modelo GNB, que es el que menor capacidad predictiva refleja.

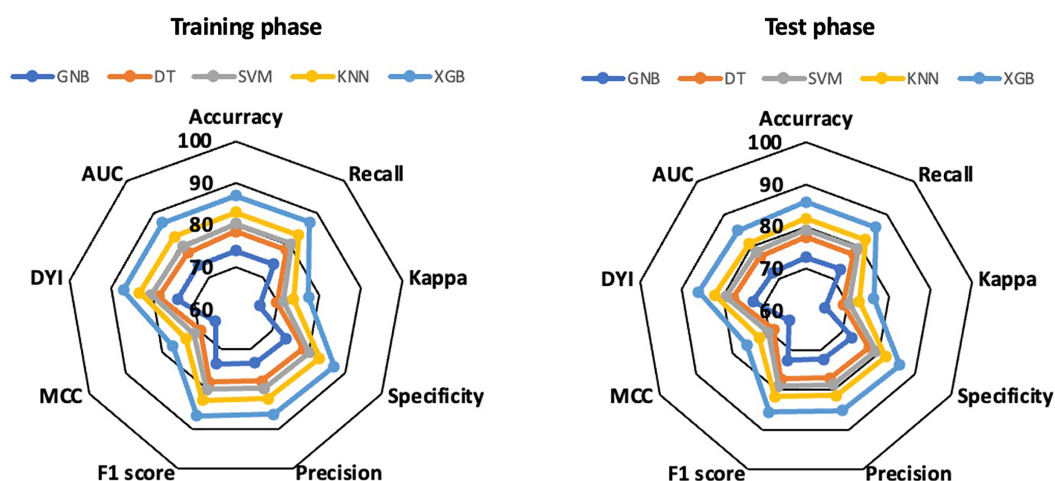


Figura 16. Diagrama de radar de la fase de entrenamiento y fase de prueba. AUC: Área Bajo la Curva. MCC: Coeficiente de correlación de Mathews. DYI: Índice de Youden. SVM: Support Vector Machine. DT: Decision Tree. GNB: Gaussian Naïve Bayes. KNN: K-Nearest Neighbors. XGB: Extreme Gradient Boosting.

5.4. Riesgo de mortalidad durante el reingreso en el primer año tras COVID-19

De los pacientes de nuestro estudio que requirieron reingreso en el primer año, un 32,4 % falleció durante este nuevo ingreso. Los polimorfismos que más se han relacionado con este factor han sido aquellos que en el sistema inflamatorio y en la ECA2. En concreto el que más se ha asociado ha sido el rs2228145 del gen *IL6R*; seguido de otros como el rs17561 y el rs1800587 del gen *IL1A*, el rs35697037, el rs2074192 y el rs879922 del gen *ACE2*, el rs1143634 del gen *IL1B*, el rs1800896 y el rs1800872 del gen *IL10* y el rs2227306 del gen *IL8*, todos ellos recogidos en la Figura 17.

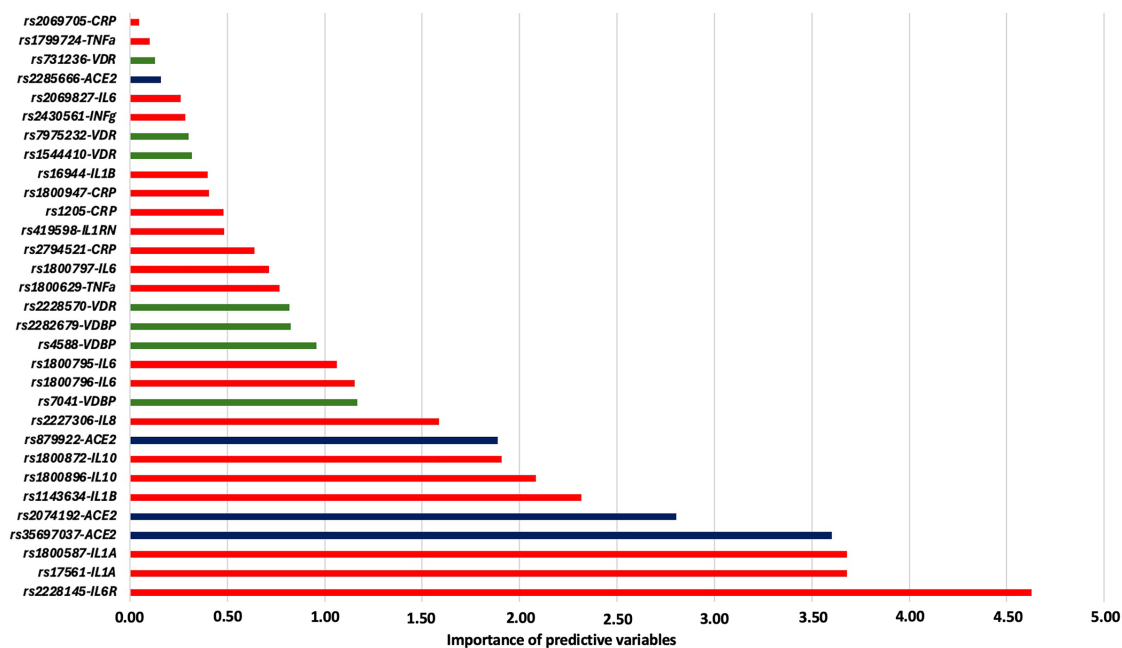


Figura 17. Polimorfismos genéticos relacionados con la mortalidad durante el reingreso en el primer año tras COVID-19. Color rojo respuesta inflamatoria, color verde metabolismo vitamina D y color azul SRAA.

Los modelos de aprendizaje automático evaluados en el estudio y sus valores que muestran su capacidad predictiva del riesgo de mortalidad durante el reingreso están reflejados en la Tabla 10.

Tabla 10. Resultados de los diferentes métodos de aprendizaje automático utilizados para evaluar el riesgo de mortalidad durante el reingreso en el primer año tras COVID-19.

	Método	BA (%)	Recall	Precision	AUC	F1score	MCC	DYI	Kappa
Mortalidad durante el reingreso	SVM	79,85	79,94	79,28	0,80	79,61	70,85	79,85	71,09
	DT	78,12	78,21	77,56	0,78	77,88	69,31	78,12	69,54
	GNB	70,83	70,91	70,32	0,71	70,62	62,85	70,83	63,05
	KNN	81,17	81,27	80,59	0,81	80,93	72,02	81,17	72,26
	XGB	86,85	86,72	86,23	0,86	85,59	77,06	86,85	77,32

BA: Balanced Accuracy. AUC: Área Bajo la Curva. MCC: Coeficiente de correlación de Mathews. DYI: Índice de Youden. SVM: Support Vector Machine. DT: Decision Tree. GNB: Gaussian Naïve Bayes. KNN: K-Nearest Neighbors. XGB: Extreme Gradient Boosting.

En las curvas ROC obtenidas en la evaluación de los modelos de aprendizaje (Figura 18) se demuestra que el modelo XGB presenta un AUC superior al de los demás modelos, siendo este de 0,86.

