Universidad Autónoma de Querétaro

Facultad de Informática.

Modelo en aprendizaje automático para predicción de blancos de micro-RNAs de cáncer de mama.

Tesis

Que como parte de los requisitos para obtener el Grado de

Maestro en Ciencias de la computación

Presenta

Jorge Alberto Contreras Rodríguez

Dirigido por:

Dra. Diana Margarita Córdova Esparza

Co-Director:

Dra. Macrina Beatriz Silva Cázares

Querétaro, Qro. a 15 de marzo de 2024

Modelo en aprendizaje automático para predicción
de blancos de micro-RNAs de cáncer de mama.

por

Jorge Alberto Contreras Rodriguez

Universidad Autónoma de Querétaro
Facultad de Informática
Maestría

Modelo en aprendizaje automático para predicción de blancos de micro-RNAs de cáncer de mama.

Tesis

Que como parte de los requisitos para obtener el Grado de Maestro en Ciencias de la computación.

Presenta

Jorge Alberto Contreras Rodríguez

Dirigido por:

Dra. Diana Margarita Córdova Esparza

Co-dirigido por:

Dra. Macrina Beatriz Silva Cázares

Dra. Diana Margarita Córdova Esparza
Presidente

Dra. Macrina Beatriz Silva Cázares
Secretario

Dr. Julio Alejandro Romero González
Vocal

Dra. Ana Marcela Herrera Navarro
Suplente

Dr. Fidel González Gutiérrez
Suplente

Centro Universitario, Querétaro, Qro.
Fecha de aprobación por el Consejo Universitario (marzo 2024)
México

Agradecimientos

A lo largo de esta trayectoria académica, he tenido el privilegio de contar con el apoyo y la colaboración de diversas instituciones y personas, a las cuales deseo expresar mi más sincero agradecimiento.

A la **Universidad Autónoma de Querétaro**, por brindarme la oportunidad de formar parte de su comunidad educativa. Su compromiso con la excelencia académica, su ambiente de aprendizaje enriquecedor y apoyo económico con la beca de manutención han sido fundamentales en mi desarrollo profesional.

A la **Facultad de Informática**, por su dedicación a la formación de profesionales en el campo de la informática. El apoyo recibido, tanto a nivel académico como personal, ha sido sustancial para alcanzar mis metas.

Agradezco a los distinguidos maestros de la carrera de Ciencias de la Computación de la Universidad Autónoma de Querétaro. Su conocimiento, orientación y enseñanza han sido una motivación constante.

Mi gratitud se extiende al sínodo, director y asesor externo de mi tesis. Su guía experta y valiosos aportes han contribuido significativamente al éxito de este trabajo de investigación.

Quiero reconocer al **Consejo Nacional de Ciencia y Tecnología (CONAHCYT)**, cuyo apoyo financiero hizo posible llevar a cabo este proyecto. Su compromiso con el avance científico y tecnológico es fundamental para el progreso de la sociedad.

En resumen, agradezco a cada persona e institución que ha formado parte de mi trayectoria académica. Este logro no habría sido posible sin su contribución y apoyo constante.

Resumen

Esta tesis se centra en el desarrollo de un modelo de aprendizaje automático para la predicción de blancos de micro-RNAs (miRNAs) relacionados con el cáncer de mama. El objetivo principal es la creación de un algoritmo basado en un modelo de aprendizaje automático que emplea un clasificador para identificar blancos predictivos de miRNAs en el subtipo de cáncer de mama conocido como BRCA.

La metodología utilizada se basó en la recopilación de datos a partir de la plataforma BioPortal, específicamente de la fuente TCGA. Se seleccionó información relevante relacionada con la expresión de miRNAs en los subtipos luminal y basalike del cáncer de mama. Se implementó la técnica de aprendizaje automático conocida como "Bosques Aleatorios" para llevar a cabo la clasificación de los miRNAs de interés, es decir, aquellos que actúan como blancos predictivos de cáncer de mama.

Para evaluar el rendimiento del modelo, se aplicaron métricas clave, incluyendo la precisión, sensibilidad y especificidad. Los resultados revelaron una alta precisión del 95%, indicando que el modelo realiza predicciones precisas. Sin embargo, se observó una sensibilidad del 20%, lo que sugiere que el modelo tiene dificultades para identificar correctamente la mayoría de las muestras positivas. Por otro lado, se logró una especificidad del 100%, lo que indica que el modelo es eficaz en la identificación de las muestras negativas, también se determinó que el bajo rendimiento en la sensibilidad puede ocasionarse debido a las pocas muestras de clase positiva con las que fue entrenado este modelo aun así se logró identificar una clase positiva la cual es de suma importancia para esta investigación.

