UNIVERSIDAD AUTÓNOMA DE BAJA CALIFORNIA

Facultad de Ingeniería, Arquitectura y Diseño

Maestría y Doctorado en Ciencias e Ingeniería

Identificación y predicción de farmacorresistencia en genoma de *Mycobacterium tuberculosis* utilizando métodos de *ML*

TESIS

PARA OBTENER EL GRADO DE

DOCTOR EN CIENCIAS

Presenta

RICARDO PEREA JACOBO

ENSENADA, BAJA CALIFORNIA, MÉXICO 2024

UNIVERSIDAD AUTÓNOMA DE BAJA CALIFORNIA

FACULTAD DE INGENIERÍA, ARQUITECTURA Y DISEÑO

MAESTRÍA Y DOCTORADO EN CIENCIAS E INGENIERÍA

Identificación y predicción de farmacorresistencia en genoma de *Mycobacterium tuberculosis* utilizando métodos de ML

TESIS

Que para obtener el grado de Doctorado en Ciencias presenta:

Ricardo Perea Jacobo

Aprobada por:

Dra. Dora-Luz Flores Gutiérrez
Director de tesis

Dra. Raquel Muñiz Salazar
Codirectora /Miembro del comité

Dr. Roberto Zenteno Cuevas
Miembro del comité

Dra. Dayanira Sheira Paniagua Meza
Miembro del comité

Dr. Dante Alberto Magdaleno Moncayo
Miembro del comité

Ensenada Baja California, México. Noviembre 2024

Índice

Resumen.

La tuberculosis (TB) es una enfermedad infecciosa que sigue representando un reto considerable para la salud pública global, debido a la complejidad en su diagnóstico temprano y en el manejo adecuado del seguimiento y tratamiento de los pacientes. La rápida identificación de cepas de *Mycobacterium tuberculosis* que sean resistentes o susceptibles a determinados fármacos es esencial para garantizar un tratamiento efectivo, minimizar las complicaciones y reducir de manera significativa la duración del tratamiento. En este contexto, la capacidad de diagnosticar la resistencia a los fármacos de manera precisa y oportuna resulta crucial para la implementación de terapias adecuadas y, en última instancia, para disminuir la mortalidad asociada con esta enfermedad.

El objetivo principal de este proyecto fue desarrollar y evaluar modelos de *machine learning* (aprendizaje máquina, ML) para la predicción de resistencia a fármacos antituberculosis de primera y segunda línea, utilizando el análisis del genoma completo de aislados clínicos de *Mycobacterium tuberculosis*. En este trabajo se exploraron diferentes enfoques de *ML*, incluyendo redes neuronales convolucionales (CNN) y tres métodos tradicionales: bosques aleatorios (RF), máquinas de soporte vectorial (SVM) y regresión logística (RL). La representación de la información genómica en un formato compatible con estos modelos resultó ser altamente efectiva, con una precisión y un *recall* superiores al 90% objetivo.

Los resultados mostraron que el mejor rendimiento global fue alcanzado por los modelos basados en bosques aleatorios (RF), aunque los modelos de redes neuronales convolucionales también obtuvieron resultados prometedores. Sin embargo, para optimizar el rendimiento de las CNN, ¿se considera necesario explorar más a fondo la forma en que los datos biológicos son presentados, como la información de la posición genómica de las variantes, la distancia o correlación entre ellas, así como orden en que la información es introducida a los modelos. Este proyecto pretende contribuir de manera significativa al avance en el diagnóstico y tratamiento de la tuberculosis, utilizando herramientas de inteligencia artificial que puedan mejorar la precisión y la eficiencia en la predicción de resistencia a fármacos, y así, ayudar a resolver uno de los mayores desafíos en el control de esta enfermedad a nivel mundial.

Introducción

La TB es una enfermedad infecciosa causada por el complejo *Mycobacterium tuberculosis*, bacteria que afecta principalmente a los pulmones (TB pulmonar), pero también puede afectar otros órganos (TB extrapulmonar). La enfermedad se transmite de una persona a otra a través del aire mediante la tos o estornudos. La TB constituye una de las diez principales causas de muerte a nivel mundial, y la segunda causa de muerte por enfermedades infecciosas.

La TB es curable si se diagnostica a tiempo, sin embargo, las personas con TB pueden morir si no reciben el tratamiento oportuno y especifico. El diagnóstico oportuno de la TB sigue siendo un desafío significativo a nivel global, con una preocupante brecha en la detección de casos. En 2022, la OMS estimó que de los 10.6 millones de personas que desarrollaron TB, cerca de 4.2 millones no fueron diagnosticadas ni notificadas, lo que representa una "brecha del diagnóstico" del 40%. Esta brecha es aún mayor en regiones de bajos y medianos ingresos, donde la falta de acceso a servicios de salud, la escasez de personal capacitado y la limitada disponibilidad de pruebas diagnósticas eficaces, complican la detección temprana. Esta situación no solo incrementa el fracaso del tratamiento, lo que aumenta la mortalidad y la transmisión de la enfermedad, sino que también subestima la verdadera carga de la TB, dificultando los esfuerzos globales para controlar y eliminar la enfermedad (World Health Organization, 2023).

El fracaso de tratamiento es uno de los factores asociados al desarrollo de la TB farmacorresistente, lo que representa uno de los mayores obstáculos en la lucha contra esta enfermedad, exacerbando tanto su control como su tratamiento. La OMS reporta un aumento de la incidencia de la TB resistente a múltiples fármacos (MDR-TB), esta es una cepa que no responde a los dos principales medicamentos antituberculosis (rifampicina e isoniacida), y la TB extremadamente resistente a los medicamentos (XDR-TB), que es una cepa de *Mycobacterium tuberculosis* que cumplen la definición de MDR-TB y que también son resistentes a cualquier fluoroquinolona y al menos a un fármaco adicional del Grupo A (World Health Organization, 2020). Según la OMS, en 2022, se registraron aproximadamente 450,000 nuevos casos de MDR-TB a nivel mundial, pero solo una fracción de estos fueron tratados adecuadamente. La detección de los casos de TB-MDR y TB-XDR es compleja y costosa, lo que agrava la situación, especialmente en países con recursos limitados.

A nivel mundial, solo el 64% de los casos de TB son diagnosticados, es decir, de los 10 millones de nuevos casos, 3.6 millones de personas se encuentran sin tratamiento y consecuentemente contagiando a más personas. Muchos países, incluyendo México, dependen todavía de la baciloscopía para diagnosticar TB, prueba que viene utilizándose desde hace más de 100 años (World Health Organization, 2018a). La detección de TB-FR requiere de pruebas adicionales rápidas y precisas que permitan determinar el tratamiento farmacológico de acuerdo con el perfil de farmacorresistencia de la bacteria, y con ello disminuir la dispersión del patógeno en la comunidad. Por otro lado, tenemos el factor económico, debido a que el 50 % de los pacientes

de TB, y sus familias, enfrentan un efecto catastrófico sobre sus finanzas para resolver esta enfermedad, superando el 20% del ingreso familiar anual(World Health Organization, 2023).

La identificación rápida de cepas resistentes o susceptibles a ciertos fármacos es esencial para el tratamiento adecuado, evitando así complicaciones y reduciendo significativamente la duración del tratamiento. Las nuevas tecnologías de inteligencia artificia o *ML* y *deep learning (DL)* brindan una posibilidad de abordaje para el análisis de un importante número de genomas completos de *M. Tuberculosis* depositadas en diversas bases de datos, estas plataformas de análisis están específicamente diseñadas para manejar grandes cantidades de información y a partir esto generar predicciones que permiten una detección rápida de mutaciones asociadas con resistencia a fármacos y con ello apoyar en la toma de decisiones clínicas y contribuir al diagnóstico rápido y oportuno de esta resistencia.

Los métodos de *ML* se han aplicado a una gran variedad de problemas en genómica y genética. El aprendizaje automático se ha utilizado para generar una anotar una amplia variedad de anotaciones derivadas de los elementos presentes en las secuencias genómicas.

Los algoritmos pueden ser entrenados para identifica elementos de secuencia de un tipo dado, y con ello entrenar un método de aprendizaje automático para reconocer esos elementos. Además, los modelos que reconocen un tipo individual de elementos genómicos se pueden combinar mediante una técnica conocida como "*ensamble learning*". Esta permite que la lógica aprendida sobre sus ubicaciones relativas de las variantes puedan ser usadas para construir sistemas de aprendizaje automático capaces de anotar genes, incluida su estructura completa de UTR/ intrón/ exón, a lo largo del genoma (Libbrecht & Noble, 2015).

grandes volúmenes de datos genómicos permitens neuronales en el desarrollo de herramientas de apoyo al diagnóstico de resistencia en TB, ya que su capacidad para procesar y analizar grandes volúmenes de datos genómicos permite identificar patrones complejos que podrían no ser detectables mediante métodos tradicionales. Estas herramientas avanzadas tendrían el potencial de mejorar significativamente la precisión y rapidez en la detección de enfermedades, como la tuberculosis farmacorresistente, facilitando un diagnóstico más temprano y una toma de decisiones clínicas más informada. La investigación en este campo podría abrir nuevas vías para la implementación de soluciones más eficaces en la práctica clínica diaria. El objetivo de este proyecto es explorar estas aproximaciones en la construcción de modelos de predicción de farmacorresistencia en datos genómicos.

Justificación

La OMS ha declarado que la farmacorresistencia a los antimicrobianos es una de las 10 principales amenazas de salud pública a las que se enfrenta la humanidad. La farmacorresistencia en cepas bacterianas, especialmente en *Mycobacterium tuberculosis*, representa uno de los mayores desafíos. (Murray et al., 2022)]. La alta prevalencia de cepas farmacorresistentes incrementa la tasa de mortalidad, y dificulta el tratamiento eficaz de la tuberculosis. La capacidad de diagnosticar rápidamente la farmacorresistencia es crucial para implementar tratamientos adecuados y reducir la mortalidad asociada. Esta problemática subraya la urgente necesidad de desarrollar métodos diagnósticos rápidos, precisos y estandarizados para la detección de farmacorresistencia (World Health Organization, 2023).

Los métodos fenotípicos, aunque precisos, son laboriosos y requieren tiempo, mientras que los métodos moleculares pueden no capturar toda la variabilidad genética responsable de la resistencia. Esta discrepancia entre métodos resalta la necesidad de innovar y mejorar las técnicas diagnósticas disponibles. La detección de mutaciones que confieren resistencia mediante métodos moleculares es una alternativa rápida y precisa a las pruebas fenotípicas de fármaco-sensibilidad fenotípica (PFS). La secuenciación de genoma completo (SGC) brinda información genética integral sobre los genes relacionados con la resistencia a los fármacos antituberculosis. Los avances en la SGC de *Mycobacterium tuberculosis* han permitido incrementar la rapidez del ensayo de semanas a 2-5 días (Alaridah et al., 2019; Coll et al., 2018; Colman et al., 2019; Comas & Gil, 2016; Jeanes & O'Grady, 2016; Kazumi & Mitarai, 2012; Madrazo-Moya et al., 2019; Roy et al., 2018; Walker et al., 2013). La OMS ha publicado recientemente una guía para el uso de las tecnologías de SGC para la detección de mutaciones asociadas con TB-DR(World Health Organization, 2018b). Sin embargo, la alta variabilidad de la farmacorresistencia en *Mycobacterium tuberculosis* requiere un enfoque diagnóstico que pueda considerar la diversidad de los perfiles genéticos y las variaciones entre regiones del mundo donde se pueda aislar a la bacteria. Considerando esto, México debe tomar acción para incorporar de manera acelerada esta tecnología emergente como el apoyo al diagnóstico de la inteligencia artificial en sus diversas formas (*ML, DL, CNN*), con el fin de disminuir la incidencia de infecciones y complicaciones de la tuberculosis.

Aunque los estudios de asociación del genoma completo (GWAS) han proporcionado valiosa información sobre los mecanismos de resistencia, su aplicación en el diagnóstico clínico sigue siendo limitada. Con un número creciente de aislados clínicos de *Mycobacterium tuberculosis* sometidos a SGC y un número creciente de variantes identificadas asociadas a la farmacorresistencia, la inteligencia artificial (IA) ofrece un enfoque complementario para los estudios de asociación de genoma completo (GWAS), ya que tiene una capacidad superior para adaptarse al creciente número de datos clínicos y biológicos.

La IA no solo puede complementar los hallazgos de GWAS, sino también proporcionar una herramienta práctica y eficiente para el diagnóstico en tiempo real, mejorando la capacidad de respuesta clínica. En comparación con la GWAS, la IA en especial los métodos de aprendizaje automático no paramétricos proporcionan una mayor flexibilidad para resolver problemas de predicción en espacios variables de alta dimensión, cuando cada variable individual puede contener información limitada y las interacciones variables son importantes (Deelder et al., 2019; Friedman et al., 2008; Lunetta et al., 2004; Witten et al., 2009).

La discrepancia observada entre los resultados de métodos moleculares y fenotípicos para el diagnóstico de la farmacorresistencia ha sido una barrera significativa en la lucha contra la tuberculosis(Ahmad et al., 2016; Brandao et al., 2020; Kang et al., 2019. La integración de IA puede ayudar a cerrar esta brecha, proporcionando un método diagnóstico que combina la precisión de los análisis moleculares con la relevancia clínica de los fenotípicos.

El desarrollo de un método de análisis que permita clasificar el agente etiológico con respecto a su perfil de farmacorresistencia puede llevarse a cabo por métodos de *ML* con un costo menor en tiempo y una precisión mayor que los métodos tradicionales fenotípicos. Una vez generado, este método de análisis *ML* podrá brindar una caracterización más amplia de los patrones genéticos en los diferentes genomas de *Mycobacterium tuberculosis* y permitirá observar además el flujo de los aislados. La presencia de diferentes linajes y brotes epidemiológicos predominantes o circulantes en la región. Reduciendo la carga nacional y posteriormente global de la TB mediante la innovación tecnológica y la mejora de las prácticas diagnósticas actuales.