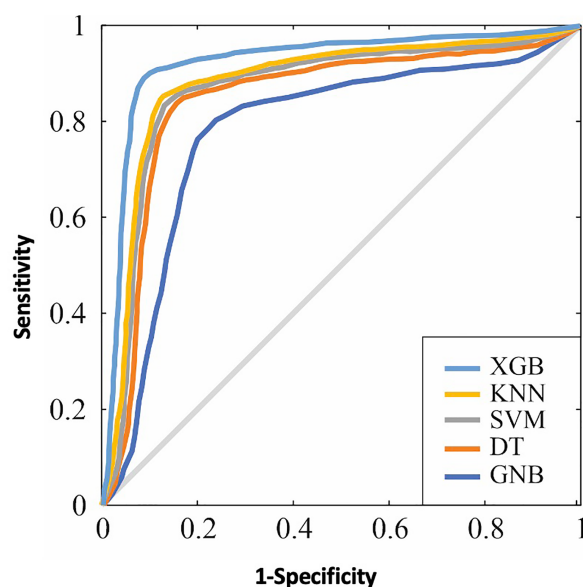


Figura 18. Comparativa de las curvas de ROC de los diferentes modelos de aprendizaje automático evaluados en la predicción del riesgo de mortalidad durante el reingreso tras un año de COVID-19.

SVM: Support Vector Machine. DT: Decision Tree. GNB: Gaussian Naïve Bayes. KNN: K-Nearest Neighbors. XGB: Extreme Gradient Boosting.

Comparando de nuevo los resultados de las métricas obtenidos mediante los diagramas de radar de cada modelo de aprendizaje observamos que el modelo XGB es el que mejores resultados obtiene tanto en la fase de entrenamiento como en la fase de prueba (Figura 19), es decir, presenta una mejor capacidad predictiva del riesgo de mortalidad durante el reingreso y además no existe sobreajuste entre ambas fases.

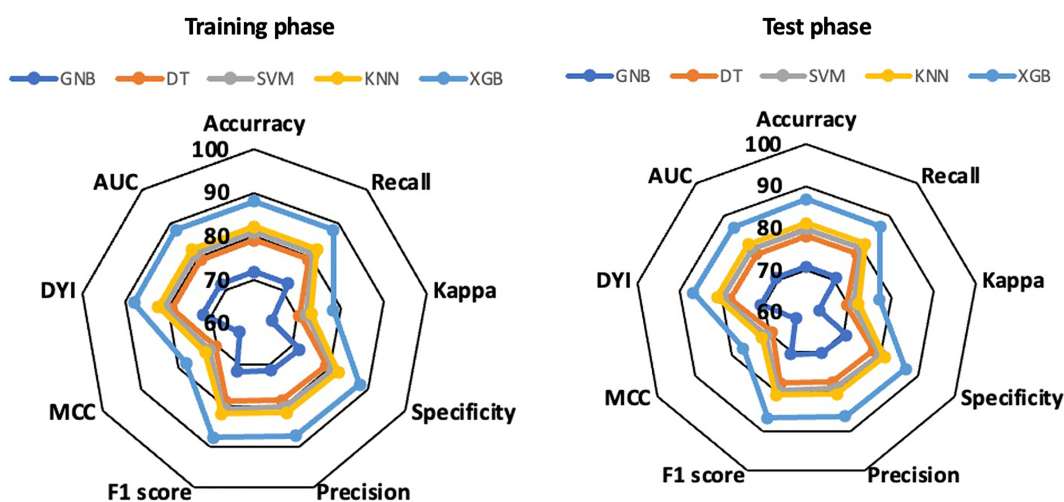


Figura 19. Diagrama de radar de la fase de entrenamiento y de la fase de prueba. AUC: Área Bajo la Curva.

MCC: Coeficiente de correlación de Mathews. DYI: Índice de Youden. SVM: Support Vector Machine. DT: Decision Tree. GNB: Gaussian Naïve Bayes. KNN: K-Nearest Neighbors. XGB: Extreme Gradient Boosting.

Tras el análisis de los diferentes SNPs en cada situación, evaluados por diversos modelos de aprendizaje automático, los resultados demuestran que el modelo XGB es el más preciso para predecir tanto el riesgo de neumonía por COVID-19 como su mortalidad, así como el riesgo de reingreso tras el primer año y la mortalidad durante este reingreso. Además de demostrar más precisión que los otros modelos, también comprobamos su comportamiento estable tanto en la fase de entrenamiento como en la fase de prueba. Así mismo, demuestra un equilibrio constante y adecuado entre la sensibilidad y la especificidad, lo cual es importante para identificar pacientes con riesgo real sin incrementar el número de falsos positivos. Todo ello nos ha permitido desarrollar unos modelos de predicción, en concreto el modelo XGB, con unas AUC de 0,86 para el riesgo de neumonía por COVID-19 y 0,86 para el riesgo de mortalidad por esta causa, así como un AUC de 0,85 y 0,86, respectivamente, para el riesgo de reingreso durante el primer año tras el COVID-19 y la mortalidad en este reingreso.

DISCUSIÓN

La enfermedad COVID-19 es una enfermedad infecciosa causada por el virus SARS-CoV-2, culpable de la pandemia iniciada en marzo de 2020 que supuso un reto a la comunidad científica en cuanto al manejo y control de esta, y que a día de hoy ha cambiado la percepción de la población sobre este tipo de enfermedades (1,2). Su vía principal de transmisión es la vía aérea y presenta una elevada tasa de contagio debido a su alta capacidad de replicación en la vía aérea del huésped (90). La presentación de la COVID-19 es variable; puede presentarse desde casos asintomáticos o leves, como sucede en la mayoría de la población infectada, hasta casos que requieren de hospitalización e ingreso en unidades de cuidados intensivos y necesidad de ventilación mecánica invasiva, debido a una de las complicaciones más frecuentes, como es el SDRA. Los síntomas más comunes recogidos en un metaanálisis son fiebre, presente en un 81,2 % de los casos; tos, con un 58,5 %; y fatiga y disnea, con un 38,5 % y un 26,1 % respectivamente (25). Estos datos son superponibles a nuestra muestra, ya que el síntoma más frecuente fue la febrícula con un 77,97 %, seguido de la disnea y de la tos con un 53,57 % y un 51,93 % respectivamente. En cuanto al manejo de la enfermedad, varios estudios han considerado diversos tratamientos como antivirales, corticoides y anticuerpos monoclonales sin un claro tratamiento efectivo. Bien es cierto que ningún resultado mostró un tratamiento único efectivo, aunque sí que la elección de una combinación de estos en función del momento en el que se encuentre la enfermedad podría tener mejores resultados; para ello, es crucial conocer la fisiopatología y la forma de evolución de la enfermedad (91, 92).

Desde el inicio de la aparición de la enfermedad se ha establecido que la variabilidad clínica que se observa está relacionada con diferentes factores como la edad, comorbilidades o el sexo (36-38). En este sentido, también se ha descrito que la variabilidad genética, asociada a los factores previamente expuestos, desempeña un papel importante en el desarrollo de la enfermedad. Conforme avancemos en estos conocimientos genéticos, podremos comprender mejor la presentación de la enfermedad en la población y los mecanismos patogénicos, sirviendo de ayuda para predecir el riesgo de la enfermedad, así como su abordaje terapéutico (41,42). Ya se conoce de forma más profunda el funcionamiento del virus, su relación con el SRAA tanto para la infección del huésped como para la regulación de la inflamación y edema del intersticio pulmonar; así como la acción inmunomoduladora de la vitamina D a nivel de la inmunidad innata como

de la adquirida; como la respuesta inflamatoria con un aumento de citoquinas inflamatorias y linfocitos T en el tejido pulmonar. Todo ello ha permitido que se comience a estudiar cómo esa variabilidad clínica y de la severidad en la evolución de los pacientes tienen un importante componente genético (64,65).