Este estudio tiene importantes implicaciones para la predicción de blancos predictivos de miRNAs en el cáncer de mama. La combinación de aprendizaje automático y el análisis de expresión génica podría mejorar significativamente los tratamientos y la detección temprana de esta enfermedad, lo que representa un avance significativo en la lucha contra el cáncer de mama.

Palabras clave: Cáncer de mama, Micro-RNAs (miRNAs), Predicción de blancos, Aprendizaje automático, Bosques aleatorios.

Abstract

This thesis focuses on the development of a machine learning model for the prediction of micro-RNA (miRNA) targets related to breast cancer. The main objective is to create an algorithm based on a machine learning model that employs a classifier to identify predictive targets of miRNAs in the breast cancer subtype known as BRCA. The methodology used was based on data collection from the BioPortal platform, specifically from the TCGA source. Relevant information related to miRNA expression in the luminal and basal-like subtypes of breast cancer was selected. The machine learning technique known as "Random Forests" was implemented to classify miRNAs of interest, those that act as predictive targets for breast cancer. To evaluate the model's performance, key metrics, including accuracy, sensitivity, and specificity, were applied. The results revealed high accuracy of 95%, indicating that the model makes accurate predictions. However, a sensitivity of 20% was observed, suggesting that the model has difficulty correctly identifying most positive samples. On the other hand, a specificity of 100% was achieved, indicating that the model is effective at identifying negative samples. It was also determined that the low sensitivity performance may be due to the limited number of positive class samples with which this model was trained. Nonetheless, the model successfully identified a positive class that is of utmost importance to this research. This study has significant implications for the prediction of miRNA predictive targets in breast cancer. The combination of machine learning and gene expression analysis could significantly improve treatments and early detection of this disease, representing a significant advancement in the fight against breast cancer.

Keywords: Breast cancer, Micro-RNAs (miRNAs), Target prediction, Machine learning, Random Forests.

ÍNDICE DE ABREVIATURAS

AD	Árboles de Decisión
ADN	Ácido Desoxirribonucleico (en inglés, Deoxyribonucleic Acid)
AUC	Área Bajo la Curva ROC (en inglés, Area Under the ROC Curve)
AVAD	Años de vida afectadas por discapacidad
BRCA	Gen de Susceptibilidad al Cáncer de Mama (en inglés, Breast Cancer Susceptibility Gene)
cDNA-RLM	Transcriptasa Inversa de ADN Complementario con Lugar de Inicio Específico (en inglés, Reverse Transcriptase of Complementary DNA with Random Hexamer Priming)
DISCR	Análisis Discriminante
DMTN	Dinámica Molecular de Red (en inglés, Molecular Dynamics Network)
DT	Árboles de Decisión (en inglés, Decision Trees)
F1	Puntuación F1
GDC1	Datos genómicos comunes
GEPIA	Expresión Génica Interactiva de Proyección de Expresión Génica en Cáncer (en inglés, Gene Expression Profiling Interactive Analysis)
GO	Ontología Génica (en inglés, Gene Ontology)
IA	Inteligencia Artificial (en inglés, Artificial Intelligence)
KNN	Vecinos más Cercanos (en inglés, k-Nearest Neighbors)
MEC	Matriz Extracelular
MEE	miRNAs estadísticamente expresados
miRNA	MicroARN (en inglés, microRNA)
ML	Aprendizaje Automático (en inglés, Machine Learning)
MVSs	Sistemas de Visión por Máquina (en inglés, Machine Vision Systems)
NB	Naïve Bayes