Marco Teórico

La TB como enfermedad global

La tuberculosis, causada por las bacterias del complejo *Mycobacterium tuberculosis*, ha sido una amenaza de salud constante a lo largo de la historia humana, junto con otras enfermedades como la malaria. Históricamente ha causado entre uno y un millón y medio de muertes de manera constante en los últimos 20 años, por lo que se mantiene como una preocupación global debido a la emergencia de pandemias y la reemergencia de enfermedades. La transmisión de la TB generalmente ocurre por vía aérea desde pacientes que presentan lesiones pulmonares "abiertas" que, al toser, liberan aerosoles que contienen pequeñas partículas líquidas, conocidas como gotas de Flügge, cada una encapsulando uno o dos bacilos. Cuando estas gotas se evaporan, los núcleos de bacilos quedan suspendidos en el aire, moviéndose con las corrientes y pudiendo ser inhalados por otras personas. Las partículas mayores de 10 µm son capturadas por la barrera mucosa de las vías respiratorias superiores y eliminadas por el sistema mucociliar. Sin embargo, las partículas más pequeñas, de entre 1 y 5 µm, pueden alcanzar los alvéolos, donde inician la primoinfección.

Cuando los bacilos de *Mycobacterium tuberculosis* alcanzan los alvéolos, la mayoría son eliminados por los macrófagos. Sin embargo, aproximadamente el 10% de las personas infectadas desarrollarán la enfermedad, con la mitad de estos casos manifestándose a los meses después de la infección y el resto, aproximadamente un 5%, después de un período prolongado que puede extenderse por décadas, durante el cual la micobacteria latente y presente en lesiones aparentemente resueltas en el pulmón, se reactiva y reinicia el proceso infeccioso. La infección inicial, o primoinfección tuberculosa, comienza con un foco de alveolitis exudativa donde los macrófagos atacan a las micobacterias. Si la carga bacteriana no es excesiva, la infección puede no avanzar más allá de esta fase local. Si la infección se disemina a través de las vías linfáticas hasta los ganglios regionales, se forma un complejo bipolar que puede incluir la diseminación de bacilos a órganos distantes como los riñones, hígado y huesos. Generalmente, estas diseminaciones son controladas localmente sin mayores consecuencias clínicas(Lozano, 2002) .

La inmunodeficiencia humana (VIH) presente en un paciente con TB, exacerbado el impacto de ambas enfermedades, especialmente en las regiones más desfavorecidas. La coexistencia de la infección por el virus de VIH y TB representa un desafío significativo para la salud global. Tratar ambas infecciones simultáneamente implica el uso combinado de terapia antirretroviral (TAR) y medicamentos antituberculosos, lo que puede generar interacciones entre medicamentos que afectan la eficacia y seguridad del tratamiento, así como el bienestar del paciente(Navasardyan et al., 2024).

el incremento reciente de la relación entre la TB y la diabetes mellitus (DM) también está representando un desafío creciente para la salud pública global, especialmente en países en desarrollo donde ambos problemas de salud son prevalentes. Investigaciones han evidenciado que la DM tipo 2 aumenta el riesgo de desarrollar TB activa, lo que complica los esfuerzos para controlar la propagación de la TB. Este vínculo podría complicar aún más el control de la TB a nivel mundial, ya que los mecanismos inmunológicos alterados en pacientes diabéticos podrían facilitar la infección por *Mycobacterium tuberculosis* (Kumar Nathella & Babu, 2017). Se sugiere que un enfoque integrado, similar al empleado en la co-infección TB-VIH, para la prevención, detección y tratamiento podría ser efectivo. Este enfoque requiere una planificación cuidadosa y una colaboración intensiva entre los programas de salud pública y las entidades encargadas de controlar estas enfermedades para manejar adecuadamente la doble carga de la TB y la DM.

A pesar de los esfuerzos globales, la TB sigue siendo prevalente, reflejando las desigualdades socioeconómicas entre países. Los países desarrollados han reducido significativamente su prevalencia, pero la globalización y la migración han reintroducido y dispersado la enfermedad a áreas donde previamente se consideraba controlada. En 2014 y 2015, todos los Estados miembros de la OMS y de las Naciones Unidas se comprometieron a poner fin a la epidemia de TB mediante la adopción de la Estrategia Fin a la TB de la OMS y los Objetivos de Desarrollo Sostenible (ODS) de las Naciones Unidas (World Health Organization, 2015.

La estrategia incluía hitos y objetivos) para lograr grandes reducciones en la tasa de incidencia de la TB, el número absoluto de muertes causadas por la TB y los costes a los que se enfrentan los pacientes con TB y sus hogares. Los requisitos clave para alcanzar los hitos y objetivos se definieron dentro de los tres pilares de la Estrategia Fin a la TB. Incluían la prestación de servicios de prevención, diagnóstico y tratamiento de la TB en el contexto del progreso hacia la cobertura sanitaria universal y la protección social; acciones multisectoriales para abordar los determinantes sociales y económicos más amplios de la TB; y avances tecnológicos (World Health Organization, 2023).

Uno de los objetivos principales de *“la Estrategia Hacia el Fin de la Tuberculosis”,* es que ningún paciente con TB y sus familias tengan que hacer frente a costes totales catastróficos como consecuencia de la enfermedad, se estableció la necesidad de la eliminación de las barreras financieras y económicas para acceder al diagnóstico y tratamiento de la tuberculosis. Catastrófico" se define como los gastos médicos directos, los gastos no médicos directos y los costes indirectos (por ejemplo, pérdidas de ingresos) que suman más del 20% de los ingresos familiares. Con este acuerdo se declara la necesidad de la innovación tecnológica que permita acortar el tiempo de diagnóstico, mejor el seguimiento y acceso a los servicios de salud con el fin de disminuir el impacto económico en las familias afectadas por la TB(World Health Organization, 2023a.

El diagnóstico de TB sigue siendo un aspecto crítico y el más vulnerable dentro de la atención de esta enfermedad. La falta de diagnóstico adecuado, o la realización de diagnósticos sin las pruebas de susceptibilidad a medicamentos necesarias, incrementa la morbilidad y el riesgo de mortalidad. En 2020, las muertes atribuidas a la TB aumentaron por primera vez desde 2005, alcanzando aproximadamente a 1,5 millones de personas, incluyendo a 214.000 individuos con VIH (World Health Organization, 2021a, 2023). De los 10 millones estimados que desarrollaron TB activa ese año, solo 5,8 millones fueron diagnosticados oficialmente, reflejando una caída del 18% comparado con 2019 debido al impacto de la COVID-19. Estos reportes destacan la urgente necesidad de mejorar el acceso a herramientas avanzadas para la detección y el diagnóstico de la TB.

Los retos incluyen la falta de herramientas adecuadas, los altos costos, el estigma y otros factores socioeconómicos que restringen el acceso a los servicios de salud, así como las deficiencias en los sistemas de salud y la lenta adopción de herramientas disponibles por parte de los programas nacionales (David Branigan, 2020; Branigan et al., 2021. Superar estas barreras, iniciando el diagnóstico en las etapas tempranas de síntomas y ofreciendo pruebas universales de resistencia a los medicamentos, es crucial para cerrar la brecha en el diagnóstico de TB y reducir su transmisión [Drain 2018]. La Organización Mundial de la Salud (OMS) está mejorando los protocolos reguladores y trabajando en revisión de nuevas herramientas diagnósticas. A pesar de los avances, es crucial reconocer que ninguna herramienta por sí sola solucionará la brecha diagnóstica ni erradicará la TB. Se necesitan estrategias integradas que incluyan una variedad de herramientas diseñadas para diferentes contextos y que sean económicamente accesibles para asegurar su adopción a nivel mundial (Branigan et al., 2021

Farmacorresistencia en *Mycobacterium tuberculosis*

La farmacorresistencia en *Mycobacterium tuberculosis* es una de las principales barreras en el tratamiento eficaz de la TB. La farmacorresistencia a los medicamentos antituberculosis se produce cuando las cepas desarrollan mutaciones genéticas que les permiten sobrevivir a la exposición a estos fármacos. Esta resistencia puede ser adquirida durante el tratamiento inadecuado o puede ser transmitida de persona a persona. Numerosos genes han sido identificados como responsables de la farmacorresistencia en *Mycobacterium tuberculosis*. Algunos de los genes clave incluyen:

- *rpo*B: Asociado con la resistencia a la rifampicina. Las mutaciones en el gen *rpo*B afectan la subunidad beta de la ARN polimerasa.
- katG: Relacionado con la resistencia a la isoniazida. Las mutaciones en *kat*G, que codifica la catalasa-peroxidasa, impiden la activación del fármaco.

- *inh*A: También asociado con la resistencia a la isoniazida. Las mutaciones en *inh*A afectan la enoyl-ACP reductasa, una enzima esencial en la síntesis de ácidos micólicos.
- *emb*B: Asociado con la resistencia al etambutol. Las mutaciones en *emb*B afectan la arabinosil transferasa.
- *pnc*A: Relacionado con la resistencia a la pirazinamida. Las mutaciones en *pnc*A afectan la pirazinamidasa, necesaria para la conversión de pirazinamida a su forma activa.
- *gyr*A y *gyr*B: Relacionados con la resistencia a las fluoroquinolonas. Las mutaciones en estos genes afectan la ADN girasa.

Este grupo de genes ha sido clasificado por OMS en dos categorías: *"Tier 1"* y *"Tier 2".* Esta clasificación es un un sistema de priorización para analizar regiones del genoma, basado en la probabilidad de que contengan mutaciones relacionadas con la resistencia a tratamientos o medicamentos. El *Tier 1,* Se refiere a los genes y regiones genómicas que son considerados los más probables de contener mutaciones de resistencia. Estas mutaciones son las que cumplen los criterios más estrictos para ser consideradas relevantes durante la evaluación inicial de confianza. El *Tier 2*, comprende genes que, aunque no sean tan prioritarios como los de *Tier 1*, tienen una probabilidad razonable de contener mutaciones de resistencia. También incluye secuencias promotoras específicas que han sido definidas por la literatura científica como relevantes. El análisis de estas regiones de *Tier 2* solo se recomiendas si las mutaciones de *Tier 1* no permiten una interpretación clara. Es decir, si en un aislamiento no se detectan mutaciones que ya sean concluyentes en *Tier 1*, se pasa a investigar las mutaciones de *Tier 2 (World Healt Organizartion, 2021)*.

Clasificación de la Farmacorresistencia

La farmacorresistencia en *Mycobacterium tuberculosis* se clasifica principalmente en las siguientes categorías:

- Monorresistencia: Resistencia a un solo fármaco antituberculoso.
- Polirresistencia: Resistencia a más de un fármaco antituberculoso, pero no incluye la combinación de isoniazida y rifampicina.
- Multirresistencia (MDR-TB): Resistencia al menos a isoniazida y rifampicina, los dos fármacos de primera línea más potentes.
- Resistencia Extensiva (XDR-TB): Resistencia a isoniazida y rifampicina, además de cualquier fluoroquinolona y al menos uno de los fármacos inyectables de segunda línea (amikacina, kanamicina o capreomicina).
- Extremadamente resistentes a fármacos (XXDR-TB): Resistencia a prácticamente todos los fármacos disponibles, dejando pocas o ninguna opción de tratamiento.

Cepas, Linajes y Tipos de Resistencia

Las cepas de *Mycobacterium tuberculosis* se agrupan en diferentes linajes genéticos, los cuales pueden tener patrones de resistencia distintos.

- Linaje Euro-Americano: Común en Europa y América del Norte.
- Linaje Este asiático: Predominante en Asia Oriental.
- Linaje Indo-Oceánico: Encontrado en el subcontinente indio y las islas del Océano Índico.
- Linaje de África Occidental: Presente principalmente en África Occidental.

Cada linaje puede presentar diferentes tasas de mutaciones y, por lo tanto, variar en sus perfiles de farmacorresistencia. Esta diversidad genética contribuye a la complejidad en el diagnóstico y tratamiento de la TB. Debido a la esta diversidad, la identificación de la presencia de los polimorfismos asociados a resistencia de *Mycobacterium tuberculosis* es un desafío significativo en el tratamiento de la TB. Esta diversidad se debe a las diferentes combinaciones de mutaciones en los genes asociados a la resistencia y a las diferencias en los linajes genéticos. Los métodos fenotípicos y moleculares actuales a menudo muestran discrepancias en los resultados debido a esta diversidad.

Principios de farmacogenética en el tratamiento de TB.

En el tratamiento primario de la TB se utilizan fármacos de primera línea: Isoniacida (H), Rifampicina (R), Pirazinamida (Z), Etambutol (E) y Estreptomicina (S). Para tratar la TB-FR se utilizan fármacos de segunda línea: kanamicina (Km), amikacina (Am), capreomicina (Cm), etionamida (Eto), protionamida (Pto), ofloxacina (Ofx), levofloxacina (Lfx), moxifloxacina (Mfx) y ciclocerina (Cs). (WHO, 2019). Los fármacos antituberculosis actúan sobre la *Mycobacterium tuberculosis* por tres mecanismos: 1) bloquean las enzimas que sintetizan la pared celular, 2) interrumpen la síntesis a nivel ribosomal y 3) interfieren en procesos en síntesis de RNA/DNA (Nasiri et al., 2017).