En este contexto, nuestro estudio refuerza el trasfondo genético de la patogénesis de la enfermedad, sobre todo aquellas variantes genéticas relacionadas con el SRAA, el metabolismo de la vitamina D y el sistema inflamatorio. Nos hemos apoyado en la IA, en concreto, en el aprendizaje automático para identificar las variables genéticas que más influyen en la presentación de la enfermedad y desarrollar una herramienta que nos ayude a predecir e identificar pacientes de alto riesgo. Desde el inicio del desarrollo de la IA en la medicina en los años 1960 hasta la actualidad, ha demostrado ser un mecanismo de apoyo para el clínico, tanto a nivel diagnóstico como en la toma de decisiones terapéuticas y, en concreto, en la investigación biomédica. Esta forma de trabajo ha ido evolucionando de forma progresiva, mejorando su acceso a los datos gracias a la disminución de la codificación manual, al desarrollo de *redes neuronales* y árboles de decisión, a la mejora en las técnicas de almacenamiento de datos y su procesamiento y al desarrollo de algoritmos (75,78).

Nuestros resultados identificaron polimorfismos en determinados genes como los más relevantes en el curso de la enfermedad, particularmente diecinueve variantes genéticas en diez genes diferentes fueron las más influyentes en el pronóstico de los pacientes con COVID-19, destacando las siguientes: *IL6R* (rs2228145), *IL6* (rs1800797 y rs1800796), *PCR* (rs2794521 y rs1800947), *IFN- γ* (rs2430561), *IL1A* (rs17561 y rs1800587), *TNF- α* (rs1800629), *IL1R* (rs419598), *ECA2* (rs35697037 y rs2285666), *VDR* (rs1544410, rs7975232 y rs2228570) y *VDBP* (rs7041, rs2282679 y rs4588). Estos hallazgos ponen de manifiesto el papel fundamental de los polimorfismos genéticos relacionados con la inflamación, el metabolismo de la vitamina D y el receptor ECA2 en la evolución clínica de la COVID-19, y su relación con el riesgo de neumonía, mortalidad, reingreso durante el primer año tras la infección y mortalidad durante ese reingreso.

Como hemos descrito previamente, el SRAA actúa tanto en la capacidad de infección del virus como en el daño pulmonar. El virus SARS-CoV-2 penetra en las células del huésped, en concreto, en la mucosa nasofaríngea y los neumocitos alveolares tipo II, que expresan en su superficie receptores ECA2 por los cuales este virus presenta gran

afinidad, favoreciendo su escisión y replicación (46,47). A su vez, gracias a la ECA se produce una activación de la angiotensina II y, por consiguiente, de sus receptores AT1, cuya función implica vasoconstricción e inflamación pulmonar, favoreciendo en estos pacientes el SDRA y, por consiguiente, una peor evolución clínica (49-51). Se ha descrito que diferentes polimorfismos en el gen *ECA2* pueden alterar su capacidad de infección y expresión de la enfermedad, produciendo una variación en la susceptibilidad y severidad de la enfermedad (64-66). Con este estudio y gracias a los métodos de aprendizaje automático utilizados, hemos demostrado que los polimorfismos rs35697037 y rs2285666 del gen *ECA2* fueron los más influyentes en la hospitalización y la mortalidad de nuestros pacientes. Estos SNPs, en concreto el rs2285666, han sido estudiados por diferentes grupos científicos, concluyendo también que favorecen la capacidad de infección del virus así como el desarrollo de una enfermedad grave por esta causa debido a la modificación de la respuesta inflamatoria (66,93,94).

En cuanto al metabolismo de la vitamina D, su papel en la fisiopatología y en el desarrollo de la COVID-19 es esencial. Además de las funciones de esta vitamina a nivel del metabolismo óseo, presenta un papel importante en la inmunidad innata y adaptativa, así como en la regulación del proceso inflamatorio (95,96). En cuanto a su papel en la inmunidad innata, esta vitamina favorece la expresión de receptores tipo Toll que detectan proteínas patógenas; además, activa sustancias como la catelicidina y las beta-defensinas, y células como los macrófagos y neutrófilos, que dañan la membrana del virus limitando su replicación. Por otro lado, la función de la vitamina D en la inmunidad adquirida produce un descenso de linfocitos T, favoreciendo también su cambio de Th1 a Th2, disminuyendo los niveles de citoquinas proinflamatorias (IL-12,IL-6) a la vez que se aumentan los niveles de citoquinas antiinflamatorias (IL-4,IL-10). También influye en el aumento de linfocitos T reguladores que presentan una acción antiinflamatoria. Se ha descrito cómo el metabolismo de la vitamina D está condicionado por la variabilidad genética, sobre todo por variaciones en los genes *CYP2R1*, *CYP24R*, *DHCR7*, *VDR* y *VDBP* (68,97). En nuestro estudio se demostró que diversos polimorfismos en los genes *VDR* y *VDBP* presentaban un papel importante en el riesgo de neumonía por COVID-19, mortalidad y reingreso durante el primer año tras la infección. Estos polimorfismos son rs1544410, rs7975232 y rs2228570 en el gen *VDR* y rs7041, rs2282679 y rs4588 en el gen *VDBP*, que ya han sido estudiados en otros trabajos mostrando resultados que apoyan su relación con un mayor riesgo de severidad y gravedad de la COVID-19, reforzando la

hipótesis del papel fundamental que juega la vitamina D en la fisiopatología de esta enfermedad (67, 68, 98).

La infección por SARS-CoV-2 desencadena una respuesta hiperinflamatoria conocida como tormenta de citoquinas en la que se produce un aumento de interleucinas como la IL-2 y la IL-7, el interferón- γ , el TNF- α y el factor estimulante de colonias de granulocitos. Esta situación inflamatoria atrae a la zona infectada linfocitos T CD4⁺ y CD8⁺ y macrófagos, favoreciendo la situación inflamatoria, el daño pulmonar y una disfunción multiorgánica (99). Ante este escenario, diversos estudios ya han relacionado polimorfismos en genes implicados en procesos inflamatorios con la severidad de la enfermedad (63-65); sobre todo en genes como *IL6*, *IL6R*, *PCR*, *INF- γ* , *IL1A*, *TNF- α* e *IL1R*. Es conocido que durante el desarrollo de la enfermedad existe un aumento de las concentraciones séricas de PCR y se relaciona con un peor pronóstico (100), pero pocos estudios han demostrado el impacto de sus polimorfismos en la gravedad de la enfermedad. Nuestro estudio demuestra, gracias al aprendizaje automático, que los SNPs en el gen *PCR*, como por ejemplo rs2794521 y rs1800947, tienen una importancia en el pronóstico de estos pacientes, particularmente en la severidad y mortalidad por COVID-19 y en el riesgo de reingreso. Otros estudios han observado que variantes en el gen *IL6R*, como el SNP rs2228145, están relacionadas con un aumento de los niveles de IL-6 en estos pacientes y una mayor mortalidad (101); existe, a su vez, evidencia en otras investigaciones de que SNPs en el gen *IL6*, como rs1800796, se relacionan con una mayor severidad de la enfermedad (102,103). Los resultados en nuestro estudio son superponibles a estos hallazgos, ya que tanto los SNPs del gen *IL6R* como los del gen *IL6* se han visto relacionados en nuestra población con el riesgo de desarrollo de neumonía, mortalidad y riesgo de reingreso y mortalidad durante ese reingreso. Por otro lado, SNPs en genes como *IFN- γ* y *TNF- α* , tales como rs2430561 y rs1800629, respectivamente, se relacionan según investigaciones con mayor susceptibilidad a adquirir una infección por COVID-19, así como mayor susceptibilidad de desarrollo de síntomas (104,105). En nuestro estudio hemos relacionado esos polimorfismos sobre todo al riesgo de desarrollo de neumonía por COVID-19 y de mortalidad asociada.