PALB-2	Gen de la Proteína Ligadora de BRCA2 (en inglés, Partner and Localizer of BRCA2)
RB	Proteína del Retinoblastoma (en inglés, Retinoblastoma Protein)
RBs	Redes Bayesianas (en inglés, Bayesian Networks)
RF	Bosques Aleatorios (en inglés, Random Forests)
RNA	Ácido Ribonucleico (en inglés, Ribonucleic Acid)
RNAm	ARN Mensajero (en inglés, Messenger RNA)
RNAs	Redes Neuronales Artificiales (en inglés, Artificial Neural Networks)
ROC	Curva Característica de Operación del Receptor (en inglés, Receiver Operating Characteristic)
SVM	Máquinas de Vectores de Soporte (en inglés, Support Vector Machines)
TCGA	Atlas del Genoma del Cáncer, en ingles
TNBC	Cáncer de Mama Triple Negativo (en inglés, Triple-Negative Breast Cancer)
TP53	Gen TP53 (también conocido como p53)
UTR	Región No Traducida (en inglés, Untranslated Region)
VPH	Virus del Papiloma Humano (en inglés, Human Papillomavirus)

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

Índice

CAPITULO I

1. Introducción

El cáncer se caracteriza por una desregulación celular que puede modificarse mediante control genético en los niveles postranscripcional y transduccional, que pueden regularse por medio del control del ciclo celular sobre los niveles de expresión de genes relacionados. Por lo tanto, las modificaciones se describen principalmente por procesos transcripcionales y de metilación de microRNA (miRNA) (Zhi et al., 2011).

El cáncer de mama (BRCA) es un tipo de cáncer que afecta las células epiteliales de la glándula mamaria, donde la multiplicación celular ocurre de manera anormal y descontrolada, desarrollando así la formación de tumores malignos. Las células de cáncer de mama surgen de las glándulas productoras de leche llamadas lobulillos y conductos (Figura 1), que son canales responsables de transportar la leche secretada por los lobulillos hacia el pezón (Breastcancer.org, 2021).

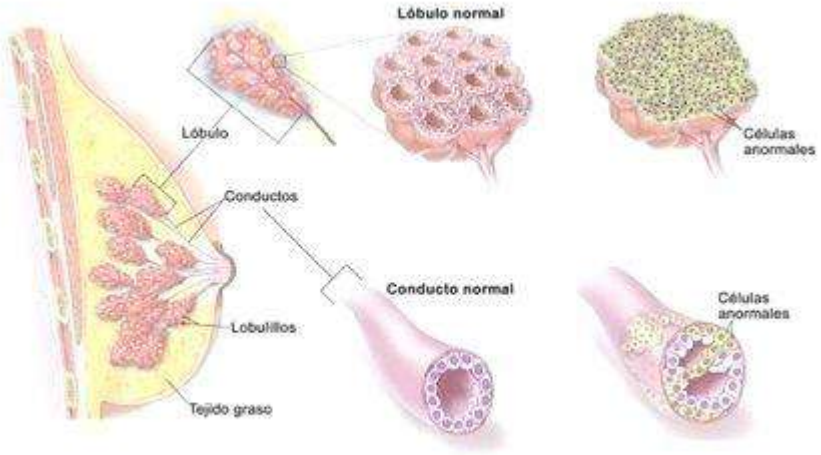


Figura 1 ***Anatomía de la glándula mamaria.*** **Fuente: Netter (2019).**

Los miRNA son pequeños RNA no codificantes que funcionan como importantes reguladores genéticos postranscripcionales de diversas funciones biológicas. En general, los miRNA regulan negativamente la expresión génica al unirse a sus RNA mensajeros selectivos (mRNA), lo que puede conducir a la degradación o

inhibición de la traducción del mRNA, dependiendo de los niveles de complementación con la secuencia blanco. La expresión anormal de estos miRNA se ha implicado en la etiología de varias enfermedades humanas (Loh et al., 2019).

Estos pequeños RNA no codificantes fueron descubiertos por primera vez por Ambros et al., en *C.elegans*, quienes encontraron que *lin-4*, el gen que controla el crecimiento en *C.elegans*, codifica una proteína, que produce dos pequeños RNA, uno de 22 nt de largo y el otro de 61 nt de largo. El RNA más largo adopta una estructura circular y es un precursor más corto, estos RNAs muestran regiones antisentido que son complementarias a varios sitios en la región 3' UTR de los genes 1 y 2 de lin*14* los investigadores demostraron que se observó una reducción de proteínas sin una disminución en los niveles de mRNA (Catalanotto et al., 2016).