Se ha observado que la presencia de polimorfismos de un solo nucleótido (SNPs) resultan en modificaciones en el metabolismo de fármacos específicos como en el metabolismo intrínseco de la bacteria limitando su supervivencia (Palomino & Martin, 2014). Una mutación de resistencia puede modificar directamente la acción de un fármaco o compensar a través de la activación de una ruta alternativa, estas mutaciones pueden causar resistencia a múltiples fármacos y generar complejas interacciones gen-gen (Safi et al., 2013; Trauner et al., 2014; Gygli et al., 2017

Mecanismos moleculares generadores de farmacorresistencia en fármacos de primera línea

La isoniacida actúa durante la fase de replicación activa por medio de la enzima catalasaperoxidasa, y posteriormente uniéndose a la enzima inhA, provocando la inhibición de la síntesis del ácido micólico de la pared bacteriana aumentando la sensibilidad a la acción de los

radicales libres de oxígeno reactivo. La catalasa peroxidasa es codificada por el gen katG. Aunque el gen tiene un tamaño de 1771 pares de bases, más del 65% de las mutaciones relacionadas con farmacorresistencia se localizan en 3 codones siendo el más importante el codón 315 donde la Serina (Ser) cambia a treonina (Thr), asparagina(Asn) o arginina (Arg)(Cuevas-Córdoba and Zenteno-Cuevas, 2010). La enoil-tioéster reductasa *InhA* cataliza un paso esencial en la biosíntesis de ácidos grasos, esta enzima cataliza la reducción de enoil-ACP dependiente de NADH en la biosíntesis de ácidos grasos y ácidos micólicos, que forman un componente esencial de la membrana y la pared celular de *Mycobacterium tuberculosis* [Pecho silva 2019]. La OMS reporta que mutaciones en 5 codones de este gen participan en la inhibición de la unión de la enzima con isoniacida favoreciendo la resistencia al fármaco. Sin embargo, en la literatura se reporta cepas con fenotipo resistente a isoniacida que no tienen mutaciones en *katG* o *InhA*, sugiriendo otros posibles mecanismos de resistencia entre los que se han propuesto mutaciones en los genes *ahpC, ndh, fabG1, Rv1258c, mshA, Rv2752c, Rv1258c*, sin embargo, las mutaciones en estos genes aún no han sido asociadas oficialmente a farmacorresistencia por parte de la OMS (World Healt Organizartion, 2021).

La rifampicina se une a la RNA polimerasa inhibiendo la síntesis del ácido nucleico durante la replicación bacteriana. La RNA polimerasa es codificada por los genes *rpoA, rpoB, rpoC, rpoD*. Mutaciones en una región de 81 pb del gen *rpoB* han demostrado que disminuyen la afinidad de la rifampicina a la RNA polimerasa dando lugar a la resistencia. La OMS reconoce 24 codones en rpoB relacionados al origen de la farmacorresistencia, sin embargo, se ha reportado la posible contribución de los genes Rv2752c, *rpoA, rpoC, embB* en el origen a la resistencia al mismo fármaco por otros mecanismos.

Diagnóstico de *Mycobacterium tuberculosis*

La OMS y la Fundación para Nuevos Diagnósticos Innovadores (FIND) han desarrollado una serie de Perfiles de Productos Objetivo (Target producto profiles,TPP) para brindar orientación a los desarrolladores de pruebas diagnósticos sobre los tipos de nuevas herramientas que se necesitan según los diferentes casos de uso. Al momento FIND determina las necesidades, con su respectivo TPP. Estos TPP establecen los criterios operativos y de rendimiento óptimos y mínimos que estas herramientas deben cumplir, incluida la precisión, la configuración, el tiempo para obtener resultados y el precio. Analizar los métodos diagnósticos disponibles con el enfoque de los TPP permite identificar puntos críticos de mejora y las necesidades para alcanzar las metas de los ODS (Tabla 1). En el contexto de este proyecto, su análisis permite contextualizar la utilidad de las tecnologías emergentes de información, como la inteligencia artificial (IA).

Pruebas de detección y clasificación de la tuberculosis

Las herramientas para la detección (*screening*) de TB, como la detección de síntomas y la radiografía de tórax, pueden identificar a las personas que pueden tener TB deban someterse a una evaluación adicional para diagnosticar TB. Estas herramientas también se pueden utilizar para clasificar a las personas que acuden a recibir atención en centros de salud para identificar si una persona debe ser evaluada más a fondo. En marzo de 2021, la OMS publicó directrices sobre la detección de TB recomendando la implementación de pruebas de detección sistemáticas entre poblaciones de alto riesgo y en comunidades con alta carga de TB (World Health Organization, 2021b). Esta esta ocasión fue la primera vez que se recomendó el uso de la detección asistida por computadora (CAD) para ayudar en la interpretación de las radiografías de tórax. Estas recomendaciones permiten ampliar en gran medida las posibilidades de detección de la tuberculosis, incluso antes de la aparición de síntomas (TB subclínica), incorporando algoritmos de IA para apoyar toma de decisiones clínicas (Branigan et al., 2021).

Prueba de biomarcadores distintos del esputo en el lugar de atención

La mayoría de los casos de TB pulmonar se diagnostican mediante baciloscopia. Sin embargo, esta tiene una sensibilidad subóptima en los niños y las personas infectadas con VIH. Además, se pueden presentar tienen dificultades para obtener muestras de esputo de buena calidad. Una prueba de clasificación (*triage*) en el lugar de atención para descartar la tuberculosis, que debería ser una prueba sencilla y de bajo costo que puedan utilizar los proveedores de atención médica de primer contacto para identificar a quienes necesitan más pruebas. Se recomienda una prueba no basada en esputo en el lugar de atención, capaz de detectar todas las formas de TB mediante la identificación de biomarcadores característicos. (World Health Organization, 2014) Las pruebas moleculares rápidas son capaces de diagnosticar y confirmar microbiológicamente la presencia de TB de forma rápida y precisa (Haraka et al., 2021). Sin embargo, en 2022, sólo el 63% de las personas diagnosticadas con TB pulmonar fueron microbiológicamente confirmadas, ya sea mediante el uso de una prueba molecular rápida recomendada por la OMS o por baciloscopia a pesar de que ya no se recomienda esta última (World Health Organization, 2023). Las pruebas moleculares rápidas recomendadas por la OMS tienen una sensibilidad para la TB de hasta el 90% (Gene Xpert MTB/RIF Ultra) en comparación con la baciloscopia con una sensibilidad media del 50% (World Health Organization, 2009) . Sin embargo, en 2020, sólo 3.5 millones o el 47% de los 7.5 millones de personas diagnosticadas con TB recibieron una prueba molecular rápida recomendada por la OMS como prueba inicial (World Health Organization, 2023) Al momento las dos pruebas recomendadas por la OMS son: GeneXpert de Cepheid y TB truenat de Molbio, estas pruebas tienen la ventaja de además de confirmar el diagnostico de TB permiten identificar un perfil básico de farmacorresistencia. GeneXpert es una plataforma de reacción en cadena de la polimerasa (PCR) en tiempo real automatizada, integra la extracción, purificación, amplificación y detección dentro de un cartucho dando resultados en menos de 3 horas. Pero a pesar de estar disponible desde 2010 no se ha tenido acceso a esta tecnología al ritmo deseado y recomendado por la OMS, esto debido a los altos costos del servicio, requisitos de infraestructura,

mantenimiento y posiblemente a la ausencia de presión competitiva (Branigan et al., 2021). TB truenat es un chip de basado en PCR de tiempo real que requiere una extracción de ADN previa, a pesar de ser menos automatizada tiene la ventaja de poder operar a temperaturas de hasta 40 °C, disminuyendo los requerimientos de infraestructura y permitiendo colocarse como una solución en laboratorios más cercanos al punto de atención con respecto a GeneXpert. La Alianza "*Stop TB partnership* han publicado recientemente una guía de implementación de Truenat donde indica que pueden ampliarse y aplicarse en los países junto con la infraestructura existente de GeneXpert (Stop TB Partnership, 2021. Tras la introducción de Truenat, otras empresas también están desarrollando pruebas moleculares rápidas para la TB, entre ellas SD Biosensor, Bioneer y LumiraDx, con el principal objetivo de aumentar la portabilidad, velocidad y asequibilidad.

Pruebas de susceptibilidad a medicamentos de TB de próxima generación en centros periféricos

La Estrategia Fin a la TB exige un diagnóstico temprano y un tratamiento rápido de todas las personas de todas las edades con cualquier forma de tuberculosis. Esto requiere garantizar el acceso universal a las pruebas de susceptibilidad a los medicamentos (PSD) para todas las personas con signos y síntomas de TB y ya no dar prioridad únicamente a las personas en mayor riesgo de TB MDR y/o TB asociada al VIH.

La OMS recomienda el acceso universal a las PSD al menos para rifampicina, y una posterior PSD para al menos las fluoroquinolonas entre todos los pacientes con TB con resistencia a la rifampicina. Los métodos fenotípicos de PSD son actualmente el estándar de oro para la detección de resistencia a los medicamentos, pero estos métodos requieren mucho tiempo, una infraestructura de laboratorio sofisticada, personal calificado y un estricto control de calidad.

Para realizar la PSD fenotípica, las micobacterias a menudo se cultivan inicialmente en una variedad de medios de cultivo sólidos o líquidos. Los medios más utilizados son Löwenstein-Jensen (LJ), agar Middlebrook 7H10 (7H10), agar enriquecido Middlebrook 7H11 (7H11) y caldo Middlebrook 7H9. Este último se utiliza como medio para el sistema automatizado de cultivo de *Mycobacterium tuberculosis* con tubo indicador de crecimiento de micobacterias (MGIT). Los sistemas de cultivo líquidos comerciales para la PSD reducen el tiempo para obtener resultados a tan solo 10 días, en comparación con los 28 a 42 días necesarios para la PSD en los medios sólidos (World Health Organization, 2018a).

El uso de tecnologías de secuenciación de próxima generación para la detección de mutaciones asociadas con la resistencia a medicamentos en el complejo Mycobacterium tuberculosis

La farmacorresistencia se diagnostica tradicionalmente mediante medios de cultivo y pruebas fenotípicas proceso lento y costoso, presentando dificultad de reproductibilidad e imprecisiones (Farhat et al., 2016). La secuenciación de próxima generación (NGS) tiene un gran potencial como método para diagnosticar rápidamente la TB farmacorresistente (TB-DR).

Sin embargo, la adopción de estas tecnologías para el diagnóstico de la TB-DR se ha visto obstaculizada por los costos, la integración en los flujos de trabajo de laboratorio existentes, la capacitación técnica y los requisitos de habilidades para la utilización de la tecnología, además de la necesidad de orientación experta con respecto al manejo y la interpretación clínica de los datos de secuenciación (World Health Organization, 2023b).

A diferencia de otros ensayos moleculares para el diagnóstico de resistencia en TB, que se basan en la identificación indirecta de MTB y un conjunto limitado de mutaciones de resistencia a través de la hibridación de sondas con secuencias genéticas específicas, los ensayos NGS pueden proporcionar información de secuencia detallada para múltiples regiones genéticas o genomas completos de interés. Todas las plataformas de secuenciación dependen de un flujo de trabajo básico similar para obtener lecturas de secuenciación de genomas de TB presentes en muestras clínicas; (1) primero se extrae el ADN de muestras clínicas o aislados cultivados; (2) el ADN pasa por un procesamiento enzimático de sintesis; (3) se determina el nucleótido presente en múltiples fragmentos de ADN en paralelo, (4) y luego se utilizan análisis bioinformáticos para mapear las lecturas individuales mediante su comparación con un genoma de referencia(Dolinger et al., 2016)

Específicamente, los ensayos NGS pueden confirmar la presencia o ausencia de inserciones y deleciones (indeles) y evaluar la aparición de mutaciones raras y otros datos que pueden no ser detectados mediante otros ensayos moleculares, como la presencia de heterorresistencia o una mezcla de múltiples poblaciones genéticas, en una muestra clínica (Colman et al., 2015). Los ensayos NGS son flexibles, ya que pueden programarse para una variedad de aplicaciones, incluida la evaluación de información genética para organismos adicionales que pueden estar presentes en una muestra clínica. A pesar de las ventajas de la NGS sobre otros métodos moleculares para la identificación y caracterización de la TB-DR, la adopción de estas tecnologías se ha visto obstaculizada, especialmente en países de ingresos bajos y medianos, principalmente por la falta de soluciones de almacenamiento y análisis de datos fácilmente disponibles, y la falta de soluciones "plug-and-play" capaces de obtener información de secuenciación directamente de muestras clínicas primarias. (Dolinger et al., 2016).

En Julio del 2023 la OMS emitió un comunicado sobre las tecnologías NGS, en donde, después de una revisión sistemática concluyo que la NGS es precisa, rentable según el contexto, aceptable e implementable en condiciones de rutina a pesar de la complejidad inherente. La OMS respalda el uso de NGS dirigidas para detectar la resistencia a los medicamentos después del diagnóstico de tuberculosis, para guiar la toma de decisiones clínicas para el tratamiento de la TB resistente a los medicamentos World Health Organization, 2023b). Destacando 3 productos comerciales para el uso clínico:

- Deeplex® Myc-TB (GenoScreen): para rifampicina, isoniazida, pirazinamida, etambutol, fluoroquinolonas, bedaquilina, linezolid, clofazimina, amikacina y estreptomicin.

- NanoTB® (Oxford Nanopore Technologies): para rifampicina, isoniazida, fluoroquinolonas, linezolid, amikacina y estreptomicina.
- TBseq® (ShengTing Biotech): para etambutol.

De estos tres, Deeplex® Myc-TB, ha tenido más presencia en el mercado mexicano, lo cual lo hace el más relevante para comparar los modelos de predictivos que desarrollen. Esta es una prueba basada en la secuenciación de nueva generación (NGS) diseñada para la identificación y análisis de cepas del complejo *Mycobacterium tuberculosis* (MTBC). Esta herramienta permite la identificación de especies micobacterianas, la genotipificación a nivel de sublinajes y la predicción de la resistencia a fármacos antituberculosos, todo a partir de una única muestra clínica. La prueba se basa en la secuenciación profunda de 24 amplicones que cubren 18 genes principales asociados con la resistencia a medicamentos de primera y segunda línea.

