Uso de *machine learning* como apoyo al diagnóstico del complejo *Mycobacterium tuberculosis*: Una revisión sistemática

M. A. Guerrero-Chevannier[1], R. Perea-Jacobo[1,2], D. L. Flores[1], R. Muñiz-Salazar[2*]

[1] Facultad de Ingeniería, Arquitectura y Diseño, Universidad Autónoma de Baja California, México.
[2] Escuela de Ciencias de la Salud, Universidad Autónoma de Baja California, México.
* *ramusal@uabc.edu.mx*

Resumen

La implementación de modelos de machine learning (ML) como apoyo al diagnóstico médico permite facilitar el análisis y disminuir errores en el diagnóstico. En este artículo se realiza una revisión sistemática de la eficacia y precisión de los modelos de ML implementados actualmente. Se implementó la metodología PRISMA para la revisión de artículos, por medio de la plataforma www.covidence.org. Se realizó la búsqueda de artículos que utilizaron modelos de ML aplicados al diagnóstico radiológico y al análisis de genoma completo de *M. tuberculosis*, y que hayan sido publicados en los últimos cinco años. La búsqueda se realizó en las plataformas de Medline/Pubmed, Embase y ScienceDirect. De un total de 135 artículos compatibles con los criterios de búsqueda, sólo 18 artículos cumplieron con los criterios y objetivos de la revisión. Los artículos se organizaron de acuerdo a su aplicación, 1) detección de farmacorresistencia en genoma completo y 2) diagnóstico radiológico. El método más utilizado fue *support vector machine* (SVM) con una exactitud variable de 73% hasta 93.89%, *artificial neural networks* (ANN) presentó la exactitud más alta de 100% pero solo para un gen específico. Con el creciente aumento del registro y disponibilidad de datos clínicos, se requiere implementar protocolos de análisis eficaces. Los modelos de ML son efectivos e incluso en algunos casos superiores a los métodos de asociación directa. Los modelos presentan mejor rendimiento al integrarse varios en el mismo protocolo.

Palabras clave: Diagnóstico médico, Farmacoresistencia, Machine learning, Prisma, Tuberculosis.

1. Introducción

La tuberculosis (TB) es una enfermedad infecciosa que presenta grandes retos en la salud pública y el tratamiento adecuado, depende de un diagnóstico rápido y preciso. Es una enfermedad bacteriana, generalmente respiratoria, con transmisión persona a persona ocasionada el Complejo Micobacterium tuberculosis el cual incluye múltiples cepas relacionadas. En la mayoría de los casos, la TB es tratable y curable; sin embargo, las personas con TB pueden morir si no reciben el tratamiento adecuado. 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 de la baciloscopia para diagnosticar TB, prueba que viene utilizándose desde hace más de 100 años [1]. La baciloscopia permite detectar la presencia de la bacteria en los pacientes, pero no caracteriza la cepa.

La identificación rápida de cepas resistentes o susceptibles a ciertos fármacos es esencial para el tratamiento adecuado evitando complicaciones y reduciendo significativamente la duración del

dx.doi.org/10.24254/CNIB.20.54

tratamiento. El cultivo bacteriano y la prueba de sensibilidad a drogas (PSD) son los métodos clásicos para determinar la resistencia bacteriana, estos requieren de 2 a 4 semanas para crecimiento y entrega resultados. Estas pruebas se deben de realizar en laboratorios con bioseguridad nivel 3, los cuales solo están disponibles en centros de referencia regionales, incrementando el tiempo de espera de resultados. Debido a esto la mayoría de los pacientes inicia tratamiento primario de 6 meses sin pruebas de sensibilidad, la OMS estima que solo el 51% de los pacientes a nivel mundial son examinados para resistencia farmacológica [1].

En 2018 se han reportado a nivel mundial 186 772 casos de TB Multidrogo resistente/Resistente a Rifampicina (MDR/RR-TB) un incremento del 13.7 % con respecto al año previo [1]. El tratamiento de MDR/RR-TB requiere de utilización de fármacos de segunda línea por un periodo de 9 a 20 meses, incrementando las complicaciones por consumo de fármacos y disminuyendo el apego al tratamiento. Las pruebas moleculares rápidas, los datos clínicos y las tecnologías de secuenciación se encuentran disponibles en cuestión de días, un tiempo mucho menor a los métodos basados en cultivo.