Cada miRNA puede tener múltiples mRNA blanco y cada mRNA puede ser regulado por múltiples miRNA. Debido a la dificultad de identificar experimentalmente blancos de miRNA, ha aumentado la predicción de blancos bioinformáticos. Los resultados iniciales de estas predicciones son complejos, aunque los últimos análisis sugieren que se puede predecir con certeza al menos una clase de blancos de miRNA y que, en el momento del análisis, se podría predecir la clase de miRNA, que contiene el 30 % o más de genes humanos (Ta et al., 2019).

Existe una hipótesis generalizada de que múltiples miRNA pueden trabajar juntos para regular el mismo mRNA. Esta idea fue confirmada por experimentos *in vitro* y observaciones de que varias regiones 3'UTR contenían diferentes sitios de unión de miRNA. En algunos casos, también se observó que múltiples sitios de unión de este miRNA provocaban una mayor represión, que se incorporó a varios algoritmos de predicción.

Según Fan y Kurgan (2015), el campo de la predicción de miRNA blancos, basado en secuencias de bases de datos públicas, se encuentra en constante actualización. Los predictores difieren en muchos aspectos, en cuanto a los métodos de predicción subyacentes que tienen en cuenta los detalles de la unión de miRNA-mRNA, incluyendo:

- El uso de complementación de seguimiento de bases de datos
- Disponibilidad del sitio y mantenimiento evolutivo
- Evaluación empírica (conjunto de datos y procedimiento de evaluación; tipo de modelo predictivo utilizado)
- Disponibilidad (facilidad de uso)
- Popularidad e impacto
- Desempeño predictivo

Una de las contribuciones más relevantes en la identificación de nuevos blancos ha sido el descubrimiento de secuencias cortas contiguas de 6-8 pb en miRNAs que se unen al mRNA blancos de la cadena complementaria. Estos sitios de unión se encuentran comúnmente en la región 5' de los miRNA y se conocen como región semilla o *seed sequence* (Navarro, 2008).

Otros estudios han demostrado que además de los miRNAs que se unen a la región 3'UTR, también pueden unirse a regiones de codificación genética como la 5'UTR, que codifican una serie de proteínas ribosómicas para facilitar la traducción del código; esto sugiere otro sitio funcional para miRNAs (Albíztegui et al., 2014).

El uso de computadoras (y aprendizaje automático) en la previsión y pronóstico es parte de la creciente tendencia hacia la medicina personalizada y predictiva. Esta transición a la medicina es importante no solo para los pacientes (en la vida) sino también para los médicos (en las decisiones de tratamiento) (Cruz & Wishart, 2007).

Los progresos en el campo de la computación han contribuido en automatizar actividades humanas. Una de las áreas en donde se ha logrado un mejor adelanto es en la inteligencia artificial (IA). En la actualidad los expertos en las diferentes áreas de investigación del cáncer de mama evalúan la importancia del uso de los métodos de ML, esperan predecir mejor el riesgo aprovechando la información de múltiples niveles de “grandes datos”, para identificar nuevos marcadores genéticos, avanzar en la orientación precisa de la prevención y la detección temprana del cáncer de mama.

Sin embargo, prever de manera precisa la evolución de una enfermedad representa uno de los desafíos más emocionantes y complejos para los expertos médicos. Por esta razón, los enfoques de aprendizaje automático se han convertido en una herramienta cada vez más popular en la comunidad de investigadores médicos.

En el ámbito de la salud, la inteligencia artificial ya se está empleando como una solución para reducir la carga de trabajo de los profesionales, mejorando así tanto la atención médica como el proceso de diagnóstico. Es crucial resaltar la relevancia de la inteligencia artificial en la medicina, especialmente en lo que respecta a la detección temprana del cáncer de mama, ya que esta puede ser una medida efectiva para prevenir pérdidas de vidas debido a esta afección cuando se detecta a tiempo. La IA ha demostrado ser efectiva, rápida y precisa, ya se ha demostrado en varios escenarios clínicos que se pueden lograr buenos resultados (Sánchez & Valencia Orozco, 2020).

Para el proceso de identificar los blancos de miRNA, los enfoques de ML no abordan las propiedades de miRNA-mRNA, como la secuencia o la estabilidad termodinámica, sino que intentan reconocer posibles objetivos de miRNA al abordar las interacciones miRNA-mRNA biológicamente relevantes. Los algoritmos se entrenan utilizando interacciones miRNA-mRNA probadas experimentalmente como ejemplos positivos y ejemplos negativos generados

artificialmente. De esta forma, el software ML intenta reconocer los patrones que distinguen los objetivos reales de los objetivos falsos. En presencia de un nuevo conjunto de datos nunca visto, estos modelos se pueden usar para predecir correctamente si un blanco es "verdadero" o no (Riolo et al., 2020).