La capacidad de secuenciación profunda permite la detección de variantes en subpoblaciones heterorresistentes que representan tan solo el 1-3% de la muestra, lo que supera la capacidad de otras pruebas moleculares rápidas. El sistema está diseñado para funcionar con muestras de ADN extraídas de muestras clínicas inactivadas por calor o etanol y tiene una sensibilidad para detectar genomas micobacterianos a niveles por debajo del límite de detección de la microscopía clásica. El análisis directo de las variantes de genoma completo permite la identificación de especies mediante el análisis del gen hsp65; la genotipificación de cepas de *Mycobacterium tuberculosis* a nivel de sublinajes, utilizando el locus CRISPR/Direct Repeat y SNPs filogenéticos; y la predicción de resistencia a fármacos comparando mutaciones detectadas con bases de datos de referencia, incluyendo el catálogo de mutaciones de la OMS (World Healt Organizartion, 2021; Sibandze et al., 2022; 'From clinical sample to drug resistance profile Deeplex ® Myc-TB USER MANUAL', 2023).

Recientemente se ha incrementado el uso del análisis del genoma completo como herramienta diagnóstica para identificar rápidamente un amplio panel de mutaciones que brindan información clínica para la toma de decisiones (Satta et al., 2018). La SGC puede usarse para identificar loci de resistencia a través de estudios de asociación amplia del genoma (GWAS) y puntos de convergencia evolutiva por medio los árboles basados en filogenia (Coll et al., 2018). Los GWAS utilizan métodos clásicos de regresión con o sin incorporación de técnicas de regularización, estos métodos pueden fallar para detectar interacciones entre covariantes y podrían ser subóptimos para analizar grandes bases de datos con altas dimensiones (Lunetta et al., 2004; Heidema et al., 2006).

Fundamentos de *ML* aplicados a la bioinformática

La introducción de la tecnología de la información en el campo de la asistencia sanitaria ha proporcionado mejoras en la toma de decisiones clínicas. Con el creciente uso de tecnologías digitales se ha recopilado una mayor cantidad de datos de los que se pueden analizar, sin embargo, con los últimos avances en análisis de datos y sistemas de toma de decisiones, superar este desafío parece ser finalmente factible (Alsuliman, Humaidan and Sliman, 2020).

La inteligencia artificial (IA), se refiere a la capacidad de un sistema para interpretar datos externos correctamente, aprender de dichos datos y usar esos aprendizajes para lograr objetivos y tareas específicos utilizando una adaptación flexible (Alsuliman, Humaidan and Sliman, 2020). La utilidad de la IA se ha estudiado en múltiples áreas de la salud y de la práctica médica incluyendo medicina de precisión, salud poblacional, procesamiento de imágenes en radiología, dermatología y oftalmología transformando la práctica médica (Kulkarni and Jha, 2020). La IA permite analizar grandes conjuntos de datos digitales que se encuentran actualmente a disposición para generar modelos capaces de capacitarse a sí mismos en una tarea específica (Kulkarni and Jha, 2020).

El componente principal de AI es el aprendizaje automático (*ML*). El aprendizaje automático se da cuando las computadoras son utilizadas para aplicar modelos estadísticos a los datos. Es una subdisciplina de la IA, donde los programas de computadora (algoritmos) aprenden las relaciones entre los datos de entrada y salida, ofrece una manera eficiente de capturar el conocimiento mediante la información contenida en los datos, para mejorar de forma gradual el rendimiento de modelos predictivos y tomar decisiones basadas en dichos datos (Kononenko, 2001). Se pueden distinguir entre tres categorías de algoritmos de aprendizaje automático: supervisado, no supervisado y de refuerzo.

En el aprendizaje supervisado, los programas de computadora aprenden asociaciones mediante el análisis de muestras de datos definidas por un supervisor (generalmente un experto humano) en un proceso llamado Capacitación. Una vez que se han aprendido las asociaciones, se pueden usar para predecir ejemplos futuros en un proceso llamado pruebas (Panch, Szolovits and Atun, 2018).

En el aprendizaje no supervisado, los programas informáticos aprenden asociaciones en los datos sin una definición externa de asociaciones. A menudo se usa para la agrupación, es decir, extraer correlaciones no descubiertas en los datos de entrada de tal manera que se formen subconjuntos de datos que comparten características comunes (Alsuliman, Humaidan and Sliman, 2020).

En el aprendizaje por refuerzo, el sistema aprende a comportarse basado en una señal escalar de recompensa / castigo. El castigo puede considerarse como una señal de recompensa negativa que refuerza una acción que evita su entrega (Doya, 2007).

Hay un campo particular del aprendizaje automático, que se usa para grandes conjuntos de procesamiento de datos llamados aprendizaje profundo (*deep learning*, DL). DL es un sistema computacional basado en las neuronas que determina las correlaciones entre los datos mediante pruebas evolutivas o por capas para reducir una función de costo. El aprendizaje profundo es una

herramienta poderosa para aprender problemas cognitivos complejos (De Fauw et al., 2018). Sin embargo, el bajo volumen de datos, la alta dispersión y la baja calidad de los datos pueden limitar la eficacia de los métodos de aprendizaje profundo (Miotto et al., 2018).

Estado del arte.

Actualmente para el análisis del genoma completo los métodos más utilizados de *ML* son *support vector machine* (SVM), *logistic regression* (LR), *product-of-marginals* (PM), random forest (RF) *gradient boosting tree* (GBT), *class-conditional Bernoulli mixture model* (CBMM), *k-nearest neighbor* (kNN), *artificial neural network* (ANN), *sequential minimization optimization* (SMO), *neuronal network* (NN) y algoritmos naive Bayes (NB).

El más usado de acuerdo a seis estudios, fue SVM reportando una precisión desde el 93.89% hasta 73% (Yang et al., 2018; Chowdhury, Khaledian and Broschat, 2019; Duffy et al., 2019; Kouchaki et al., 2019, 2020; Jamal et al., 2020; Kavvas et al., 2020). Kavvas 2018 reporta al utilizar únicamente SVM una precisión de SVM del 73%, en contraste Chowdhury 2019 reporta precisión mayor al 80% utilizando SVM y LR(Chowdhury, Khaledian and Broschat, 2019); Kouchaki 2019 reporta la precisión más alta de 93.89% al utilizar una combinación de SVM, LR, PM (Kouchaki et al., 2019).

Todos los estudios utilizan como genoma de referencia el genoma H37rv, pero presentan objetivos de genes de resistencia diferentes, por lo que la precisión para cada fármaco es diferente a la precisión global, en ambos estudios el método de *ML* fue más sensible que el de asociación directa, pero la utilización de los tres métodos juntos mostró un incremento mayor en la sensibilidad principalmente en Pirazinamida (PZA) pero muy poca diferencia en el resto de fármacos de primera línea y los de segunda línea (44.29% para PZA, 30.42% para CIP, 12% para AK, MOX, y OFX, 8% para EMB y KAN, y 4% para SM y CAP).

Deelder (2019) implementó un modelo basado en GBT y LR el cual presentó una sensibilidad muy alta para los fármacos de primera línea en especial para Isoniacida (INH) y Rifampicina (RIF) (Deelder et al., 2019). La sensibilidad del método de Deelder fue para RIF (88,8%) e INH (91,1%) fue mayor que para EMB (82,8%) y PZA (69,7%), en las fluoroquinolonas fue más alta para CIP (85,7%), seguida por OFL (81,0%) y MOX (53,3%). Duffy 2019 utilizó un modelo basado en LR reportando una sensibilidad 81.2% a 82.5% un resultado menor al método de Deelder. De acuerdo con estos resultados se observa mejores rendimientos cuando se utilizan modelos combinados que individuales (Duffy et al., 2019).

Yang (2019) utiliza un método basado en siete modelos de *ML* (LR-L1 y LR-L2, SVM-L2 y VM-RBF, RF, PM, CBMM) el cual reporta una precisión superior del 90% para todos los fármacos, presentado una mejora de 2 a 4% para INH, de 97% para RIF y EMB, 96% para ciprofloxacino (CIP), siendo uno de los métodos que reportan mayor sensibilidad (Yang et al., 2019).

Jamal 2020 utilizó cuatro NB, kNN, SVM, y ANN, siendo este uno de los pocos artículos que se utiliza ANN, la precisión global que reporta es del 70%, sin embargo, se encuentra que el mejor de los modelos es ANN con una precisión del 81.81% (InhA) al 100% (gyrA) (Jamal et al., 2020).

Objetivo general

Desarrollar modelos de *ML* para la predicción de resistencia a fármacos antituberculosis de primera y segunda línea a partir del análisis del genoma completo de aislados clínicos de *Mycobacterium tuberculosis*.

Objetivos Específicos

1. Proponer la estructura de una representación de la información genómica compatible con los modelos de inteligencia artificial.
2. Diseñar modelos de *ML* para la predicción de farmacorresistencia de aislados clínicos de *Mycobacterium tuberculosis a partir de genoma completo*.
3. Validar los modelos de *ML* diseñados para la predicción de farmacorresistencia de aislados clínicos de *Mycobacterium tuberculosis*. Mediante la concordancia de las predicciones con los patrones obtenidos mediante métodos fenotípicos.
4. Predecir nuevas variantes asociadas a farmacorresistencia en aislados clínicos de *Mycobacterium tuberculosis* a partir de genoma completo utilizando modelos de ML.

Hipótesis

Los modelos de *ML* y *Deep learning* alcanzarán una especificidad y sensibilidad superiores al 90% en la detección de perfiles de farmacorresistencia en genomas completos de *Mycobacterium tuberculosis*, mostrando una concordancia con los patrones obtenidos mediante métodos fenotípicos.

Metodología

Diseño del estudio

Este proyecto corresponde a un estudio exploratorio y de desarrollo tecnológico en inteligencia artificial aplicada al campo de la bioinformática. El proyecto se centra en el diseño y validación de modelos predictivos con la exploración e identificación de nuevas variantes genéticas asociadas a la farmacorresistencia. Se siguió un proceso por etapas para la implementación del proyecto. La primera se basó en definir y establecer la estructura de datos adecuada para representar la información genómica de *Mycobacterium tuberculosis*. La segunda etapa comprendió el desarrollo de varios modelos de ML y la exploración de múltiples configuraciones para las redes CNN, así como la optimización de hiperparámetros y la implementación de técnicas de entrenamiento y ajuste. En la tercera etapa se evaluaron los modelos desarrollados para determina su efectividad en la predicción de la farmacorresistencia. Finalmente, en una cuarta etapa, con los resultados de los modelos, se exploraron las variantes de mayor relevancia para las predicciones de los modelos y su relación con el metabolismo de los fármacos antituberculosis.

Metodología para objetivo 1. Propuesta de la estructura de representación de la información genómica

Para este proyecto se trabajó con los conjuntos de muestras de Secuencias de Genoma Completo de *Mycobacterium tuberculosis*, publicados los siguientes repositorios digitales públicos: PATRIC DataBase https://www.bv-brc.org/, NCBI https://www.ncbi.nlm.nih.gov/ y EBI https://www.ebi.ac.uk/, ReseqTB https://www.reseqtb.org/. Como criterios de inclusión se consideraron solo aquellos genomas que contaban con metadatos relacionados con el perfil de farmacorresistencia, datos clínicos y epidemiológicos. Adicionalmente a partir de los SNPs identificados en los genes de interés, se construyó una matriz binaria usando todos los SNPs reportados en los genes relacionados a los principales fármacos utilizados en el tratamiento de la tuberculosis: Rifampicina, Isoniazida, Etambutol. Sin embargo, se incluyeron además Etionamida y Rifabutina, con el fin de explorar etiquetas con mayor asimetría.

Además, para construir una matriz de entrenamiento estándar que permitio desarrollar modelos predictivos para un número mayor de fármacos se incluyeron en la selección de características un total de 49 genes para los principales 16 fármacos (Tabla 4) utilizados en el tratamiento de TB. Esto con el fin de identificar y analizar las mutaciones clave, para entender mejor los mecanismos de resistencia, y permitir que los modelos predictivos cuenten con un panel más extenso de fármacos que permitan ser utilizados en la práctica clínica para mejorar el diagnóstico y tratamiento de la tuberculosis resistente.

La matriz generada para el entrenamiento y validación de este proyecto incluye genes de Tier 1 y Tier 2 propuestos por la OMS (Tabla 3) (World Healt Organizartion, 2021). Para la representación de los nucleótidos en el análisis de los SNPs (Single Nucleotide Polymorphisms), se implementaron diversas estrategias de codificación numérica. En primer lugar, se realizó un filtrado de los SNPs relevantes asociados a farmacorresistencia (World Healt Organizartion, 2021). y posteriormente se evaluaron diferentes métodos de codificación: (1) una codificación one-hot, donde cada nucleótido se representaba mediante un vector binario exclusivo; (2) una codificación binaria basada en la posición de cada nucleótido; y (3) una codificación numérica, en la cual se asignaron valores específicos a los nucleótidos adenina (A), timina (T), citosina (C), y guanina (G), correspondientes a los números 1, 2, 3, y 4, respectivamente. Adicionalmente, se exploraron varias configuraciones de entrada, incluyendo la representación de los datos como vectores lineales (tensores unidimensionales), tensores bidimensionales cuadrados, tensores espaciados, y matrices de características de 49 canales. En todas estas configuraciones, la posición genómica de cada SNP se utilizó como criterio para determinar su orden y ubicación dentro de las matrices de características. Se probaron sistemáticamente cada una de estas configuraciones y métodos de codificación para identificar la combinación más eficaz. La efectividad de las distintas representaciones fue evaluada mediante técnicas de reducción de dimensionalidad, específicamente Análisis de Componentes Principales (PCA) y t-SNE, lo que permitió visualizar y comparar la distribución de los datos, facilitando la identificación de la codificación y configuración óptimas para los modelos de ML.. Posteriormente, se ajustó el vector resultante para que fuera compatible con la entrada en la red neuronal convolucional (CNN).

Metodología para objetivo 2. Diseño de los modelos de ML.