Con el aumento de la disponibilidad de las nuevas tecnologías de inteligencia artificial, en particular machine learning y deep learning brindan una posibilidad de abordaje a estas complejas bases de datos clínicos, imágenes radiológicas y genoma completos, con el fin de realizar detección y clasificación rápida la enfermedad, para apoyar en la toma de decisiones clínicas y contribuir al diagnóstico rápido y oportuno.

En el presente trabajo se implementa una revisión sistemática bajo la metodología PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses) [2]. , la cual presenta un conjunto de elementos basados en evaluación de intervenciones. Prisma consta de una lista de verificación de 27 elementos y un diagrama de flujo de cuatro fases útiles para mejorar la presentación de informes de revisiones sistemática y metaanálisis.

2. Metodología

Se revisó sistemáticamente la literatura para describir la utilidad de los aportes más nuevos de los métodos computarizados de machine learning en el diagnóstico integral del complejo Mycobacterium Tuberculosis. Se enfocó la revisión en las ventajas que aporta el método para el diagnóstico, el modelo más utilizado y las limitaciones para su implementación. Para la revisión de artículos se implementó la metodología Prisma [2].

2.1 Criterios de elegibilidad

Los estudios se seleccionaron de acuerdo con los criterios que se detallan en la Tabla 1.

Tabla 1: Criterios de elegibilidad para inclusión en la revisión sistemática.

Criterio	Descripción
Año de publicación	2015 a 2020
Diseño de estudio	Analíticos
Método de comparación	Método computacional, análisis estadístico
Resultados	Exactitud de la predicción

2.2 Estrategias de búsqueda

Las estrategias de búsqueda de literatura se desarrollaron utilizando Encabezados de Temas Médicos (MeSH) y palabras de texto relacionadas con machine learning y tuberculosis. Se buscó en las bases de datos de Medline/Pubmed (National Center for Biotechnology Information, s.f.) y ScienceDirect (Elsevier, s.f.), las búsquedas fueron limitadas al idioma inglés. Se revisaron las listas de referencias de los estudios incluidos o las revisiones relevantes identificadas mediante la búsqueda. También se buscó en los perfiles personales de los autores para asegurar que se haya capturado todo el material relevante. Las palabras de búsqueda fueron las que se muestran en la Tabla 2.

Tabla 2: Palabras de búsqueda

Palabras de búsqueda
"Machine learning"
"Deep learning"
"Diagnostic"
"Tuberculosis"
"Clinic"
"X- Ray"
"Genome"
"Resistance"

2.3 Registro de los estudios

Los resultados de la búsqueda de literatura se cargaron en un archivo colaborativo, se anexaron los criterios de elegibilidad y resúmenes de citas a una base de datos. Se integró el uso de la plataforma Covidence para asistencia en la elaboración de revisiones sistemáticas. Se examinaron de forma independiente los títulos y resúmenes producidos por la búsqueda según los criterios de elegibilidad. Se obtuvieron informes completos de todos los títulos que parecen cumplir con los criterios de elegibilidad o donde existe alguna incertidumbre. Luego se examinaron los textos completos y se verifico que cumplen con los criterios de elegibilidad. Se buscó información adicional de los autores del estudio cuando fuera necesario para resolver preguntas sobre la elegibilidad. Se resolvieron los desacuerdos a través de la discusión. Se registraron los motivos para excluir los ensayos.

Utilizando formularios estandarizados para la extracción de datos en línea (Covidence). Se resolvieron los desacuerdos mediante consenso. En cada artículo se identificaron los siguientes criterios para su análisis: tipo de datos, disponibilidad de los datos, número de observaciones incluidas, método de análisis, método de validación, asociación de resultados con métodos directos (cultivo, PSD, prueba rápida). De cada estudio incluido se registró la métrica de desempeño de la predicción con respecto al valor medido directamente, se tomó ésta como el indicador de confianza de cada uno de los métodos utilizados.

3. Resultados y Discusión

Se encontraron 135 artículos en las bases de datos Medline/Pubmed y ScienceDirect, de acuerdo con los criterios de búsqueda. De estos 46 fueron eliminados por encontrarse repetidos, 47 fueron eliminados por no presentar de criterios de elegibilidad, 23 más fueron eliminados por no presentar una

metodología u objetivos de acuerdo con esta investigación. En total, se incluyeron 18 artículos para su análisis.