2. Marco Teórico

2.1. Cáncer

El cáncer, caracterizado por una proliferación celular descontrolada y eludir el sistema inmunológico, es una causa significativa de mortalidad global. A pesar de tratamientos convencionales como cirugía, quimioterapia y radioterapia, la recaída y la falta de supervivencia siguen siendo desafíos. Comprender el cáncer desde su inicio hasta la metástasis y la recurrencia es crucial para abordar este problema (Yin et al., 2021).

Comprender el origen de las células cancerosas es esencial para prevenir y evaluar el riesgo de cáncer. Estudios recientes han identificado que las células cancerosas se transforman a través de mutaciones genéticas y epigenéticas, lo que les permite proliferar y formar tumores. Este comportamiento no solo depende del genotipo del huésped, sino también de factores como la dieta, las infecciones y el tabaquismo, que contribuyen a la oncogénesis y al riesgo de cáncer. Además, las deficiencias en el sistema inmunológico en la detección y destrucción de células cancerosas recién formadas desempeñan un papel crucial en el crecimiento y la propagación de los tumores. (Yin et al., 2021).

2.2. Cáncer de Mama

El cáncer de mama se origina en las células de los conductos lácteos o los lóbulos del tejido glandular de la mama. Inicialmente, el tumor es localizado y tiene poco potencial de propagación, pero con el tiempo puede volverse invasivo y diseminarse a los ganglios linfáticos cercanos o a otros órganos. Los

tratamientos, que incluyen cirugía, radioterapia y terapia farmacológica, pueden ser efectivos si se detecta a tiempo y salvan vidas.

No se conoce una causa infecciosa para el cáncer de mama, y factores como la edad, la obesidad, el consumo de alcohol, el historial familiar y la exposición a la radiación pueden aumentar el riesgo.

Las mutaciones genéticas heredadas, como BRCA1, BRCA2 y PALB-2, también pueden contribuir al riesgo. Aunque se pueden controlar algunos factores de riesgo, ser mujer es el principal factor de riesgo.

Los procedimientos de cáncer de mama han avanzado para incluir tratamientos menos invasivos y terapias dirigidas específicas (Cáncer de mama, s/f).

2.3. Estadísticas del cáncer de mama.

El cáncer de mama es una neoplasia común y recurrente en todo el mundo, con una alta prevalencia y una mayor incidencia en países desarrollados, aunque Japón es una excepción. Es la principal causa de muerte entre las mujeres, con tasas de mortalidad en constante aumento debido a factores como el envejecimiento de la población, cambios en el estilo de vida y la relación entre el cáncer y la obesidad. Por otro lado, el cáncer de cuello uterino también es un problema global, con la mayoría de los casos diagnosticados en países en desarrollo. Aunque las tasas de mortalidad han disminuido gracias a mejoras en las condiciones sociales y la atención médica, las áreas desfavorecidas aún enfrentan tasas más altas de mortalidad. En México, se han registrado un considerable número de casos de cáncer de mama y algunas regiones, como Sonora y Nuevo León, tienen tasas de mortalidad más altas. La detección suele ocurrir alrededor de los 54.9 años, siendo el grupo de 50 a 59 años el más afectado (Centro Nacional de Equidad de Género y Salud Reproductiva, s/f).

2.4. Rasgos del cáncer.

Las células cancerosas presentan deficiencias en los sistemas reguladores que controlan la homeostasis y la proliferación celular, lo que da lugar a la formación de más de 100 tipos de cáncer, cada uno con sus propias variaciones. Se pueden identificar seis cambios esenciales en la fisiología celular que impulsan el crecimiento maligno: la capacidad de recibir señales de crecimiento de manera autónoma, la insensibilidad a señales que deberían inhibir el crecimiento, la habilidad para evitar la muerte celular programada, la capacidad de replicación ilimitada, la continua formación de nuevos vasos sanguíneos (angiogénesis), la invasión de tejidos circundantes y la formación de metástasis.

Estos cambios representan la superación del mecanismo de defensa anticancerígeno y son comunes en la mayoría de los tumores humanos, explicando la rareza relativa del cáncer a lo largo de la vida de una persona (Figura 2) (Hanahan & Weinberg, 2000).