La metodología para abordar el objetivo 2 del presente proyecto pretende diseñar y desarrollar los modelos de ML específicamente orientados a predecir la resistencia a fármacos en aislados clínicos de *Mycobacterium tuberculosis.* Este proceso incluyó la selección de algoritmos adecuados (DT, RF, NN, CNN, entre otros), la definición de hiperparámetros, y la implementación de técnicas de entrenamiento y ajuste. Una vez construida la matriz de las variantes (SNPs) se siguió en parte un modelo de desarrollo en espiral, por medio del cual se llevan a cabo refinaciones progresivas que retroalimentan nuevas fases del ciclo de diseño, implementación y pruebas. Se procedió inicialmente con la división del conjunto de datos en tres segmentos: un 70% se asignó para el entrenamiento y la búsqueda del modelo óptimo, un 20% para propósitos de validación, y el restante 10% se reservó para la fase de prueba. La optimización de la arquitectura de redes neuronales se llevó a cabo mediante el uso de *Keras Tuner,* una herramienta diseñada para realizar una búsqueda eficiente de hiperparámetros (O'Malley, 2019). En este

estudio, se emplearon dos métodos de búsqueda: la búsqueda aleatoria y la búsqueda en cuadrícula, con el fin de identificar la configuración óptima de la red neuronal que maximice el rendimiento del modelo en la tarea específica en cuestión. Se ejecutaron 1000 iteraciones para explorar diferentes configuraciones de manera aleatoria, y se definió un espacio de búsqueda que incluye los hiperparámetros clave de la red neuronal: número de capas, número de neuronas por capa, funciones de activación, tasa de aprendizaje, entre otros (Tabla 2). Este espacio de búsqueda fue diseñado para explorar una amplia gama de configuraciones posibles. Para cada conjunto de hiperparámetros seleccionado, la red neuronal fue entrenada durante un 5 de épocas para la exploración.

Se utilizó el conjunto de validación para evaluar el rendimiento de cada modelo y seleccionar el mejor conjunto de hiperparámetros. Esta metodología permitió explorar una amplia gama de configuraciones de la red neuronal, incluyendo variaciones en el número de neuronas, la cantidad de capas y las funciones de activación. Estas exploraciones se detallan en la Tabla 2. Las métricas objetivo durante la exploración fueron *sensibilidad (recall), precisión y exactitud*. Al finalizar las búsquedas, se seleccionó el modelo con el mejor rendimiento en el conjunto de validación. Después de identificar el mejor modelo mediante la optimización de hiperparámetros con *Keras Tuner*, el modelo seleccionado fue sometido a un proceso de entrenamiento más exhaustivo para maximizar su rendimiento. El modelo óptimo fue entrenado durante un máximo de 250 épocas. Para evitar el sobreentrenamiento del modelo y asegurar una buena capacidad de generalización, se implementó la técnica de *early stopping*, si la pérdida de validación no mejoraba después de 10 épocas consecutivas, el entrenamiento se detenía automáticamente. Durante el entrenamiento, se aplicó un ajuste dinámico del learning rate basado en el progreso del loss. Este ajuste fue realizado mediante la implementación de un programador de tasa de aprendizaje (learning rate scheduler) que reducía el learning rate cuando la pérdida (loss) en el conjunto de validación dejaba de mejorar durante varias épocas consecutivas(François Chollet, 2015; Kim et al., 2021). Esto ayudó a afinar los pesos del modelo con mayor precisión durante las etapas finales del entrenamiento. Dado que los datos presentaban un desbalance entre las clases, se asignaron pesos asimétricos a las clases utilizando el parámetro class_weight. Este ajuste fue crucial para asegurar que el modelo no se inclinara hacia la predicción de la clase mayoritaria y que prestara la debida atención a las clases minoritarias. Se utilizó la función de pérdida *binary cross entropy* como métrica principal para la optimización del modelo. Las predicciones se compararon con las etiquetas del método directo para determinar un porcentaje de eficacia (Deelder et al., 2019). Con fines de comparación se implementaron tres algoritmos tradicionales de ML, esto fueron *RF*, *SVM* y *RL*. Para cada uno de estos modelos, se emplearon configuraciones estándar.

Todas las exploraciones d las diversas configuraciones de la red, fueron realizadas poniendo un especial énfasis en la sensibilidad (recall) como objetivo principal. Debido al desbalance significativo presente en el conjunto de datos, se decidió aplicar una estrategia para mitigar el impacto de este desbalance en el rendimiento del modelo. Primero, se realizó un ajuste en el

peso de las clases, asignando mayor peso a la clase minoritaria para asegurar que el modelo prestara suficiente atención a esta durante el entrenamiento. Segundo, y se implementó una función de pérdida personalizada basada en *focal loss* propuesta por Lin et al. (2017) (Lin et al., 2017). Ver figura 2.

Metodología para objetivo 3. Validación de los modelos de ML

En esta fase del estudio, se llevó a cabo una evaluación exhaustiva de los modelos de redes neuronales convolucionales (CNN) desarrollados para predecir la farmacorresistencia. El objetivo principal fue determinar la efectividad de estos modelos mediante un análisis detallado de su rendimiento en el conjunto de prueba y compararlo con métodos tradicionales.

Inicialmente, los modelos de CNN fueron sometidos a un proceso de validación cruzada para asegurar la robustez y generalización de los resultados. Posteriormente, los modelos fueron evaluados utilizando el conjunto de prueba, que es independiente del conjunto de entrenamiento y validación, para obtener una estimación precisa de su rendimiento en datos no vistos. Esta evaluación se centró en el cálculo de métricas clave, incluyendo: *Accuracy (Precisión)*: La proporción de predicciones correctas en el conjunto de prueba. *Recall (Sensibilidad)*: La capacidad del modelo para identificar correctamente los casos de farmacorresistencia. *F1-Score*: La media armónica entre *accuracy y recall*, proporcionando un balance entre ambas métricas. *AUC-ROC: El área bajo la curva ROC*, que mide la capacidad del modelo para distinguir entre clases.

Se compararon los resultados obtenidos por los modelos de CNN con el diagnostico fenotípico determinado por métodos tradicionales, permitiendo identificar mejoras o desventajas del enfoque basado en redes neuronales. Se generaron matrices de confusión para cada modelo, las cuales proporcionaron una visión detallada sobre las predicciones correctas e incorrectas, desglosadas por clase. Con base en el análisis de las métricas de desempeño y las visualizaciones generadas, se determinó cuál de los modelos de CNN ofrecía el mejor balance entre *accuracy, recall, y F1-score*, y si superaba o no a los métodos tradicionales en la predicción de la farmacorresistencia.

Metodología para objetivo 4. Predicción de nuevas variantes asociadas a farmacorresistencia.

En el objetivo 4 se pretende Predecir nuevas variantes asociadas a farmacorresistencia en aislados clínicos de *Mycobacterium tuberculosis* a partir de genoma completo utilizando modelos de ML en esta fase, se utilizó el modelo de ML con mejor rendimiento para explorar nuevas variantes genéticas potencialmente asociadas a la resistencia a fármacos en *Mycobacterium tuberculosis*.

El modelo de mejor rendimiento, que resultó ser un RF, para aplicar a los datos genómicos no utilizados en las fases anteriores. Este modelo permitió evaluar la relevancia relativa de cada una de las variables en la predicción de farmacorresistencia. Se seleccionaron aquellas variantes que contribuyeron con más del 1% de la relevancia total del modelo. Posteriormente, se identificó a qué genes pertenecían estas variantes y se agrupó la predicción de resistencia según el fármaco. Finalmente, se compararon las variantes de mayor relevancia relativa con las reportadas en el catálogo de mutaciones de la OMS, para determinar si las predicciones del modelo coincidían con mutaciones conocidas asociadas a la resistencia a fármacos específicos (World Healt Organizartion, 2021).

Infraestructura

Para el almacenamiento de los datos, así como para el procesamiento, se utilizó el servidor del Laboratorio de Epidemiología y Ecología Molecular de la Escuela de Ciencias de la Salud de la UABC, así como dispositivos de almacenamientos del cuerpo académico de Bionanoingeniería de la Facultad de Ingeniería, Arquitectura y Diseño, UABC. El análisis de los datos y el desarrollo del modelo computacional se utilizó un servidor Huawei Atlas 800 (Modelo 910) proporcionado por medio de la convocatoria Alianza UNAM-HUAwei 2022-2023, con apoyo técnico de Dr. Ivan Vladimir Meza Ruiz Instituto de Investigaciones en Matemáticas Aplicadas y Sistemas, y José Fabián Romo Zamudio miembro del Grupo Especial de Innovación de la Alianza para promover el desarrollo de Capacidades Digitales en México (GEI), por parte de la UNAM. Y para la optimización de hiperparámetros se recibió apoyo técnico por parte de los Dr. Héctor Gabriel Acosta Mesa, Dr. Efrén Mezura-Montes y Dr. José Luis Morales Reyes del Instituto de Investigaciones en Inteligencia Artificial de la Universidad Veracruzana.

Resultados

Resultados objetivo 1. Procesamiento y representación de la información genómica

La cantidad final de aislados de *Mycobacterium tuberculosis* La cantidad final de aislados utilizados para la construcción de la matriz de entrenamiento se limitó a 8,743 de los 10,228 genomas disponibles por el consorcio CRYPTIC, porque solo estos presentaban los patrones fenotípicos de DST completos y accesibles para los fármacos seleccionados.

Durante el proceso de selección de características se identificaron un total de 8,438 SNPs ubicados dentro de las regiones genómicas correspondientes a los genes de interés (Tabla 3). El conjunto de datos presentó un desequilibrio importante, donde predominaron los genomas de aislados susceptibles en comparación con una minoría de aislados resistentes a los fármacos de

primera línea (Figura 1). Siendo los más cercanos al equilibro, de mayor a menor, los perfiles de INH, RIF, RFB, EMB, ETH, por lo que estos fueron los fármacos incluidos en los entrenamientos de modelos.

La distribución de mutaciones reveló que los genes con el mayor número de SNPs fueron *PPE35* (385 SNPs), seguido por *rpoB* (363 SNPs) , *gyrA* (362 SNPs), *embB* (361 SNPs), *embA* (359 SNPs), *katG* (356 SNPs), *rpoC* (350 SNPs), *rrl* (346 SNPs) y *ethA* (340 SNPs), y finalmente *gid* (320 SNPs), correspondiendo al 42.02 % de todas las características de la matriz de entrenamiento (tabla 4). ,. La alta presencia de mutaciones en genes como *rpoB* y *katG* permitiendo una cobertura de la variabilidad genética relacionada a la resistencia a RIF y INH, respectivamente. Específicamente, *rpoB* (363 SNPs) es el principal marcador de resistencia a Rifampicina, mientras que *katG* (356 SNPs) y *inhA* (62 SNPs) son críticos para la resistencia a Isoniazida. Asimismo, embB (361 SNPs) está reportado con la resistencia a Etambutol.

Con respecto a los diferentes métodos de codificación de nucleótidos y configuraciones de entrada exploradas, los métodos de codificación one-hot y binaria mostraron resultados consistentemente bajos, con un rendimiento promedio inferior al 50% en todas las métricas de evaluación (*precision, recall, F1-score*). La codificación numérica (con valores 1, 2, 3, y 4 para A, T, C, y G) demostró ser más efectiva, logrando una representación más compacta de los datos y un rendimiento mejor en las métricas de evaluación (figura 2). Al igual que con la codificación one-hot y binaria, las configuraciones de vectores lineales no superaron el umbral del 50% en las métricas clave, probablemente debido a la alta dispersión de los datos. No se observó una mejora con el uso de tensores espaciados; los modelos continuaron presentando un rendimiento bajo debido a la dispersión de los datos. La implementación de una matriz de 2D 92x92 por 1 canal, permitió un espacio de 8464 posiciones para mapear los 8438 SNPs, dejando espacio para 26 posiciones vacías, que durante los entrenamientos fueron manejadas como valores nulos. Esta configuración mostró la mayor compatibilidad con los modelos de ML, permitiendo obtener resultados viables en el entrenamiento. La estructura bidimensional permitió una mejor organización de la información genómica, lo que mejoró el rendimiento en comparación con otras configuraciones.

Resultados objetivo 2. Diseño de modelos de ML

El entrenamiento, la exploración aleatoria y en cuadrícula para todos los modelos se llevaron a cabo en servidores de la Alianza UNAM-Huawei, equipados con unidades de procesamiento neuronal (NPU) *Ascend* 910, se requirió el acceso de octubre 2021 a agosto 2023 para poder realizar todas las etapas del proyecto.

El entorno de trabajo fue desarrollado inicialmente en *TensorFlow* 2 y posteriormente exportado a la plataforma *MindSpore* utilizando *MindStudio*. La implementación de la búsqueda en cuadricula (*grid search*), exploró para cada fármaco 250 configuraciones, donde se seleccionó al azar para la primera un tamaño de filtro de 32 a 160 con un *kernel* de 2 a 3.

El número de capas ocultas tanto convolucionales como densas, para las convolucionales se seleccionaron al azar de 1 a 15 capas, con los mismos límites para filtros y *kernel* que la capa de entrada. Para las capas densas se limitó de 1 a 3 capas con configuraciones de neuronas de 32 a 1042 por capa, las funciones de activación tanto para las convolucionales como para las densas fueron *relu, sigmoid y tanh*. La compilación fue con el optimizador *adam* y *learning rate* variable de 1e-2 a 1e-, enfocando la optimización en el *recall* y *precisión* (tabla 2).

Los resultados obtenidos a partir de la exploración de configuración de red y optimización hiperparámetros realizada con *Keras Tuner* mostraron una mayor complejidad de la estructura y profundidad de las redes neuronales convolucionales (CNN) a medida que aumentaba la asimetría presentada en el conjunto de datos. En particular, se observó que las etiquetas correspondientes a fármacos como RIF y RFB, las cuales mostraron una distribución de sensibilidad y resistencia en una proporción aproximada de 70 % y 30 % respectivamente, condujeron al desarrollo de redes neuronales con una arquitectura que varió entre 2 a 4 capas ocultas. Por otro lado, en el caso de etiquetas como ETH, donde el desbalance era de un 87 % de sensibles frente a un 13 % de resistentes, *Keras Tuner* generó un modelo más complejo, compuesto por 13 capas ocultas, sin embargo, este modelo mostró rendimientos inferiores al 90 % de lo establecido en el objetivo. Los diferentes modelos CNN mostraron rendimientos superiores al 90% tanto en precisión y *recall* para los fármacos RIF, RFB y INH, pero inferiores en ambas métricas para los fármacos EMB y ETH.