La síntesis de resultados de los estudios se realizó en dos categorías: 1) estudios diagnósticos por imágenes y 2) estudios de determinación de farmacorresistencia por medio de genoma. Los criterios para la síntesis de cada una de estas áreas se basaron en la descripción del método de machine learning utilizado, su asociación a un método directo y su correspondencia al diagnóstico clínico.

3.1 Imágenes de radiografías de tórax

Para el análisis de los estudios de clasificación de imágenes de radiografías de tórax se agrupan como; descripción conjunto de datos, donde se especifica la cantidad y origen de las imágenes de rayos X, así como datos demográficos y datos clínicos; implementación del modelo de machine learning, donde se proponen distintos modelos y arquitecturas, así como las fases de entrenamiento, prueba y validación; exactitud de las predicciones, donde se evalúan y comparan los resultados de los modelos con respecto a su rendimiento.

En la referencia [3]. utilizó los conjuntos de datos Luzaka en conjunto con Zambia (917 pacientes) con una estructura SVM Multiple Instance Learning + Active Learning, obtenido resultados de 0.874 de área bajo la curva característica operativa del receptor (AUROC)[3]. En la referencia [4]. utilizó el conjunto de datos Kampala (138) en distintos experimentos para probar el desempeño de las redes VGG19, InceptionV3, ResNet50, DenseNet121, InceptionResNetV2, obteniendo un resultado de 0.9213, 0.9045, 0.08955, 0.8893, 0.8864 de AUROC respectivamente [4]. En la referencia [5]. utilizó los conjuntos de datos Mongomery (138), Shenzhen (662) y Belarús (304) con una CNN personalizada, obteniendo resultados de 0.925 AUROC [5]. En la referencia [6]. utilizó el conjunto de datos ChestX-ray8 database (874); una base de datos que incluye además de radiografías pulmonares, etiquetas para 14 patrones radiológicos anormales, y la red Quore. AI para la búsqueda de patrones radiológicos consistentes con TB, obteniendo el mejor resultado de 0.929 AUROC para dichos patrones.

En la referencia [7]. se utilizó los dataset Shenzhen (138), Mongomery Country (662) con la red DenseNet121 obteniendo un resultado de 0.937 AUROC. En la referencia [8]. utilizó los conjuntos de datos Montgomery (138), Shenzhen (662), Kenya (967) e India (306) en dos experimentos, el primero donde utilizó una red SVM HOG, GIS, SURF y el segundo donde probó las redes AlexNet, VGG16, GoogLeNet, ResNet50, obteniendo el mejor resultado con la red AlexNet de 0.95 AUROC, finalmente obtuvo un rendimiento de 0.965 AUROC con un ensamble de las redes de ambos experimentos. En la referencia [9]. se empleó las bases de datos de Montgomery (138), Shenzhen (662), Belarus (88), y Thomas Jefferson (119), para probar el rendimiento de las redes AlexNet y GoogLeNet, obteniendo como resultado un 0.99 AUROC utilizando un ensamble de ambas redes. En la referencia [10]. sé utilizó los conjuntos de datos Pediatric P (5856), RSNA (5856), Indiana (4104) y Shenzhen (662), para probar el rendimiento de las redes cCNN, VGG16, InceptionV3, inceptionResNetV2, Xception, DenseNet12, obteniendo así un notable resultado con un ensamble de las redes InceptionResNetV2, InceptionV3, y DenseNet121 de 0.995 AUROC.

En la referencia [11]. sé utilizó el conjunto de datos Yonsei (2,000), particularmente utilizó además de imágenes datos demográficos (peso, edad, género y altura) para el entrenamiento del algoritmo, reporta un resultado de 0.9213 AUROC al usar la red VGG19. En la referencia [12]. utilizó esta misma red además de una CNN con los datos Belarus (135) para la búsqueda de tuberculosis, identificando además la resistencia o sensibilidad a fármacos, obteniendo 66% de exactitud utilizando una CNN que incluye las características de forma y textura de la imagen.

Tabla 3: Principales hallazgos en este análisis de los conjuntos de datos, número de pacientes, técnica usada y la métrica de calidad (AUROC).