Figura 2. Hallmarks del cáncer. Fuente: *Hanahan D, et.al.* 2011

Mantenimiento de la señalización proliferativa.

En la investigación sobre el mantenimiento de la señalización proliferativa en células cancerosas, se destaca su habilidad para regular la proliferación continua. Las células normales controlan con precisión las señales de crecimiento, manteniendo la homeostasis en los tejidos. Sin embargo, las células cancerosas desregulan estas señales, activando dominios de tirosina quinasa intracelulares y desencadenando vías de señalización intracelular. Esto no solo afecta el crecimiento celular, sino también la supervivencia y el metabolismo.

Evadir los supresores del crecimiento.

Las células cancerosas no solo activan señales de crecimiento, sino que también suprimen genes supresores de tumores, esenciales para el control del crecimiento. La inactivación de estos genes contribuye al desarrollo del cáncer. Se han identificado varios supresores de tumores, como RB y TP53, que regulan la proliferación celular y pueden actuar como verdaderos supresores de tumores. Estos genes desempeñan un papel central en el control de la selección celular entre la proliferación y la apoptosis (Douglas Hanahan & Weinberg, 2011).

Activación de la invasión y la metástasis.

En el año 2000, la comprensión de los mecanismos detrás de la invasión y metástasis en el cáncer era limitada. Se sabía que los carcinomas epiteliales progresaban patológicamente, lo que involucraba cambios en la forma celular y su interacción con otras células y la matriz extracelular. Un factor importante era la pérdida de la molécula E-cadherina, crucial para la cohesión celular. E-cadherina ayuda a mantener la integridad de las células epiteliales y su inactividad. Su aumento se asociaba con resistencia a la invasión y metástasis, mientras que su disminución potenciaba estos fenómenos. La pérdida y mutación de E-cadherina en el carcinoma humano subrayan su papel como supresor clave de estas características cancerígenas.

Permitir la inmortalidad replicativa.

En el año 2000, se comprendió que las células cancerosas deben tener la capacidad de replicarse indefinidamente para formar tumores. Esto contrasta con las células normales del cuerpo, que tienen un límite en su número de divisiones. Esta limitación se debe al envejecimiento y a una fase de crisis, durante la cual muchas células mueren. Sin embargo, algunas células pueden superar esta crisis y adquirir un potencial replicativo ilimitado, lo que se conoce como inmortalización. Esta característica es compartida por la mayoría de las líneas celulares establecidas, permitiéndoles proliferar en cultivo sin envejecer ni experimentar crisis.

Inducción de la angiogénesis.

Los tumores, al igual que los tejidos normales, necesitan nutrientes, oxígeno y la capacidad de eliminar desechos. La angiogénesis, un proceso que genera neovasos, satisface estas necesidades. En la embriogénesis, se forman nuevas células endoteliales y vasos (vasculogénesis), y se desarrollan vasos a partir de los existentes (angiogénesis). En adultos, la angiogénesis es transitoria en procesos como cicatrización y reproducción. Sin embargo, en el crecimiento tumoral, un "interruptor angiogenético" se mantiene activado, estimulando la formación continua de vasos que respaldan el crecimiento del tumor. (D. Hanahan & Folkman, 1996).

Resistencia a la muerte celular.

La apoptosis, muerte celular programada, ha sido una barrera natural contra el cáncer durante décadas (Adams & Cory, 2007; Lowe et al., 2004; Evan & Littlewood, 1998). Se investigaron los circuitos de señalización que desencadenan la apoptosis en respuesta a tensiones fisiológicas o terapias contra el cáncer. Los desequilibrios de señalización, causados por señalización oncogénica y daño en el ADN debido a la hiperproliferación, inducen la apoptosis. Sin embargo, algunos tumores desarrollan resistencia a la terapia y

evitan la apoptosis (Adams & Cory, 2007; Lowe et al., 2004).

Características Habilitadoras Y Nuevas Señas De Identidad.