Resultados objetivo 3. Validación de los modelos de ML

Para la comparación con los modelos clásicos de *ML* se emplearon tres enfoques diferentes: *RL, RF y SVM*. Para todos los modelos, primero se normalizaron los datos utilizando *StandardScaler*, asegurando que todas las características tuvieran una escala comparable. Los datos se dividieron en conjuntos de entrenamiento y prueba en una proporción de 70/20, reservado un 10% para validación. El modelo de *regresión logística*, entrenado con un máximo de 1000 iteraciones para asegurar la convergencia, alcanzó una precisión con el conjunto validación del 79.75 % a 94.02% en el conjunto de prueba para los diversos fármacos, demostrando una notable capacidad predictiva. Por otro lado, el modelo de *RF*, configurado con 1000 árboles (n_estimators=1000) y entrenado con los mismos datos, logró una precisión del 88.68 % al 94.95 %. Finalmente, el modelo *SVM*, utilizando un kernel *RBF* con gamma='auto' y un parámetro de regularización C=1, mostró una precisión del 85.20 a 89.33% (Tabla 5).

En la predicción de resistencia a la rifampicina, el modelo de *RF* demostró ser el más efectivo, alcanzando una precisión del 94.95%, un recall del 93.23% y un F1 score de 94.02%. Esto sugiere que RF tiene una excelente capacidad para identificar tanto los casos positivos como negativos con alta precisión y equilibrio. La CNN también mostró un rendimiento con métricas muy cercanas a los modelos RF, reflejando su robustez en la captura de patrones complejos en los datos genómicos. El SVM menos preciso (89.33%) debido a que no captó todas las características presentes en el conjunto de ratos, puede ser útil en escenarios donde la simplicidad del modelo

es preferible o se presenta con limitaciones en el poder de cómputo, sin embargo, mostró un menor *recall*, indicando posibles dificultades en la detección de todos los casos positivos. La regresión logística regularizada (RL2) presentó un rendimiento similar tanto en *recall*, *precisión* y *f1-score*, pero no tan alto como el rendimiento de RF y CNN, lo que indica que mientras es eficaz, puede no capturar toda la complejidad de los datos (Tabla 5).

Para Rifabutina, el RF presentó el mejor desempeño con una *precisión* del 94.04%, un *recall* del 92.42% y un *F1 score* de 93.18%. Este modelo parece adaptarse bien a las variaciones presentes en los datos, manejando la alta dimensionalidad de manera eficiente. La CNN también fue altamente efectiva (*precisión* del 89.33%, *recall* del 79.12% y un *F1 score* de 82.18%), sugiriendo que las redes neuronales pueden modelar con éxito la resistencia a RFB a través de la captura de interacciones no lineales en los datos. El SVM mostró una menor *precisión* (88.45%) y un *recall* (78.60%), lo que podría indicar que este enfoque no es tan efectivo para esta clase de resistencia debido a la posible complejidad no lineal que SVM podría no estar capturando completamente con el kernel configurado. RL2, con una precisión del 90.43%, proporciona un rendimiento satisfactorio, pero menos preciso (Tabla 5).

En el caso de la isoniazida, tanto la CNN como el RF mostraron rendimientos similares, con precisiones alrededor del 92%. Esto sugiere que ambos modelos son adecuados para capturar las características genómicas relacionadas con la resistencia a INH. Sin embargo, la CNN tuvo un *recall* ligeramente mayor (90.99%) comparado con RF (90.51%), lo cual puede ser una indicación de que CNN es ligeramente mejor en identificar todos los casos positivos. El SVM mostró una *precisión* más baja (87.36%) y un *recall* (81.40%), reflejando su menor capacidad para manejar la complejidad de los datos genéticos en este contexto. RL2, con métricas también cercanas al 90%, sigue siendo menos efectivo en comparación con RF y CNN en términos de precisión absoluta (Tabla 5).

Para la resistencia al etambutol, el *RF* nuevamente se observó como el mejor modelo con una *precisión* del 93.16%, un *recall* del 88.89% y un *F1 score* de 90.81%. La CNN, con métricas también altas, pero ligeramente menores, se observa su utilidad en la identificación de patrones no lineales complejos. El SVM, mostró un rendimiento inferior con una precisión del 87.81% y un recall del 78.85%, lo que sugiere que puede no estar capturando todas las variaciones genómicas relevantes para la resistencia al EMB. RL2 mostron un rendimiento más equilibrado en las principales métricas (*recall, precisión y f1-score*), pero inferior en todas ellas al rendimiento de CNN y RF (Tabla 5).

La predicción de resistencia a Etionamida presentó un reto mayor para todos los modelos, pero el *RF* destacó con una *precisión* del 88.68%, un *recall* del 75.02% y un F1 score de 79.75%. La CNN mostró una menor *precisión* (79.11%) y *recall* (72.48%), lo que podría reflejar dificultades en la modelación de patrones específicos a esta resistencia con las arquitecturas usadas. El SVM tuvo un rendimiento intermedio con una *precisión* del 85.20% y un *recall* del 73.63%, mostrando cierta eficacia, aunque no tan robusta como RF. Finalmente, RL2, con una *precisión* del 82.89% y un

recall del 78.40%, sugiere que puede no ser tan efectivo en contextos de alta variabilidad genética como ETH (Tabla 5).

.

Resultados objetivo 4. Predicción de nuevas variantes asociadas a farmacorresistencia en aislados clínicos de *Mycobacterium tuberculosis* a partir de secuencias de genoma completo utilizando modelos de ML.

Para la identificación variantes genéticas que contribuyen a la farmacorresistencia, se seleccionaron los modelos de ML *RF*, debido a su rendimiento superior en comparación con otros modelos de *ML*. Una de las ventajas significativas del *RF* es su capacidad para evaluar la relevancia relativa de cada característica analizada, superando las limitaciones de las redes neuronales convolucionales (CNN), que no siempre permiten interpretar fácilmente las características procesadas. En este contexto, las características se refieren a posiciones específicas en el genoma o índices genómicos. Solo se consideraron aquellas características que aportaban al menos un 1% de importancia relativa. Posteriormente, las posiciones genómicas seleccionadas se vincularon con los genes correspondientes para confirmar su implicación en la farmacorresistencia (Figura 3).

En la determinación de farmacorresistencia a la RIF, se identificaron 13 posiciones genómicas que contribuyen significativamente a la predicción, acumulando un 47.59% de la importancia relativa. Las posiciones más relevantes corresponden a varios genes, destacándose las siguientes: rpoB en la posición 761,155 con una importancia del 15.43%, katG en la posición 2,155,168 con un 7.87%, y embB en la posición 4,247,429 con un 3.99%. Adicionalmente, otras posiciones del gen rpoB (761,139, 761,110, y 761,140) aportan un 3.54%, 2.97% y 1.12% respectivamente. Otros genes significativos incluyen rpsL (781687) con un 2.63%, gyrA (7582 y 7570) con importancias del 2.30% y 1.15%, rrs (1473246) con un 1.55%, y rpoC (764817) con un 1.09%. (Figura 7).

En el análisis de farmacorresistencia a la rifabutina (RFB) utilizando el modelo *RF*, se identificaron doce posiciones genómicas que contribuyen significativamente a la predicción, acumulando un 45.43% de la importancia relativa. Las posiciones más destacadas incluyen rpoB en la posición 761155 con una importancia del 17.62%, katG en la posición 2155168 con un 6.46%, y embB en la posición 4247429 con un 3.67%. Además, se observaron otras contribuciones notables de las posiciones *rpoB* 761139 (3.51%) y *rpsl* 781687 (2.97%). Otros genes y sus posiciones relevantes incluyen *embB* (4247431, 2.28% y 4248003, 1.40%), gyrA (7582, 2.27% y 7570, 1.27%), *rpoB* (761110, 1.34%), *rpoC* (764817, 1.34%), y *rrs* (1473246, 1.30%). Estos resultados destacan la

contribución de múltiples loci genómicos a la resistencia a la RIF y RFB, con *rpoB* siendo nuevamente el gen más relevante. (Figura 8).

En el análisis de farmacorresistencia a la isoniazida (INH), se identificaron ocho posiciones genómicas que contribuyen significativamente a la predicción, acumulando un 42.65% de la importancia relativa. La posición más destacada es katG en 2155168, que aporta un 22.30% de la importancia relativa. Otras posiciones importantes incluyen rpoB en 761155 con un 6.78%, rpsL en 781687 con un 4.24%, y embB en 4247429 con un 2.84%. Además, se observaron contribuciones de embB en 4247431 (1.96%), gyrA en 7582 (1.70%), rpoB en 761110 (1.47%), y rpsL en 781822 (1.35%). Se observa la relevancia de varios loci genómicos en la resistencia a la isoniazida, con katG siendo el gen con mayor participación. (Figura 9).

En el análisis de farmacorresistencia a la etionamida (ETH), se identificaron once posiciones genómicas que contribuyen significativamente a la predicción, acumulando un 19.28% de la importancia relativa. Las posiciones más destacadas incluyen rpoB en la posición 761155 con una importancia del 2.70%, gyrA en 7582 con un 2.39%, y rrs en 1473246 con un 1.90%. Además, se identificaron contribuciones relevantes de pncA en 2289090 (1.78%), rpA en 3878547 (1.75%), katG en 2155168 (1.57%), y embB en 4247429 (1.39%). Otras posiciones importantes incluyen rrs en 1472359 (1.29%), rpoC en 766487 (1.19%), embB en 4248003 (1.16%), gyrA en 7570 (1.14%), y rpsL en 781687 (1.02%). Estos resultados subrayan la relevancia de múltiples loci genómicos en la resistencia a la etionamida, con varias posiciones genéticas contribuyendo de manera significativa a la predicción de resistencia, ninguna de forma predominante. (Figura 10).

En el análisis de etambutol (EMB), se identificaron 15 posiciones genómicas que contribuyen significativamente a la predicción, acumulando un 39.51% de la importancia relativa. La posición más destacada es embB en 4247429, que aporta un 7.41% de la importancia relativa. Otros genes relevantes incluyen katG en 2155168 con un 5.56%, rpoB en 761155 con un 5.28%, y gyrA en 7582 con un 2.90%. Adicionalmente, se observaron contribuciones importantes de rpsL en 781687 (2.75%), embB en 4247431 (2.36%) y 4248003 (2.24%), y rpoB en 761110 (2.06%). Otros genes y sus posiciones relevantes incluyen rss en 1473246 (1.78%), gyrA en 7570 (1.41%), rpoC en 764817 (1.35%), pncA en 2289213 (1.20%), embB en 4247469 (1.15%), rrs en 1472359 (1.04%), y rpoA en 3878547 (1.02%). Estos resultados subrayan la importancia de múltiples loci genómicos en la resistencia a la etionamida, destacando a embB como el gen con la mayor contribución relativa. (Figura 11).

Discusión

Estructura y representación de la información genómica para el entrenamiento de modelos.

La prioridad de reducir la incidencia casos de TB farmacorresistente, mediante el diagnóstico oportuno y elección del tratamiento específico, ha impulsado en los últimos años la

implementación de inteligencias artificial para diagnóstico) (Libiseller-Egger et al., 2020; Sharma et al., 2022; Perea-Jacobo et al., 2023). Sin embargo, la complejidad de los datos genómicos de *Mycobacterium tuberculosis* ha requerido que sea necesario trabajar con representaciones de esta información, la cual ha tenido diversas aproximaciones (Aytan-Aktug et al., 2020; Kavvas et al., 2020; Libiseller-Egger et al., 2020; Müller et al., 2021; Zabeti et al., 2021; Zhang, Teng and Alterovitz, 2021; Deelder et al., 2022; Kuang et al., 2022).

En esta investigación se optó por la representación de todas las mutaciones encontradas en los genes asociados con la resistencia a fármacos en TB, considerando tanto aquellas que cuentan con asociación verificada como las que no. El uso de la posición genómica y codificación numérica de cada nucleótido de los genes seleccionadas dentro del conjunto de todos los genomas incluidos en el proyecto, permitió construir una matriz que permite comparar los aislados y considerar como características mutaciones con asociación a farmacorresistencia aún no determinada. Debido a su bajo rendimiento, otras codificaciones no fueron seleccionadas para las etapas posteriores. La codificación numérica demostró ser sencilla y con un bajo nivel de abstracción, lo que permite mantener correspondencia con la estructura biológica de los datos y comparación con los métodos moleculares tradicionales, además facilitó la normalización de las matrices, lo que permitió conservar mayor información relevante en el proceso de entrenamiento. A pesar de que en la literatura se han propuesto otras formas de abstracción que han mostrado alto rendimiento, estas incrementan la complejidad del procesamiento y, por lo tanto, aumentar los requerimientos para la transferencia tecnológica al medio clínico. Además, esta representación ha demostrado tener alta compatibilidad con modelos tradicionales de aprendizaje automático lo que permite identificar características relevantes para los modelos de clasificación) (Chen et al., 2019; Liu et al., 2019; Yang et al., 2019; Kouchaki et al., 2020). La configuración de una matriz 2D mostró la mayor compatibilidad con los modelos de ML, permitiendo obtener resultados viables en el entrenamiento. La combinación de la codificación numérica con la estructura de matriz 2D fue la que proporcionó los mejores resultados en términos de rendimiento y compatibilidad con los algoritmos seleccionados.