Referencia	Dataset	Pacientes	Red con mayor rendimiento	Resultado (AUROC)
[3]	Luzaka y Zambia	917	SVM Multiple Instance Learning + Active Learning	0.874
[4]	Kampala	138	VGG19	0.9213
[5]	Mongomery, Shenzhen y Belarús	1,104	CNN personalizada	0.925
[6]	ChestX-ray8 database	874	Quore.AI	0.929
[7]	Shenzhen y Mongomery Country	800	DenseNet121	0.937
[8]	Montgomery, Shenzhen, Kenya e India	967	AlexNet	0.950
[9]	Montgomery, Shenzhen, Belarus, y Thomas Jefferson	1007	Ensamble de AlexNet y GoogLeNet	0.990
[10]	Pediatric P, RSNA, Indiana y Shenzhen	37,306	Ensamble de InceptionResNetV2, InceptionV3, y DenseNet121	0.995
[11]	Yonsei	2,000	VGG19	0.9213
[12]	Belarus	135	CNN	0.66

3.2 Caracterización de genes de farmacorresistencia

Para el análisis de los diversos métodos de cada estudio del genoma se agrupan como fase de procesamiento; donde se incluye el manejo de la secuencias, ensamblaje, elaboración del pangenoma, y llamado de variantes; implementación del modelo de ML, donde se proponen los diferentes modelos, la fase de entrenamiento y prueba; precisión de las predicciones, se compara los resultados del método de machine learning con respecto a los estudios de asociación directa de genoma o las pruebas fenotípicas. Se incluyeron ocho estudios, en los cuales se implementaron diez métodos diferentes de ML, entre ellos support vector machine (SVM), logistic regression (LR), product-of-marginals (PM), 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 (seis estudios) fue SVM reportando una precisión desde el 93.89% hasta 73% [13]. [14].[15].[16].[17]. La referencia [15]. reporta al utilizar únicamente SVM una precisión de SVM del 73%, en contraste con [14]. que reporta precisión mayor al 80% utilizando SVM y LR. La Referencia [13]. reporta la precisión más alta de 93.89% al utilizar una combinación de SVM, LR, PM. 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 machine learning 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 ciprofloxacino (CIP), 12% para amikacina (AK), moxifloxacino

(MOX), y ofloxacino (OFX), 8% para etambutol (EMB) y kanamicina (KAN)). La referencia [18]. 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). La sensibilidad del método 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%) [18]. La referencia [17]. utilizó un modelo basado en LR reportando una sensibilidad 81.2% a 82.5% un resultado menor al método de [18]. De acuerdo con estos resultados se observa mejores rendimientos cuando se utilizan modelos combinados que individuales.

En la referencia [15]. utiliza un método basado en siete modelos de machine learning (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 [15].

En la referencia [16]. 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) [16].

4. Conclusiones

La cantidad reducida de artículos en el campo de estudio específico y la heterogeneidad de los resultados plantea la necesidad de más análisis para caracterizar la utilidad de los métodos de machine learning como auxiliares de diagnóstico. Se observa una poca utilización de los modelos deep learning en los estudios analizados. Para los estudios de clasificación de imágenes de rayos X, se puede concluir que el uso de ensambles de ANN mejora considerablemente el resultado con respecto a los experimentos realizados únicamente con una red neuronal simple, además se puede mejorar el resultado haciendo uso de datos clínicos y datos demográficos. Los reportes muestran que es posible diagnosticar la presencia de M. tuberculosis y caracterizar el perfil de drogo sensibilidad utilizando únicamente imágenes radiológicas simples. En el caso de los estudios que utilizan modelos para la detección de mutaciones, se puede concluir que la eficacia en la predicción para cada gen es diferente de acuerdo con el modelo implementado, por lo tanto, que la integración de múltiples modelos es más eficaz que uno solo. El modelo más utilizado es SVM. Las redes neuronales artificiales han mostrado alta precisión, pero se encontraron pocos artículos que las implementen por lo que se requiere mayor investigación para su caracterización. Los resultados de esta revisión describen que para el análisis adecuado de los datos biomédicos con respecto al diagnóstico y clasificación de la TB es recomendable la implementación de múltiples modelos simultáneamente.

Declaración de conflictos de interés

Los autores declaran no tener ningún conflicto de interés para este trabajo.