La "ventaja del cáncer" se refiere a la capacidad funcional que permite la supervivencia y propagación de las células cancerosas, adquirida a través de diversos mecanismos y momentos en la tumorigénesis. La inestabilidad genómica, que lleva a mutaciones aleatorias y reordenamientos cromosómicos, es un factor destacado en la adquisición de habilidades especiales por parte de las células cancerosas. Además, el estado inflamatorio de las lesiones premalignas puede promover el desarrollo tumoral de varias maneras. El segundo atributo es la evasión del sistema inmunológico, lo que destaca el papel dual de este sistema en la promoción y antagonismo del cáncer. Ambos atributos contribuyen al desarrollo y progresión de varios tipos de cáncer en humanos, siendo características clave de los cánceres emergentes (Negrini et al., 2010; Luo et al., 2009; Colotta et al., 2009).

Inestabilidad del genoma y mutación.

La progresión tumoral implica una secuencia de cambios genómicos en células neoplásicas que les permiten crecer y dominar localmente. Cada paso en esta progresión se representa como una expansión clonal causada por la adquisición de un genotipo mutante facilitador. Además de las mutaciones, fenotipos heredados como la inactivación de genes supresores de tumores pueden adquirirse mediante cambios epigenéticos, como la metilación del ADN y modificaciones de histonas. Esto sugiere que las expansiones clonales pueden ser desencadenadas tanto por mutaciones como por cambios no mutacionales en la regulación génica (Berdasco & Esteller, 2010; Esteller, 2007; Jones & Baylin, 2007).

Promotor de tumores Inflamación.

Los patólogos han observado que algunos tumores muestran una densa infiltración de células inmunitarias, lo que refleja un estado inflamatorio en los tejidos. Esta infiltración varía desde sutiles infiltrados que solo son detectables con marcadores específicos hasta inflamaciones evidentes. Históricamente, se creía que estas respuestas inmunitarias representaban los esfuerzos del sistema inmunitario por eliminar los tumores, y hay evidencia creciente de respuestas antitumorales contra varios tipos de tumores. Sin embargo, los tumores también desarrollan mecanismos para evitar la destrucción inmunitaria (Flier et al., 1986; Page S et al., 2010).

Reprogramación de Energía Metabolismo.

La proliferación celular descontrolada en la enfermedad neoplásica no solo involucra un crecimiento celular desregulado, sino también una alteración del metabolismo energético. En situaciones en las que hay oxígeno presente, las células normales siguen una vía metabólica que involucra la glucólisis en el citosol y la posterior utilización de las mitocondrias. Sin embargo, las células cancerosas, a pesar de la disponibilidad de oxígeno, exhiben un metabolismo energético atípico, limitando su uso y favoreciendo la glucólisis, conocido como "glucólisis aeróbica" (Warburg effect) (Warburg, 1956).

Evadir la destrucción inmunitaria.

La formación de tumores y su relación con el sistema inmunológico plantean un enigma sin resolver. La teoría de la vigilancia inmunológica sugiere que el sistema inmunológico está diseñado para detectar y eliminar las células cancerosas incipientes, evitando la formación de tumores. Sin embargo, los tumores sólidos que se desarrollan pueden eludir la detección inmunológica o limitar su eliminación, lo que aún no se comprende por completo (Douglas Hanahan & Weinberg, 2011).

2.5. MicroRNA.

El primer microRNA (miRNA), lin-4, descubierto por los grupos de Ambrose y Ruvkun en 1993 en Caenorhabditis elegans (Lee et al., 1993; Wightman et al., 1993), revolucionó el campo de la biología molecular. Antes de eso, el laboratorio de Horvitz caracterizó a lin-4 como uno de los genes que regulan el desarrollo temporal en las larvas de C. elegans (Chalfie et al., 1981; Horvitz & Sulston, 1980). Más tarde, el mismo grupo descubrió que una mutación en lin-4 tenía el fenotipo opuesto a una mutación en otro gen, lin-14 (Ambros & Horvitz, 1987; Ferguson et al., 1987). Ambros y Ruvkun continuaron estudiando lin-4 y lin-14 y luego descubrieron que lin-4 no es un RNA codificante de proteínas sino un RNA pequeño no codificante (Almeida et al., 2011; R. Lee et al., 2004). También encontraron que lin-14 estaba regulado a nivel postranscripcional a través de su región 3' no traducida (UTR) y que la secuencia lin-4 era complementaria a la secuencia 3'-UTR de lin-14 (Lee et al., 1993). Por lo tanto, sugirieron que lin-4 regula lin-14 a nivel postranscripcional (Wightman et al., 1993). Desde entonces, los miRNA se han identificado en todos los sistemas de modelos animales y algunos están altamente conservados entre especies (Davis-Dusenbery & Hata, 2010; Friedländer et al., 2014; Li et al., 2010; Pasquinelli et al., 2000). Los miRNA son pequeños RNA no codificantes con una longitud promedio de 22 nucleótidos. La mayoría interactúa con la región UTR 3' del RNA mensajero (RNAm) para inhibir su expresión (Ha & Kim, 2014). Además, se ha demostrado que los miRNA activan la expresión génica en determinadas condiciones (Broughton et al., 2016). Estudios recientes indican que los miRNA se transportan entre diferentes compartimentos subcelulares para regular la tasa de traducción e incluso la transcripción (Makarova et al., 2016). Los miRNA son críticos para el desarrollo normal de los animales y están involucrados en muchos procesos biológicos (Fu et al., 2013). La expresión anormal de miRNA está asociada con muchas enfermedades humanas (Tüfekci et al., 2014). Además, los miRNA se secretan en los fluidos extracelulares y se han informado ampliamente como biomarcadores potenciales en varias enfermedades y también como moléculas