El conjunto de entrenamiento de este estudio se centró en las mutaciones de los principales 49 genes con asociación a farmacorresistencia reportada por la OMS (World Healt Organizartion, 2021) con el fin mantener la interpretabilidad de las características a pesar del uso de redes neuronales. La inclusión de genes de Tier 1 y Tier 2 en el conjunto de datos asegura que el modelo desarrollado tenga en cuenta tanto los marcadores genéticos más estudiados y robustos, como aquellos que, aunque menos conocidos, podrían contribuir significativamente a la resistencia a múltiples fármacos. Además de los genes de alta correlación, se incluyeron otros como embA, embC y ubiA (relacionados con EMB) y genes como fgd1, fbiA, y fbiB (asociados con Delamanid, aunque sin alta correlación directa), con el fin de cubrir la mayor diversidad de mutaciones que se pueden encontrar en la población. Se consideraron todos los SNP presentes en las regiones, incluyendo mutaciones, deleciones e inserciones usando posiciones genómicas como índices, de las 8,438 características 30% corresponden a características asociadas a cada uno de los fármacos usados en el estudios, el resto se incluyó debido a el reporte de mutaciones en otros genes

también influyen en la predicción (Kouchaki et al., 2020; Deelder et al., 2022)(. Además, incluir mayor diversidad para el uso del mismo estándar de entrada para próximos modelos enfocados en fármacos de segunda línea.

La representación de la información genómica como una matriz binaria de 96x96 demostró ser altamente efectiva (precisión y recall >90%) para la predicción de farmacorresistencia utilizando modelos CNN. Este enfoque captura adecuadamente las relaciones espaciales entre mutaciones, permitiendo al modelo identificar patrones relevantes con mayor precisión. A pesar de la exploración de otras configuraciones, la simplicidad y eficacia de la representación binaria bidimensional sobresalió.

Dado que los modelos de aprendizaje automático tienden a estar sesgados hacia la clase mayoritaria, el ajuste en el peso de las clases, asignando mayor peso a la clase minoritaria, es el abordaje estándar para asegurar que se prestara suficiente atención a esta durante el entrenamiento. Sin embargo, no se alcanzó los resultados óptimos hasta la implementación de una función de pérdida personalizada basada en la *focal loss* propuesta por Lin et al. (2017). La función *focal loss* es una modificación de la función de pérdida de entropía cruzada (cross entropy) que introduce un factor de ajuste dinámico que da más importancia a los ejemplos mal clasificados, especialmente en el caso de la clase minoritaria, lo cual permitió manejar el desbalance durante el entrenamiento.

Validación de los modelos de *ML*

Las CNN lograron alcanzar objetivos de sensibilidad superiores al 90 % en los dos fármacos más importantes de primera línea (RIF y INH), cifra sugerida por la OMS para modelos de orientación diagnóstica en este ámbito. Similar al reportado en la literatura para modelos convolucionales (Chen et al., 2019; Kuang et al., 2022) . Sin embargo, las métricas reportadas en la literatura se enfocan en F1 score debido a la asimetría de los conjuntos de datos, y el área bajo la curva (Figura 2). Pocos estudios reporta sus sensibilidades, y no abordan un enfoque de entrenamiento siguiendo la sensibilidad (Libiseller-Egger et al., 2020; Sharma et al., 2022; Perea-Jacobo et al., 2023)(Libiseller-Egger et al., 2020; Sharma et al., 2022; Perea-Jacobo et al., 2023).

Por su parte los métodos tradicionales de aprendizaje automático (*RF, SMV, RL2*) mostraron diferencias importantes dependiendo el tipo de modelo. En cuanto a SVM y RL2, para ninguno de los fármacos logro superar a la CNN, observando sensibilidades de un 78.85 % a un 89.55 %. Sin embargo, RF mostró un rendimiento similar a la CNN, superior en precisión y F1-score, pero inferior en sensibilidad.

Entre los modelos se observó que la complejidad del conjunto de datos aumentaba conforme se incrementaba la asimetría de las etiquetas. En el caso de ETH el rendimiento fue el más bajo tanto para los modelos tradicionales como para CNN. Esto concuerda con reportes previos donde se sugiere que esta mayor complejidad y rendimiento se puede deber las mutaciones de resistencia

para este fármaco, como para otros de la familia están dispersos en múltiples genes a diferencia de RIF y INH que las mutaciones se focalizan en una región especifica (Yang et al., 2019; Libiseller-Egger et al., 2020; Deelder et al., 2022; Kuang et al., 2022).

El método de diagnóstico de resistencia mediante secuenciación *Deeplex* se basa en la identificación directa de mutaciones conocidas en 18 genes específicos asociados con resistencia a los fármacos, ofreciendo resultados altamente específicos al detectar mutaciones puntuales previamente catalogadas (Sibandze et al., 2022). En contraste, los modelos desarrollados en este proyecto, incluidos los tres modelos clásicos y la CNN, emplean una matriz más amplia que abarca 49 genes y 8439 SNPs, lo que permite realizar una predicción indirecta de la resistencia. Mientras que *Deeplex* está optimizado para la identificación directa de mutaciones ya conocidas, las CNN desarrolladas tienen la capacidad de integrar y analizar patrones complejos de variación genética, lo que puede facilitar la detección de nuevas asociaciones genéticas no necesariamente relacionadas con mutaciones previamente catalogadas.

Sin embargo, esta capacidad de predicción más generalizada en la CNN podría resultar en mayores tasas de falsos positivos, un factor que debe ser cuidadosamente considerado en aplicaciones clínicas. Aunque la sensibilidad de la CNN sugiere un rendimiento robusto, la especificidad y precisión pueden no ser equivalentes a las de herramientas como Deeplex, que están específicamente diseñadas para la detección directa de mutaciones conocidas. Esta comparación destaca la relevancia de utilizar tanto enfoques directos como indirectos en la evaluación de la resistencia, dependiendo de los objetivos clínicos y del tipo de datos disponibles.

Limitaciones del estudio y consideraciones futuras.

El uso de una matriz de características enfocada en todos los genes permite valorar todos los fármacos usados en el área clínica, sin embargo, al momento solo se ha incluido medicamentos de primera línea. Esto hace que el 70% de características de la matriz este sub-aprovechada. La mayoría de los estudios prueban de 11 a 13 medicamentos, indicando que los rendimientos son mejores para los fármacos más estudiados y muestreados, bajando el rendimiento en los fármacos de segunda línea menos estudiados, sin embargo, al no incluirlos en el estudio no se observa si este efecto se replica en estos modelos. Además, la reducción de dimensiones y la determinación de importancia relativa pueden ser usadas para reducir la matriz de entrenamiento y posiblemente mejorar el rendimiento.

Conclusiones

El desafío de representar de manera precisa y biológicamente relevante la información genómica es crucial para el rendimiento de los modelos de aprendizaje automático en el diagnóstico de la farmacorresistencia de la tuberculosis. Como se ha demostrado en este estudio, la elección de la representación de las mutaciones genéticas impacta significativamente en la capacidad predictiva

de los modelos, especialmente al utilizar enfoques complejos como las redes neuronales convolucionales (CNN). El enfoque para la construcción de los datos de entrenamiento centrados en los genes principales muestra un buen rendimiento tanto para la CNN como para los modelos tradicionales. Al igual que otras investigaciones, los modelos tradicionales presentan un rendimiento similar a las CNN, sin requerir abstracción del conjunto de características, permitiendo mantener su interpretabilidad biológica. Aunque las CNN pueden manejar patrones genéticos complejos y ofrecer la posibilidad de descubrir nuevas asociaciones, su rendimiento no siempre supera al de los métodos de ML clásicos cuando se emplean representaciones de menor complejidad. El incremento complejidad de los modelos de *Deep learning*, y requisitos de poder de cómputo, no parece justificable debido al buen rendimiento de modelos como RF y SVM (Chen et al., 2019; Kouchaki et al., 2020; Yang et al., 2019).

Debido a la naturaleza de los mecanismos biológicos que confieren resistencia a fármacos en TB, se han explorado la inclusión de los SNP en todos los genes de *Mycobacterium tuberculosis*, así como regiones intergénicas. Pero esto incrementa tremendamente las dimensiones, la matriz de características y el poder de cómputo necesario. Debido a que recientemente la OMS, ha revisado y publicado un catálogo con más de 17,000 mutaciones donde en< 1% se ha verificado su participación en conferir resistencia a las cepas, se observa que focalizar la construcción de la matriz de características en los genes principales reduce el poder de cómputo necesario manteniendo un rendimiento aceptable.

La diversidad de enfoques para la representación genómica en la literatura científica y en las diversas bases de datos donde los genomas son depositados ha dificultado la estandarización de metodologías, lo que a su vez complica la comparación directa y la replicabilidad de los modelos desarrollados. Este estudio subraya la necesidad de avanzar hacia la creación de estándares que permitan una evaluación más coherente y reproducible del rendimiento de los modelos, considerando tanto la precisión en la representación biológica como la complejidad computacional.

El mantenimiento de la interpretabilidad biológica de la matriz de características y la valoración del rendimiento en términos de sensibilidad y especificidad permiten la comparación con los estándares clínicos, para diseño de futuros modelos e implementaciones de transferencia de estas técnicas al campo clínico. Sin embargo, la simplificación en la representación, aunque facilita el desarrollo y la transferencia tecnológica al entorno clínico, puede limitar el potencial de modelos más avanzados como las CNN, haciendo que sus ventajas sobre los métodos clásicos sean menos evidentes. Este hallazgo resalta la importancia de seleccionar la representación de datos de manera estratégica, en función del equilibrio entre la complejidad del modelo y la necesidad de precisión en la identificación de mutaciones críticas.

Referencias

Ahmad, S. et al. (2016) 'Discordance across phenotypic and molecular methods for drug susceptibility testing of drug-resistant Mycobacterium tuberculosis isolates in a low TB incidence country', PLoS ONE, 11(4). Available at: https://doi.org/10.1371/journal.pone.0153563.

Alaridah, N. et al. (2019) 'Transmission dynamics study of tuberculosis isolates with whole genome sequencing in southern Sweden', Scientific Reports 2019 9:1, 9(1), pp. 1–9. Available at: https://doi.org/10.1038/s41598-019-39971-z.

Alsuliman, T., Humaidan, D. and Sliman, L. (2020) 'ML and artificial intelligence in the service of medicine: Necessity or potentiality?', Current research in translational medicine, 68(4), pp. 245–251. Available at: https://doi.org/10.1016/J.RETRAM.2020.01.002.

Aytan-Aktug, D. et al. (2020) 'Prediction of Acquired Antimicrobial Resistance for Multiple Bacterial Species Using Neural Networks', mSystems, 5(1). Available at: https://doi.org/10.1128/msystems.00774-19.

Brandao, A.P. et al. (2020) 'Transmission of Mycobacterium tuberculosis presenting unusually high discordance between genotypic and phenotypic resistance to rifampicin in an endemic tuberculosis setting', Tuberculosis, 125. Available at: https://doi.org/10.1016/j.tube.2020.102004.

Branigan, D. et al. (2021) 'Pipeline Report » 2021 Tuberculosis Diagnostics Tuberculosis Diagnostics'.

Chen, M.L. et al. (2019) 'Beyond multidrug resistance: Leveraging rare variants with machine and statistical learning models in Mycobacterium tuberculosis resistance prediction', EBioMedicine, 43, pp. 356–369. Available at: https://doi.org/10.1016/j.ebiom.2019.04.016.

Chowdhury, A.S., Khaledian, E. and Broschat, S.L. (2019) 'Capreomycin resistance prediction in two species of Mycobacterium using a stacked ensemble method', Journal of Applied Microbiology, 127(6), pp. 1656–1664. Available at: https://doi.org/10.1111/jam.14413.

Coll, F. et al. (2018) 'Genome-wide analysis of multi- and extensively drug-resistant Mycobacterium tuberculosis', Nature genetics, 50(2), pp. 307–316. Available at: https://doi.org/10.1038/S41588-017-0029-0.

Colman, R.E. et al. (2015) ‘Detection of Low-Level Mixed-Population Drug Resistance in Mycobacterium tuberculosis Using High Fidelity Amplicon Sequencing’, PLOS ONE, 10(5), p. e0126626. Available at: https://doi.org/10.1371/JOURNAL.PONE.0126626.

Colman, R.E. et al. (2019) ‘Whole-genome and targeted sequencing of drug-resistant Mycobacterium tuberculosis on the iSeq100 and MiSeq: A performance, ease-of-use, and cost evaluation’, PLoS medicine, 16(4). Available at: https://doi.org/10.1371/JOURNAL.PMED.1002794.

Comas, I. and Gil, A. (2016) ‘Secuenciación masiva para el diagnóstico y la epidemiología de tuberculosis’, Enfermedades Infecciosas y Microbiología Clínica, 34, pp. 32–39. Available at: https://doi.org/10.1016/S0213-005X(16)30217-8.

Cuevas-Córdoba, B. and Zenteno-Cuevas, R. (2010) ‘Tuberculosis drogorresistente: mecanismos moleculares y métodos diagnósticos’, Enfermedades Infecciosas y Microbiología Clínica, 28(9), pp. 621–628. Available at: https://doi.org/10.1016/J.EIMC.2009.12.005.

David Branigan (2020) An Activist’s Guide to Tuberculosis Diagnostic Tools – Treatment Action Group. Available at: https://www.treatmentactiongroup.org/publication/an-activists-guide-to-tuberculosis-diagnostic-tools/ (Accessed: 14 April 2024).

Deelder, W. et al. (2019) ‘ML predicts accurately mycobacterium tuberculosis drug resistance from whole genome sequencing data’, Frontiers in Genetics, 10(SEP). Available at: https://doi.org/10.3389/fgene.2019.00922.

Deelder, W. et al. (2022) ‘A modified decision tree approach to improve the prediction and mutation discovery for drug resistance in Mycobacterium tuberculosis’, BMC Genomics, 23(1). Available at: https://doi.org/10.1186/s12864-022-08291-4.

Dolinger, D.L. et al. (2016) ‘Next-generation sequencing-based user-friendly platforms for drug-resistant tuberculosis diagnosis: A promise for the near future’, International journal of mycobacteriology, 5 Suppl 1, pp. S27–S28. Available at: https://doi.org/10.1016/J.IJMYCO.2016.09.021.