Nuestro estudio se apoya en la utilización del aprendizaje automático para establecer algoritmos predictivos que se basen en las condiciones genéticas de los pacientes. La IA y el aprendizaje automático, que han experimentado una evolución exponencial en el

campo de la genética, inicialmente se centraban en modelos de clasificación para detectar ciertos patrones genómicos. Sin embargo, de forma progresiva, gracias al desarrollo del aprendizaje profundo y al mayor manejo de volúmenes de datos, se ha permitido ampliar el conocimiento de relaciones genéticas y fenotipos clínicos (106). Estos métodos han facilitado la identificación de variables de riesgo, la predicción de respuesta al tratamiento y la identificación de biomarcadores, mejorando la interpretación clínica y la personalización terapéutica en diferentes enfermedades y relaciones como la diabetes o el cáncer (107,108). Nuestro trabajo lo pone de manifiesto; establece las bases de cómo el desarrollo de algoritmos de IA basados en la integración de parámetros genéticos nos permite identificar con un alto grado de efectividad a los pacientes que tendrán una peor evolución tras sufrir infección por el SARS-CoV-2. Particularmente, hemos desarrollado unos algoritmos de predicción para evaluar el riesgo de neumonía por COVID-19, de mortalidad por dicha causa y de mortalidad tras el reingreso durante el primer año de seguimiento tras la infección por SARS-CoV-2 con un AUC de 0,86; así como para evaluar el riesgo de reingreso durante el primer año de seguimiento tras la infección por SARS-CoV-2 con un AUC de 0,85. Estos resultados se traducen en que nuestros algoritmos basados en el aprendizaje automático presentan una muy buena capacidad de discriminación para determinar qué pacientes presentan una evolución clínica más grave y tienen mayor riesgo de fallecer por dicha causa, así como de reingresar. Esto en la práctica clínica se podrá traducir en una optimización del tiempo y los recursos de cara a poder identificar a los pacientes que dadas sus características genéticas vayan a presentar una peor evolución y así tener un plan de actuación, tanto terapéutico como preventivo, de cara a modificar esta evolución clínica. En nuestro estudio, el método Extreme Gradient Boosting fue el que alcanzó mejores ratios de predicción en la evolución de los pacientes. Este modelo ha demostrado, frente a los otros valorados en nuestro estudio, mayor precisión y equilibrio a la hora de identificar pacientes de riesgo entre nuestra muestra, minimizando de esta forma resultados de falsos positivos y, a su vez, desarrollando una herramienta útil para este objetivo.

El aprendizaje automático ha demostrado ser una herramienta crucial en la investigación de la COVID-19, determinados estudios han conseguido identificar datos clínicos y de laboratorio para predecir la severidad de la enfermedad (109,110); de forma progresiva se ha ido avanzando en dichos modelos incluyendo datos genéticos y permitiendo identificar variantes genéticas que contribuyen a la susceptibilidad y gravedad de la

enfermedad (111), y que en un futuro en la práctica clínica, nos puedan ayudar a identificar pacientes de mayor riesgo e incluso apoyar en la toma de decisiones terapéuticas y estrategias preventivas, así como en la detección de biomarcadores predictivos.

La principal debilidad de nuestro estudio es el tamaño muestral; con una muestra más numerosa, y de diferentes hospitales, se podrían haber hecho comparaciones más específicas, abordando comparaciones clínicas con grupos de pacientes más homogéneos y, por lo tanto, más consistentes y aplicables en la práctica clínica. Por otra parte, la principal fortaleza de nuestro trabajo es que ponemos de manifiesto cómo la variabilidad genética es crucial en la evolución clínica tras la infección por SARS-CoV-2, en particular la influencia de polimorfismos genéticos del SRAA, del metabolismo de la vitamina D y de la respuesta inflamatoria. Además, por primera vez, se desarrollan algoritmos basados en técnicas de IA que determinan el grado de influencia de cada SNP en el riesgo de neumonía, de mortalidad, de reingreso durante el primer año de seguimiento y de mortalidad durante este reingreso. Dichos algoritmos de predicción presentan una gran eficiencia, con AUC superiores al 85%. Este trabajo sienta las bases para que se puedan validar los algoritmos de predicción que hemos diseñado en otras muestras de pacientes y así establecer de forma más consistente su validez.

CONCLUSIONES

En este estudio se pone de manifiesto la importancia del componente genético en el desarrollo de la COVID-19, en concreto en el riesgo de desarrollo de neumonía, de mortalidad, riesgo de reingreso durante el primer año de seguimiento y de mortalidad durante este reingreso; teniendo especial relevancia los SNPs de los genes relacionados con la ECA2, el sistema inflamatorio y el metabolismo de la vitamina D.

Se ha desarrollado mediante métodos de aprendizaje automático una herramienta predictiva que, basándose en esa información genética, ha permitido identificar pacientes con peor pronóstico para las variables previamente descritas. Nuestros algoritmos predictivos que han integrado diversos SNPs han mostrado unos resultados con un AUC superior a 0,85, lo cual se traduce en un alto valor predictivo de estas herramientas para identificar a estos pacientes.

De esta forma, este estudio desarrolla por primera vez un método de predicción de la severidad de la COVID-19 basado en la genética gracias al aprendizaje automático. Además, esta investigación permite en un futuro incorporar una cohorte más amplia, y por otro lado, ampliar los conocimientos desarrollando estudios funcionales en biología molecular y celular que permitan mejorar los conocimientos y la precisión de este modelo.