Referencias

[1] World Health Organization, “Global report Tuberculosis 2019,” WHO., Geneva., GE, Switzerland, *WHO/CDS/TB/2019.* 15, 2019.

[2] D. Moher *et al*. "Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement," *Syst Rev,* vol. 4, no. 1, pp. 2046-4053, 2015.
[3] J. Melendez, *et al.* "An automated tuberculosis screening strategy combining X-ray-based computer-aided detection and clinical information," *Sci Rep*, vol. 6, no. 1, pp. 25265, 2016.
[4] A. S. Becker, *et al.* "Detection of tuberculosis patterns in digital photographs of chest X-ray images using Deep Learning: feasibility study," *Int J Tuberc Lung Dis,* vol. 22, no. 3, pp. 328-335, 2018.
[5] F. Pasa, *et al.* "Efficient Deep Network Architectures for Fast Chest X-Ray Tuberculosis Screening and Visualization," *Sci Rep,* vol. 9, no. 1, pp. 6268, 2019.
[6] R. Singh, *et al.* "Deep learning in chest radiography: Detection of findings and presence of change," *Plos One*, vol. 13, no. 10, 2018.
[7] O. Gozes and H. Greenspan, "Deep Feature Learning from a Hospital-Scale Chest X-ray Dataset with Application to TB Detection on a Small-Scale Dataset," in *41st Annual Int Conference of the IEEE Eng Med Biol Soc*., Berlin, DE, Germany, 2019, pp. 4076-4079.
[8] S. Rajaraman, *et al.* "A novel stacked generalization of models for improved TB detection in chest radiographs," in *Conf Proc IEEE Eng Med Biol Soc*., Honolulu, HI, United Satets, 2018, pp. 718-721.
[9] P. Lakhani and B. Sundaram, "Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks," *Radiol,* vol. 284, no. 2, pp. 574-582, 2017.
[10] S. Rajaraman and S. K. Antani, "Modality-Specific Deep Learning Model Ensembles Toward Improving TB Detection in Chest Radiographs." *IEEE Access,* vol. 8, no, 1, pp. 27318-27326, 2020.
[11] SJ. Heo, *et al. "*Deep Learning Algorithms with Demographic Information Help to Detect Tuberculosis in Chest Radiographs in Annual Workers' Health Examination Data," *Int J Environ Res Public Health,* vol. 16, no. 2, pp. 250¸ 2019.
[12] S. Jaeger, *et al.* "Detecting drug-resistant tuberculosis in chest radiographs," *Int J Comput Assist Radiol Surg,* vol. 13, no. 12, pp. 1915-1925, 2018.
[13] S. Kouchaki, *et al.* "Application of machine learning techniques to tuberculosis drug resistance analysis" *Bioinformatics*, vol. 35, no. 13, pp. 2276-2282, 2019.
[14] AS. Chowdhury, *et al.* "Capreomycin resistance prediction in two species of Mycobacterium using a stacked ensemble method," *J Appl Microbiol,* vol. 127, no. 1, pp. 1656-1664. 2019.
[15] ES. Kavvas, *et al.* "Machine learning and structural analysis of Mycobacterium tuberculosis pan-genome identifies genetic signatures of antibiotic resistance," *Nat Commun,* vol. 9, no. 1, pp. 4306, 2018.
[16] Y. Yang, *et al.* "Machine learning for classifying tuberculosis drug-resistance from DNA sequencing data," *Bioinformatics,* vol. 34, no. 10, pp. 1666-1671, 2018.
[17] S. Jamal, *et al*. "Artificial Intelligence and Machine learning based prediction of resistant and susceptible mutations in Mycobacterium tuberculosis," *Sci Rep,* vol. 10, no. 1, pp. 5487, 2020.
[18] FJ. Duffy, EG. Thompson, TJ Scriba, DE. Zak, *"*Multinomial modelling of TB/HIV co-infection yields a robust predictive signature and generates hypotheses about the HIV+TB+ disease state," *PLoS One,* vol. 14, no. 7, 2019.
[19] W. Deelder, *et al.* "Machine learning predicts accurately Mycobacterium tuberculosis drug resistance from whole genome sequencing data," *Front Gen,* vol. 10, no. 1, pp. 922, 2019.