Doya, K. (2007) ‘Reinforcement learning: Computational theory and biological mechanisms’, HFSP journal, 1(1), pp. 30–40. Available at: https://doi.org/10.2976/1.2732246/10.2976/1.

Duffy, F.J. et al. (2019) ‘Multinomial modelling of TB/HIV co-infection yields a robust predictive signature and generates hypotheses about the HIV+TB+ disease state’, PLoS ONE, 14(7), pp. 1–17. Available at: https://doi.org/10.1371/journal.pone.0219322.

Farhat, M.R. et al. (2016) ‘Genetic determinants of drug resistance in mycobacterium tuberculosis and their diagnostic value’, American Journal of Respiratory and Critical Care Medicine, 194(5), pp. 621–630. Available at: https://doi.org/10.1164/rccm.201510-2091OC.

De Fauw, J. et al. (2018) ‘Clinically applicable deep learning for diagnosis and referral in retinal disease’, Nature medicine, 24(9), pp. 1342–1350. Available at: https://doi.org/10.1038/S41591-018-0107-6.

Friedman, J., Hastie, T. and Tibshirani, R. (2008) ‘Sparse inverse covariance estimation with the graphical lasso’, Biostatistics (Oxford, England), 9(3), pp. 432–441. Available at: https://doi.org/10.1093/BIOSTATISTICS/KXM045.

‘From clinical sample to drug resistance profile Deeplex ® Myc-TB USER MANUAL’ (2023).

Gygli, S.M. et al. (2017) ‘Antimicrobial resistance in Mycobacterium tuberculosis: mechanistic and evolutionary perspectives’, FEMS microbiology reviews, 41(3), pp. 354–373. Available at: https://doi.org/10.1093/FEMSRE/FUX011.

Haraka, F. et al. (2021) ‘Impact of the diagnostic test Xpert MTB/RIF on patient outcomes for tuberculosis’, The Cochrane Database of Systematic Reviews, 2021(5). Available at: https://doi.org/10.1002/14651858.CD012972.PUB2.

Heidema, A.G. et al. (2006) ‘The challenge for genetic epidemiologists: how to analyze large numbers of SNPs in relation to complex diseases’, BMC Genetics, 7(23), p. 23. Available at: https://doi.org/10.1186/1471-2156-7-23.

Jamal, S. et al. (2020) 'Artificial Intelligence and ML based prediction of resistant and susceptible mutations in Mycobacterium tuberculosis', Scientific Reports, 10(1), pp. 1–16. Available at: https://doi.org/10.1038/s41598-020-62368-2.

Jeanes, C. and O'Grady, J. (2016) 'Diagnosing tuberculosis in the 21st century - Dawn of a genomics revolution?', International journal of mycobacteriology, 5(4), pp. 384–391. Available at: https://doi.org/10.1016/J.IJMYCO.2016.11.028.

Kang, J.Y. et al. (2019) 'Clinical implications of discrepant results between genotypic MTBDRplus and phenotypic Löwenstein-Jensen method for isoniazid or rifampicin drug susceptibility tests in tuberculosis patients', Journal of Thoracic Disease, 11(2), pp. 400–409. Available at: https://doi.org/10.21037/jtd.2019.01.58.

Kavvas, E.S. et al. (2020) 'A biochemically-interpretable ML classifier for microbial GWAS', Nature Communications, 11(1). Available at: https://doi.org/10.1038/S41467-020-16310-9.

Kazumi, Y. and Mitarai, S. (2012) 'The evaluation of an identification algorithm for Mycobacterium species using the 16S rRNA coding gene and rpoB', International journal of mycobacteriology, 1(1), pp. 21–28. Available at: https://doi.org/10.1016/J.IJMYCO.2012.01.004.

Kononenko, I. (2001) 'ML for medical diagnosis: History, state of the art and perspective', Artificial Intelligence in Medicine, 23(1), pp. 89–109. Available at: https://doi.org/10.1016/S0933-3657(01)00077-X.

Kouchaki, S. et al. (2019) 'Application of ML techniques to tuberculosis drug resistance analysis', Bioinformatics, 35(13). Available at: https://doi.org/10.1093/bioinformatics/bty949.

Kouchaki, S. et al. (2020) 'Multi-Label Random Forest Model for Tuberculosis Drug Resistance Classification and Mutation Ranking', Frontiers in Microbiology. Available at: https://doi.org/10.3389/fmicb.2020.00667.

Kuang, X. et al. (2022) 'Accurate and rapid prediction of tuberculosis drug resistance from genome sequence data using traditional ML algorithms and CNN', Scientific Reports, 12(1). Available at: https://doi.org/10.1038/S41598-022-06449-4.

Kulkarni, S. and Jha, S. (2020) 'Artificial Intelligence, Radiology, and Tuberculosis: A Review', Academic radiology. Acad Radiol, pp. 71–75. Available at: https://doi.org/10.1016/J.ACRA.2019.10.003.

Kumar Nathella, P. and Babu, S. (2017) 'Influence of diabetes mellitus on immunity to human tuberculosis', Immunology, 152(1), pp. 13–24. Available at: https://doi.org/10.1111/imm.12762.

Libbrecht, M.W. and Noble, W.S. (2015) 'ML applications in genetics and genomics', Nature reviews. Genetics, 16(6), pp. 321–332. Available at: https://doi.org/10.1038/NRG3920.

Libiseller-Egger, J. et al. (2020) 'Robust detection of point mutations involved in multidrug-resistant Mycobacterium tuberculosis in the presence of co-occurrent resistance markers', PLOS Computational Biology, 16(12), p. e1008518. Available at: https://doi.org/10.1371/JOURNAL.PCBI.1008518.

Lin, T.Y. et al. (2017) 'Focal Loss for Dense Object Detection', IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(2), pp. 318–327. Available at: https://doi.org/10.1109/TPAMI.2018.2858826.

Liu, Y. et al. (2019) 'Phenotype Prediction and Genome-Wide Association Study Using Deep Convolutional Neural Network of Soybean', Frontiers in Genetics, 10(November), pp. 1–10. Available at: https://doi.org/10.3389/fgene.2019.01091.

Lozano, J.A. (2002) 'Tuberculosis. Patogenia, diagnóstico y tratamiento', Offarm, 21(8), pp. 102–110. Available at: https://www.elsevier.es/es-revista-offarm-4-articulo-tuberculosis-patogenia-diagnostico-tratamiento-13035870 (Accessed: 13 April 2024).

Lunetta, K.L. et al. (2004) 'Screening large-scale association study data: Exploiting interactions using random forests', BMC Genetics, 5(1), pp. 1–13. Available at: https://doi.org/10.1186/1471-2156-5-32/FIGURES/5.

Madrazo-Moya, C.F. et al. (2019) 'Whole genomic sequencing as a tool for diagnosis of drug and multidrug-resistance tuberculosis in an endemic region in Mexico', PLoS ONE, 14(6). Available at: https://doi.org/10.1371/journal.pone.0213046.

Miotto, R. et al. (2018) ‘Deep learning for healthcare: review, opportunities and challenges’, Briefings in bioinformatics, 19(6), pp. 1236–1246. Available at: https://doi.org/10.1093/BIB/BBX044.

Müller, S.J. et al. (2021) ‘First-line drug resistance profiling of Mycobacterium tuberculosis: a ML approach’, AMIA Annual Symposium Proceedings, 2021, p. 891. Available at: /pmc/articles/PMC8861754/ (Accessed: 9 May 2023).

Murray, C.J. et al. (2022) ‘Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis’, The Lancet, 399(10325), pp. 629–655. Available at: https://doi.org/10.1016/S0140-6736(21)02724-0.

Navasardyan, I. et al. (2024) ‘HIV-TB Coinfection: Current Therapeutic Approaches and Drug Interactions’, Viruses, 16(3), p. 321. Available at: https://doi.org/10.3390/V16030321.

Panch, T., Szolovits, P. and Atun, R. (2018) ‘Artificial intelligence, ML and health systems’, Journal of global health, 8(2). Available at: https://doi.org/10.7189/JOGH.08.020303.

Perea-Jacobo, R. et al. (2019) ‘Rifampin pharmacokinetics in tuberculosis-diabetes mellitus patients: A pilot study from Baja California, Mexico’, International Journal of Tuberculosis and Lung Disease, 23(9), pp. 1012–1016. Available at: https://doi.org/10.5588/ijtld.18.0739.

Perea-Jacobo, R. et al. (2023) ‘ML of the Whole Genome Sequence of Mycobacterium tuberculosis: A Scoping PRISMA-Based Review’, Microorganisms 2023, Vol. 11, Page 1872, 11(8), p. 1872. Available at: https://doi.org/10.3390/MICROORGANISMS11081872.

Roy, S. et al. (2018) ‘Standards and Guidelines for Validating Next-Generation Sequencing Bioinformatics Pipelines: A Joint Recommendation of the Association for Molecular Pathology and the College of American Pathologists’, Journal of Molecular Diagnostics, 20(1), pp. 4–27. Available at: https://doi.org/10.1016/J.JMOLDX.2017.11.003.

Safi, H. et al. (2013) ‘Evolution of high-level ethambutol-resistant tuberculosis through interacting mutations in decaprenylphosphoryl-β-D-arabinose biosynthetic and utilization pathway genes’, Nature genetics, 45(10), pp. 1190–1197. Available at: https://doi.org/10.1038/NG.2743.

Satta, G. et al. (2018) 'Mycobacterium tuberculosis and whole-genome sequencing: how close are we to unleashing its full potential?', Clinical Microbiology and Infection. Elsevier B.V., pp. 604–609. Available at: https://doi.org/10.1016/j.cmi.2017.10.030.

Sharma, A. et al. (2022) 'Tuberculosis drug resistance profiling based on ML: A literature review', The Brazilian Journal of Infectious Diseases, 26(1), p. 102332. Available at: https://doi.org/10.1016/J.BJID.2022.102332.

Sibandze, D.B. et al. (2022) 'Rapid molecular diagnostics of tuberculosis resistance by targeted stool sequencing', Genome medicine, 14(1). Available at: https://doi.org/10.1186/S13073-022-01054-6.

Stop TB Partnership, U. and the G.L.Initiative. (2021) 'Practical Guide to Implementation of Truenat Tests for the Detection of TB and Rifampicin Resistance Practical Guide to Implementation of Truenat TM Tests for the Detection of TB and Rifampicin Resistance', Geneva: Stop TB Partnership [Preprint].

Trauner, A. et al. (2014) 'Evolution of drug resistance in tuberculosis: recent progress and implications for diagnosis and therapy', Drugs, 74(10), pp. 1063–1072. Available at: https://doi.org/10.1007/S40265-014-0248-Y.

Walker, T.M. et al. (2013) 'Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: A retrospective observational study', The Lancet Infectious Diseases, 13(2), pp. 137–146. Available at: https://doi.org/10.1016/S1473-3099(12)70277-3.

Witten, D.M., Tibshirani, R. and Hastie, T. (2009) 'A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis', Biostatistics (Oxford, England), 10(3), pp. 515–534. Available at: https://doi.org/10.1093/BIOSTATISTICS/KXP008.

World Healt Organizartion (2021) Catalogue of mutations in Mycobacterium tuberculosis complex and their association with drug resistance.

World Health Organization (2009) Approaches to improve sputum smear microscopy for tuberculosis diagnosis expert group meeting report, 2009. Available at:

https://stoptb.org/wg/gli/assets/documents/EGM%20Report%20on%20Microscopy%20Methods%20FINAL%20November%202009.pdf (Accessed: 14 April 2024).

World Health Organization (2014) High-priority target product profi les for new tuberculosis diagnostics: report of a consensus meeting. Available at: www.who.int (Accessed: 14 April 2024).

World Health Organization (2015) Global strategy and targets for tuberculosis prevention, care and control after 2015. Available at: https://iris.who.int/handle/10665/162760 (Accessed: 13 April 2024).

World Health Organization (2018a) Technical manual for drug susceptibility testing of medicines used in the treatment of tuberculosis. Available at: https://www.who.int/publications/i/item/9789241514842 (Accessed: 15 April 2024).

World Health Organization (2018b) The use of next-generation sequencing technologies for the detection of mutations associated with drug resistance in Mycobacterium tuberculosis complex: technical guide. Available at: https://www.who.int/publications/i/item/WHO-CDS-TB-2018.19 (Accessed: 16 April 2024).

World Health Organization (2021a) GLOBAL TUBERCULOSIS REPORT 2021. Available at: http://apps.who.int/bookorders.

World Health Organization (2021b) WHO consolidated guidelines on tuberculosis: module 2: screening: systematic screening for tuberculosis disease. Available at: https://www.who.int/publications/i/item/9789240022676 (Accessed: 14 April 2024).

World Health Organization (2023a) 'Global Tuberculosis Report 2023', World Health Organization [Preprint].

World Health Organization (2023b) Use of targeted next-generation sequencingto detect drug-resistant tuberculosis. Available at: https://www.who.int/publications/i/item/9789240076372 (Accessed: 10 April 2024).

Yang, Y. et al. (2018) ‘ML for classifying tuberculosis drug-resistance from DNA sequencing data’, Bioinformatics, 34(10), pp. 1666–1671. Available at: https://doi.org/10.1093/bioinformatics/btx801.

Yang, Y. et al. (2019) ‘DeepAMR for predicting co-occurrent resistance of Mycobacterium tuberculosis’, Bioinformatics, 35(18), pp. 3240–3249. Available at: https://doi.org/10.1093/bioinformatics/btz067.

Zabeti, H. et al. (2021) ‘INGOT-DR: an interpretable classifier for predicting drug resistance in *Mycobacterium tuberculosis*’, Algorithms for Molecular Biology, 16(1). Available at: https://doi.org/10.1186/s13015-021-00198-1.

Zhang, A., Teng, L. and Alterovitz, G. (2021) ‘An explainable ML platform for pyrazinamide resistance prediction and genetic feature identification of Mycobacterium tuberculosis’, Journal of the American Medical Informatics Association : JAMIA, 28(3), pp. 533–540. Available at: https://doi.org/10.1093/JAMIA/OCAA233.