BIBLIOGRAFÍA

1. Guan WJ, Ni ZY, Hu Y, Liang WH, Ou CQ, He JX, et al. Clinical characteristics of coronavirus disease 2019 in China. *N Engl J Med*. 2020 Apr 30;382(18):1708–20.
2. Paules CI, Marston HD, Fauci AS. Coronavirus infections—more than just the common cold. *JAMA*. 2020;323(8):707–8. doi:10.1001/jama.2020.0757.
3. Gong Z, Song T, Hu M, Che Q, Guo J, Zhang H, et al. Natural and socio-environmental factors in the transmission of COVID-19: a comprehensive analysis of epidemiology and mechanisms. *BMC Public Health*. 2024 Aug 13;24(1):2196.
4. Meslé MMI, Knock ES, Russell TW, Funk S, Abbott S, et al. Estimated number of lives directly saved by COVID-19 vaccination programmes in the WHO European Region from December 2020 to March 2023: a retrospective surveillance study. *Lancet Respir Med*. 2024 Sep;12(9):714–27. doi:10.1016/S2213-2600(24)00179-6. Erratum in: *Lancet Respir Med*. 2025 Jun 23;S2213-2600(25)00235-8. doi:10.1016/S2213-2600(25)00235-8.
5. Lauer SA, Grantz KH, Bi Q, Jones FK, Zheng Q, Meredith HR, et al. The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: estimation and application. *Ann Intern Med*. 2020;172(9):577–82. doi:10.7326/M20-0504.
6. Rabenau HF, Cinatl J, Morgenstern B, Bauer G, Preiser W, Doerr HW. Stability and inactivation of SARS coronavirus. *Med Microbiol Immunol*. 2005;194(1):1–6.
7. Gerlach M, Wolff S, Ludwig S, Schäfer W, Keiner B, Roth NJ, et al. Rapid SARS-CoV-2 inactivation by commonly available chemicals on inanimate surfaces. *J Hosp Infect*. 2020;106(3):633–5.
8. Parasa S, Desai M, Thoguluva Chandrasekar V, Patel HK, Kennedy KF, Roesch T, et al. Prevalence of gastrointestinal symptoms and fecal viral shedding in patients with coronavirus disease 2019: a systematic review and meta-analysis. *JAMA Netw Open*. 2020;3(6):e2011335. doi:10.1001/jamanetworkopen.2020.11335.
9. Monserrat Sanz J, Gómez Lahoz AM, Oliva Martín R. Papel del sistema inmune en la infección por el SARS-CoV-2: inmunopatología de la COVID-19. *Medicine (B Aires)*. 2021;13(33):1917–31. doi:10.1016/j.med.2021.05.005.

10. Song Z, Xu Y, Bao L, Zhang L, Yu P, Qu Y, et al. From SARS to MERS, thrusting coronaviruses into the spotlight. *Viruses*. 2019;11(1):59. doi:10.3390/v11010059.
11. Guo YR, Cao QD, Hong ZS, Tan YY, Chen SD, Jin HJ, et al. The origin, transmission and clinical therapies on coronavirus disease 2019 (COVID-19) outbreak: an update on the status. *Mil Med Res*. 2020;7(1):11. doi:10.1186/s40779-020-00240-0.
12. Tay MZ, Poh CM, Rénia L, MacAry PA, Ng LFP. The trinity of COVID-19: immunity, inflammation and intervention. *Nat Rev Immunol*. 2020 Jun;20(6):363–74. doi:10.1038/s41577-020-0311-8.
13. Enjuanes L. Development of protection against coronavirus induced diseases: a review. In: Siddell SG, editor. *Corona- and related viruses: current concepts in molecular biology and pathogenesis*. Boston: Springer; 1995. p. 197–211.
14. Siu YL, Teoh KT, Lo J, Chan CM, Kien F, Escriou N, et al. The M, E, and N structural proteins of the severe acute respiratory syndrome coronavirus are required for efficient assembly, trafficking, and release of virus-like particles. *J Virol*. 2008 Nov;82(22):11318–30. doi:10.1128/JVI.01052-08.
15. Masters PS. Genetic and molecular biological analysis of protein–protein interactions in coronavirus assembly. In: Snijder EJ, Spaan WJM, Ziebuhr J, editors. *The Nidoviruses: toward control of SARS and other nidovirus diseases*. Boston: Springer; 2006. p. 163–73.
16. Perlman S. Pathogenesis of coronavirus-induced infections: review of pathological and immunological aspects. In: Siddell SG, editor. *Coronaviruses and arteriviruses*. New York: Springer; 1998. p. 503–13.
17. Maguiña Vargas C, Gastelo Acosta R, Tequen Bernilla A. El nuevo coronavirus y la pandemia del COVID-19. *Rev Med Hered*. 2020 Abr;31(2):125–31. doi:10.20453/rmh.v31i2.3687.
18. Vilema Paillacho L. Influencia de factores funcionales en los aspectos celulares y moleculares del sistema inmunológico. [tesis de licenciatura]. Quito: Universidad Central del Ecuador; 2024.
19. Adamo S, Ladell K, Davies A, Gostick E, Schober K, Bruncker K, et al. Signature of long-lived memory CD8+ T cells in acute SARS-CoV-2 infection. *Nature*. 2022 Feb;602(7895):148–55. doi:10.1038/s41586-021-04330-4.

20. Kusnadi A, Ramírez-Suástegui C, Fajardo V, Chee SJ, Meckiff BJ, Simon H, et al. Severely ill patients with COVID-19 display impaired exhaustion features in SARS-CoV-2-reactive CD8⁺ T cells. *Sci Immunol.* 2021;6(55):eabe4782. doi:10.1126/sciimmunol.abe4782.
21. Iwasaki S, Fujisawa S, Nakakubo S, Kamada K, Yamashita Y, Fukumoto T, et al. Comparison of SARS-CoV-2 detection in nasopharyngeal swab and saliva. *J Infect.* 2020;81(2):e145–7. doi:10.1016/j.jinf.2020.05.071.
22. Lynch KL, Whitman JD, Lacanienta NP, Beckerdite EW, Kastner SA, Shy BR, et al. Magnitude and kinetics of anti-severe acute respiratory syndrome coronavirus 2 antibody responses and their relationship to disease severity. *Clin Infect Dis.* 2021;72(2):301–8. doi:10.1093/cid/ciaa979.
23. Choe PG, Kang CK, Suh HJ, Jung J, Song KH, Bang JH, et al. Waning antibody responses in asymptomatic and symptomatic SARS-CoV-2 infection. *Emerg Infect Dis.* 2021;27(1):327–9. doi:10.3201/eid2701.203515.
24. Liu T, Hu J, Xiao J, He G, Kang M, Rong Z, et al. Time-varying transmission dynamics of novel coronavirus pneumonia in China. *bioRxiv [Preprint].* 2020 Jan 25. doi:10.1101/2020.01.25.919787.
25. Alimohamadi Y, Sepandi M, Taghdir M, Hosamirudsari H. Determine the most common clinical symptoms in COVID-19 patients: a systematic review and meta-analysis. *J Prev Med Hyg.* 2020;61(3):E304–12. doi:10.15167/2421-4248/jpmh2020.61.3.1530.
26. Oran DP, Topol EJ. The proportion of SARS-CoV-2 infections that are asymptomatic: a systematic review. *Ann Intern Med.* 2021;174(5):655–62. doi:10.7326/M20-6976.
27. Omoush SA, Alzyoud J. The prevalence and impact of coinfection and superinfection on the severity and outcome of COVID-19 infection: an updated literature review. *Pathogens.* 2022;11(4):445. doi:10.3390/pathogens11040445.
28. Marr KA, Platt A, Tornheim JA, Zhang SX, Datta K, Cardozo C, et al. Aspergillosis complicating severe coronavirus disease. *Emerg Infect Dis.* 2021;27(1):18–25. doi:10.3201/eid2701.202896.
29. Rutsaert L, Steinfert N, Van Hunsel T, Bomans P, Naesens R, Mertes H, et al. COVID-19-associated invasive pulmonary aspergillosis. *Ann Intensive Care.* 2020;10(1):71. doi:10.1186/s13613-020-00686-4.

30. Wu J, Wu X, Zeng W, Guo D, Fang Z, Chen L, et al. Chest CT findings in patients with coronavirus disease 2019 and its relationship with clinical features. *Investig Radiol.* 2020;55(5):257–61. doi:10.1097/RLI.0000000000000670.
31. Jenner WJ, Kanji R, Mirsadraee S, Gue YX, Price S, Prasad S, et al. Thrombotic complications in 2928 patients with COVID-19 treated in intensive care: a systematic review. *J Thromb Thrombolysis.* 2021;51:595–607. doi:10.1007/s11239-020-02380-7.
32. Klok FA, Kruip MJHA, van der Meer NJM, Arbous MS, Gommers DAMPJ, Kant KM, et al. Confirmation of the high cumulative incidence of thrombotic complications in critically ill ICU patients with COVID-19: an updated analysis. *Thromb Res.* 2020;191:148–50. doi:10.1016/j.thromres.2020.04.041.
33. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet.* 2020;395(10223):497–506. doi:10.1016/S0140-6736(20)30183-5.
34. Xiong S, Liu L, Lin F, Shi J, Han C, Wang B, et al. Clinical characteristics of 116 hospitalized patients with COVID-19 in Wuhan, China: a single-centered, retrospective, observational study. *BMC Infect Dis.* 2020;20(1):787. doi:10.1186/s12879-020-05495-4.
35. Portela-Sánchez S, Sánchez-Porras R, Cuadrado ML, Alguacil-Guillén M, Sánchez-Sánchez R, Marín-Corral J, et al. Neurological complications of COVID-19 in hospitalized patients: the registry of a neurology department in the first wave of the pandemic. *Eur J Neurol.* 2021;28(10):3339–47. doi:10.1111/ene.14983.
36. Ge E, Li X, Liu J, Yang Y. Association of pre-existing comorbidities with mortality and disease severity among 167,500 individuals with COVID-19 in Canada: a population-based cohort study. *PLoS One.* 2021;16(10):e0258154. doi:10.1371/journal.pone.0258154.
37. Hendren NS, de Lemos JA, Ayers C, Das SR, Rao A, Carter S, et al. Association of body mass index and age with morbidity and mortality in patients hospitalized with COVID-19: results from the American Heart Association COVID-19 Cardiovascular Disease Registry. *Circulation.* 2021;143(2):135–44. doi:10.1161/CIRCULATIONAHA.120.051936.

38. Pezzullo AM, Axfors C, Contopoulos-Ioannidis DG, Apostolatos A, Ioannidis JPA. Age-stratified infection fatality rate of COVID-19 in the non-elderly population. *Environ Res.* 2023;216:114655. doi:10.1016/j.envres.2022.114655.
39. Jose RJ, Manuel A. COVID-19 cytokine storm: the interplay between inflammation and coagulation. *Lancet Respir Med.* 2020;8(6):e46–7. doi:10.1016/S2213-2600(20)30216-2.
40. Camps-Vilaró A, Pinsach-Abuin M, Degano IR, Ramos R, Martí-Lluch R, Elosua R, et al. Genetic characteristics involved in COVID-19 severity: the CARGENCORS case-control study and meta-analysis. *J Med Virol.* 2024;96(2):e29404. doi:10.1002/jmv.29404.
41. Velavan TP, Pallerla SR, Rüter J, Augustin Y, Kremsner PG, Krishna S, et al. Host genetic factors determining COVID-19 susceptibility and severity. *EBioMedicine.* 2021;72:103629. doi:10.1016/j.ebiom.2021.103629.
42. Niemi MEK, Daly MJ, Ganna A. The human genetic epidemiology of COVID-19. *Nat Rev Genet.* 2022;23(9):533–46. doi:10.1038/s41576-022-00469-0.
43. Garcia-Carretero R, Vazquez-Gomez O, Gil-Prieto R, Gil-de-Miguel A. Hospitalization burden and epidemiology of the COVID-19 pandemic in Spain (2020–2021). *BMC Infect Dis.* 2023;23(1):476. doi:10.1186/s12879-023-08454-y.
44. Fu L, Wang B, Yuan T, Chen X, Ao Y, Fitzpatrick T, et al. Clinical characteristics of coronavirus disease 2019 (COVID-19) in China: a systematic review and meta-analysis. *J Infect.* 2020 Jun;80(6):656–65. doi:10.1016/j.jinf.2020.03.041.
45. Ingraham NE, Barakat AG, Reilkoff R, Bezdicek T, Schacker T, Chipman JG, et al. Understanding the renin-angiotensin-aldosterone-SARS-CoV axis: a comprehensive review. *Eur Respir J.* 2020 Jul;56(1):2000912. doi:10.1183/13993003.00912-2020.
46. Yan R, Zhang Y, Li Y, Xia L, Guo Y, Zhou Q. Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. *Science.* 2020 Mar 27;367(6485):1444–8. doi:10.1126/science.abb2762.

47. Shang J, Ye G, Shi K, Wan Y, Luo C, Aihara H, et al. Structural basis of receptor recognition by SARS-CoV-2. *Nature*. 2020 May;581(7807):221–4. doi:10.1038/s41586-020-2179-y.
48. Pastrian-Soto G. Presencia y expresión del receptor ACE2 (target de SARS-CoV-2) en tejidos humanos y cavidad oral. Posibles rutas de infección en órganos orales. *Int J Odontostomat*. 2020 Dec;14(4):501-507.
49. Imai Y, Kuba K, Rao S, Huan Y, Guo F, Guan B, et al. Angiotensin-converting enzyme 2 protects from severe acute lung failure. *Nature*. 2005 Jul;436(7047):112–6. doi:10.1038/nature03712.
50. Annoni F, Orbegozo D, Rahmania L, Irazabal M, Mendoza M, De Backer D, et al. Angiotensin-converting enzymes in acute respiratory distress syndrome. *Intensive Care Med*. 2019 Aug 1;45(8):1159–60. doi:10.1007/s00134-019-05626-2.
51. Kuba K, Imai Y, Rao S, Gao H, Guo F, Guan B, et al. A crucial role of angiotensin converting enzyme 2 (ACE2) in SARS coronavirus–induced lung injury. *Nat Med*. 2005 Aug;11(8):875–9. doi:10.1038/nm1267.
52. Mohan M, Cherian JJ, Sharma A. Exploring links between vitamin D deficiency and COVID-19. *PLoS Pathog*. 2020 Sep 18;16(9):e1008874. doi:10.1371/journal.ppat.1008874.
53. Weir EK, Thenappan T, Bhargava M, Chen Y. Does vitamin D deficiency increase the severity of COVID-19? *Clin Med (Lond)*. 2020 Jul;20(4):e107–8. doi:10.7861/clinmed.2020-0280.
54. Minton K. Vitamin D shuts down T cell-mediated inflammation. *Nat Rev Immunol*. 2022 Jan;22(1):1. doi:10.1038/s41577-021-00671-x.
55. Colotta F, Jansson B, Bonelli F. Modulation of inflammatory and immune responses by vitamin D. *J Autoimmun*. 2017 Dec;85:78–97. doi:10.1016/j.jaut.2017.07.008.
56. Zdrenghea MT, Makrinioti H, Bagacean C, Bush A, Johnston SL, Stanciu LA. Vitamin D modulation of innate immune responses to respiratory viral infections. *Rev Med Virol*. 2017 Jan;27(1):e1909. doi:10.1002/rmv.1909.
57. Radujkovic A, Hippchen T, Tiwari-Heckler S, Dreher S, Boxberger M, Merle U. Vitamin D deficiency and outcome of COVID-19 patients. *Nutrients*. 2020 Sep 10;12(9):2757. doi:10.3390/nu12092757.

58. Bryce C, Grimes Z, Pujadas E, Ahuja S, Beasley MB, Albrecht R, et al. Pathophysiology of SARS-CoV-2: targeting of endothelial cells renders a complex disease with thrombotic microangiopathy and aberrant immune response. The Mount Sinai COVID-19 autopsy experience. medRxiv. 2020. doi:10.1101/2020.05.18.20099960.
59. Gustine JN, Jones D. Immunopathology of hyperinflammation in COVID-19. *Am J Pathol*. 2021 Jan 1;191(1):4–17. doi:10.1016/j.ajpath.2020.08.009.
60. Tan LY, Komarasamy TV, Balasubramaniam RM. Hyperinflammatory immune response and COVID-19: a double edged sword. *Front Immunol*. 2021 Sep 30;12:742941. doi:10.3389/fimmu.2021.742941.
61. Mehta P, McAuley DF, Brown M, Sanchez E, Tattersall RS, Manson JJ, et al. COVID-19: consider cytokine storm syndromes and immunosuppression. *Lancet*. 2020 Mar 28;395(10229):1033–4. doi:10.1016/S0140-6736(20)30628-0.
62. Jose RJ, Manuel A. COVID-19 cytokine storm: the interplay between inflammation and coagulation. *Lancet Respir Med*. 2020 Jun;8(6):e46–7. doi:10.1016/S2213-2600(20)30216-2.
63. Yip JQ, Oo A, Ng YL, Chin KL, Tan KK, Chu JJH, AbuBakar S, Zainal N. The role of inflammatory gene polymorphisms in severe COVID-19: a review. *Virol J*. 2024 Dec 20;21(1):327. doi:10.1186/s12985-024-02597-3.
64. Ren H, Lin Y, Huang L, Xu W, Luo D, Zhang C. Association of genetic polymorphisms with COVID-19 infection and outcomes: an updated meta-analysis based on 62 studies. *Heliyon*. 2023 Dec 14;10(1):e23662. doi:10.1016/j.heliyon.2023.e23662.
65. Abobaker A, Nagib T, Alsoufi A. The impact of certain genetic variants (single nucleotide polymorphisms) on incidence and severity of COVID-19. *J Gene Med*. 2021 Feb;23(2):e3310. doi:10.1002/jgm.3310.
66. Sabater Molina M, Nicolás Rocamora E, Bendicho AI, Vázquez EG, Zorio E, Rodriguez FD, et al. Polymorphisms in ACE, ACE2, AGTR1 genes and severity of COVID-19 disease. *PLoS One*. 2022;17(1):e0263140. doi:10.1371/journal.pone.0263140.
67. Dobrijevic Z, Robajac D, Gligorijevic N, Šunderic M, Penezic A, Miljuš G, et al. The association of ACE1, ACE2, TMPRSS2, IFITM3 and VDR polymorphisms with COVID-19 severity: A systematic review and meta-

- analysis. *EXCLI J.* 2022 Jun 20;21:818-839. doi: 10.17179/excli2022-4976. PMID: 35949487.
68. Tentolouris N, Achilla C, Anastasiou IA, Eleftheriadou I, Tentolouris A, Basoulis D, et al. The Association of Vitamin D Receptor Polymorphisms with COVID-19 Severity. *Nutrients.* 2024 Mar 2;16(5):727. doi: 10.3390/nu16050727. PMID: 38474855.
 69. McCarthy J, Minsky ML, Rochester N, Shannon CE. A proposal for the Dartmouth summer research project on artificial intelligence. *AI Mag.* 2006;27:12–14.
 70. Pople HE. The formation of composite hypotheses in diagnostic problem-solving: an exercise in synthetic reasoning. *Proceedings of the Third International Joint Conference on Artificial Intelligence.* 1973:1030–1037.
 71. Miller RA, Pople HE, Myers JD. INTERNIST-I, an experimental computer-based diagnostic consultant for general internal medicine. *N Engl J Med.* 1982 Aug 19;307(8):468–76.
 72. Shortliffe EH. MYCIN: A rule-based computer program for advising physicians regarding antimicrobial therapy selection. PhD thesis. Stanford University; 1975.
 73. Heckerman D. Probabilistic interpretation of MYCIN's certainty factors. *Uncertainty in Artificial Intelligence.* 1990:167–196.
 74. Pearl J. Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann; 1988.
 75. Kononenko I. Machine learning for medical diagnosis: history, state of the art and perspective. *Artif Intell Med.* 2001 Jan;23(1):89–109.
 76. Zadeh LA. Fuzzy logic. *Computer.* 1988 Apr;21(4):83–93.
 77. Bellazzi R, Zupan B. Predictive data mining in clinical medicine: a focus on selected methods and applications. *Artif Intell Med.* 2008 Apr;34(2):79–91.
 78. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med.* 2019 Apr 4;380(14):1347–58.
 79. Litmanovich DE, Chung M, Kirkbride RR, Kicska G, Kanne JP. Review of chest radiograph findings of COVID-19 pneumonia and suggested reporting language. *J Thorac Imaging.* 2020 Sep;35(5):354–60.
 80. Schleinitz D, Distefano JK, Kovacs P. Targeted SNP genotyping using the TaqMan® assay. *Methods Mol Biol.* 2011;700:77–87.

81. Cervantes J, Garcia-Lamont F, Rodríguez-Mazahua L, Lopez A. A comprehensive survey on support vector machine classification: applications, challenges and trends. *Neurocomputing*. 2020 Sep 30;408:189–215.
82. Charbuty B, Abdulazeez A. Classification based on decision tree algorithm for machine learning. *J Appl Sci Technol Trends*. 2021 Mar 24;2(1):20–8.
83. Ampomah EK, Nyame G, Qin Z, Addo PC, Gyamfi EO, Gyan M. Stock market prediction with Gaussian Naïve Bayes machine learning algorithm. *Informatica*. 2021 Jun 15;45(2).
84. Cubillos M, Wøhlk S, Wulff JN. A bi-objective k-nearest-neighbors-based imputation method for multilevel data. 2022 [cited 2025 Apr 23].
85. Chen T, Benesty M, Khotilovich V, Tang Y, Cho H, Chen K, et al. XGBoost: extreme gradient boosting. R package version 0.4-2. 2015;1(4):1–4.
86. Feurer M, Hutter F. Hyperparameter optimization. In: Hutter F, Kotthoff L, Vanschoren J, editors. *Automated machine learning: methods, systems, challenges*. Cham: Springer International Publishing; 2019. p. 3–33.
87. Chen RC, Dewi C, Huang SW, Caraka RE. Selecting critical features for data classification based on machine learning methods. *J Big Data*. 2020 Jul 23;7(1):52.
88. Herland M, Khoshgoftaar TM, Wald R. A review of data mining using big data in health informatics. *J Big Data*. 2014 Jun 24;1(1):2.
89. MATLAB Runtime - MATLAB Compiler [Internet]. Available from: <https://es.mathworks.com/products/compiler/matlab-runtime.html>
90. Subong BJJ, Forteza IL. SARS CoV 2 replication revisited: molecular insights and current and emerging antiviral strategies. *COVID*. 2025;5(6):85. doi:10.3390/covid5060085.
91. Dong Y, Shamsuddin A, Campbell H, Theodoratou E. Current COVID-19 treatments: rapid review of the literature. *J Glob Health*. 2021 Apr 24;11:10003. doi:10.7189/jogh.11.10003.
92. Torrente-López A, Hermosilla J, Navas N, Cuadros-Rodríguez L, Cabeza J, Salmerón-García A. The relevance of monoclonal antibodies in the treatment of COVID-19. *Vaccines (Basel)*. 2021 May 26;9(6):557. doi:10.3390/vaccines9060557.

93. Gómez J, Albaiceta GM, García-Clemente M, López-Larrea C, Amado-Rodríguez L, Lopez-Alonso I, et al. Angiotensin-converting enzymes (ACE, ACE2) gene variants and COVID-19 outcome. *Gene*. 2020;762:145102.
94. Saengsiwaritt W, Jittikoon J, Chaikledkaew U, Udomsinprasert W. Genetic polymorphisms of ACE1, ACE2, and TMPRSS2 associated with COVID-19 severity: a systematic review with meta-analysis. *Rev Med Virol*. 2022;32(1):e2323.
95. Bischoff-Ferrari HA, Dawson-Hughes B, Stöcklin E, Sidelnikov E, Willett WC, Edel JO, et al. Oral supplementation with 25(OH)D3 versus vitamin D3: effects on 25(OH)D levels, lower extremity function, blood pressure, and markers of innate immunity. *J Bone Miner Res*. 2012;27(1):160-169.
96. Malmberg HR, Hanel A, Taipale M, Heikkinen S, Carlberg C. Vitamin D treatment sequence is critical for transcriptome modulation of immune challenged primary human cells. *Front Immunol*. 2021;12:754056.
97. Saccone D, Asani F, Bornman L. Regulation of the vitamin D receptor gene by environment, genetics and epigenetics. *Gene*. 2015;561(2):171-180.
98. Karcioğlu Batur L, Dokur M, Koç S, Karabay M, Akcay ZN, Gunger E, et al. Investigation of the relationship between vitamin D deficiency and vitamin D-binding protein polymorphisms in severe COVID-19 patients. *Diagnostics*. 2024;14(9):1941.
99. Merad M, Blish CA, Sallusto F, Iwasaki A. The immunology and immunopathology of COVID-19. *Science*. 2022;375(6585):1122-1127.
100. Vogt V, Haschka D, Forer L, Schwendinger S, Petzer V, Coassin S, et al. Severe COVID-19 disease is associated with genetic factors affecting plasma ACE2 receptor and CRP concentrations. *Sci Rep*. 2025;15:4708.
101. Smieszek SP, Przychodzen BP, Polymeropoulos VM, Polymeropoulos CM, Polymeropoulos MH. Assessing the potential correlation of polymorphisms in the IL6R with relative IL6 elevation in severely ill COVID-19 patients. *Cytokine*. 2021 Dec;148:155662. doi: 10.1016/j.cyto.2021.155662. Epub 2021 Jul 29. Erratum in: *Cytokine*. 2022 Jan;149:155752. doi: 10.1016/j.cyto.2021.155752. PMID: 34353696.
102. Torres-Poveda K, Bahena-Román M, Contreras-Ochoa CO, Lagunas-Martínez A, Bermúdez-Morales VH, Pando-Robles V, et al. High nasopharyngeal and serum IL-6 levels and the -573G>C polymorphism

- (rs1800796) are linked with the risk of severe COVID-19 in a Mexican population: a case-control study. *BMC Infect Dis.* 2025 Mar 5;25(1):315. doi: 10.1186/s12879-025-10695-y. PMID: 40045221.
103. Cekin N, Akin S, Pinarbasi E, Doğan OH. Impact of IL-6 rs1800795 and rs1800796 polymorphisms on clinical outcomes of COVID-19: a study on severity of disease in Turkish population. *Mamm Genome.* 2025 Mar;36(1):213-229. doi: 10.1007/s00335-024-10085-w. Epub 2024 Nov 20. PMID: 39567384.
 104. Sarges KML, Póvoa da Costa F, Santos EFD, Cantanhede MHD, da Silva R, Veríssimo AOL, et al. Association of the IFNG +874T/A polymorphism with symptomatic COVID-19 susceptibility. *Viruses.* 2024 Apr 22;16(4):650. doi: 10.3390/v16040650. PMID: 38675991.
 105. Alsayed BA, Mir R, Mir MM, Alnour TMS, Fawzy S, Ahmed MM, et al. Molecular determination of tumor necrosis factor-alpha, interleukin-8, interleukin-10, and C-X-C chemokine receptor-2 genetic variations and their association with disease susceptibility and mortality in COVID-19 patients. *Curr Genomics.* 2024 Feb 23;25(1):12-25. doi: 10.2174/0113892029272497240103052359. PMID: 38544825.
 106. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. *Nat Rev Genet.* 2015 Jun;16(6):321-32.
 107. Bellot P, et al. Machine learning approaches for the prediction of complex diseases using genetic data. *Hum Genet.* 2020 Feb;139(2):237-50.
 108. Esteva A, et al. Deep learning-enabled medical computer vision. *Nature.* 2019 Dec;572(7767):485-90.
 109. Kang J, Chen T, Luo H, Luo Y, Du G, Jiming-Yang M. Machine learning predictive model for severe COVID-19. *Infect Genet Evol.* 2021 Jun;90:104737. doi: 10.1016/j.meegid.2021.104737. Epub 2021 Jan 28. Erratum in: *Infect Genet Evol.* 2022 Sep;103:105330. doi: 10.1016/j.meegid.2022.105330. Erratum in: *Infect Genet Evol.* 2022 Jul 25:105335. doi: 10.1016/j.meegid.2022.105335. PMID: 33515712.
 110. Kamel FO, Magadmi R, Qutub S, Badawi M, Badawi M, Madani TA, et al. Machine learning-based prediction of COVID-19 prognosis using clinical and hematologic data. *Cureus.* 2023 Dec 9;15(12):e50212. doi: 10.7759/cureus.50212. PMID: 38089943.

111. Sethi S, Shakyawar S, Reddy AS, Patel JC, Guda C. A machine learning model for the prediction of COVID-19 severity using RNA-Seq, clinical, and co-morbidity data. *Diagnostics (Basel)*. 2024 Jun 18;14(12):1284. doi: 10.3390/diagnostics14121284. PMID: 38928699.

ANEXOS

El presente trabajo para optar al título de Doctora por la Universidad de Valladolid en el programa de Doctorado en Investigación en Ciencias de la Salud, ha sido publicado como:

Jaurrieta-Largo S, Miramontes-González JP, Corral-Gudino L, Gabella-Martín M, Pérez-Arroyo S, Torres AM, et al. A machine learning approach to understanding the genetic role in COVID-19 prognosis: the influence of gene polymorphisms related to inflammation, vitamin D, and ACE2. *Int J Mol Sci*. 2025 Aug 18;26(16):7975. doi: 10.3390/ijms26167975. PMID: 40869297.

International Journal of Molecular Sciences. Factor de impacto 4,9.