de señalización que median la comunicación de célula a célula (Hayes et al., 2014; Paul et al., 2018).

2.6. Modelos *in silico* o bioinformáticos

La biología continuará experimentando un aumento en el uso de matemáticas y simulaciones por computadora, tal como ha sucedido en otros campos del conocimiento. Ya se ha observado esta tendencia, y es probable que continúe. Muchos otros campos de la ciencia y la ingeniería han desarrollado la ciencia de sistemas y simulaciones matemáticas complicadas hasta un alto nivel de sofisticación. Estas capacidades influyen en nuestra vida cotidiana. Los productos químicos que usamos todos provienen de refinerías y otros procesos químicos altamente integrados con estructuras de control complejas que rivalizan con las de las células vivas.

El proceso de construcción de modelos matemáticos de procesos biológicos complejos y su simulación por computadora será iterativo. Se comenzará a construir "organismos *in silico*" que son representaciones computarizadas de sus contrapartes in vivo. Las versiones iniciales se sintetizarán utilizando datos genómicos, bioquímicos y fisiológicos. Estos modelos tendrán algunas capacidades interpretativas y predictivas. Sin embargo, debido al conocimiento incompleto de las restricciones y la anotación errónea, estos modelos iniciales solo podrán representar algunas funciones del organismo correctamente.

En el proceso de construir modelos iterativos de organismos *in silico*, se debe aprender a aceptar el fracaso. La principal diferencia entre un organismo *in silico* e in vivo es que la versión *in silico* carece de algunas características. Por lo tanto, es necesario formular hipótesis experimentales basadas en el análisis *in silico*, llevar a cabo los experimentos y actualizar los modelos. Este proceso iterativo para construir organismos *in silico* probablemente tendrá dos bucles de retroalimentación. Uno será el bucle experimental clásico y el otro, el *in silico*. Es

probable que muchas correcciones y ajustes para estos modelos provengan del análisis y la búsqueda de las cada vez más disponibles bases de datos bioinformáticas. ¿Qué se hará con estos modelos *in silico*? Es probable que tengan algún uso científico básico, como la genómica comparativa y estudios evolutivos.

Un problema adicional que merece ser comentado en este proceso iterativo de construcción de modelos es el enfoque de "necesidad de saber todo" que generan las tecnologías de alto rendimiento. Sin embargo, como se ha demostrado en otros campos, se pueden construir modelos de computadora poderosos y útiles sin "saber todo". Si se insistiera en tener modelos de computadora que explicaran cada detalle de un proceso en estudio, no se podrían construir aviones ni refinerías. De hecho, una de las habilidades en la construcción de modelos es determinar lo que se necesita para sintetizar un modelo de computadora perspicaz y útil. Es probable que las lecciones aprendidas de otros campos beneficien la construcción de modelos en biología (Palsson, 2000).

2.7. Métodos de predicción de blancos de miRNA

Enfoques computacionales para predecir los posibles blancos de los miRNAs pueden simplificar el procedimiento, permitiendo una selección inicial para reducir el número de sitios de destino que se validan experimentalmente. Existen varias herramientas para el análisis computacional, cada una utilizando una estrategia diferente para predecir los posibles blancos de los miRNAs, y su número está constantemente aumentando. Ahora, los usuarios tienen la oportunidad de acceder a una amplia variedad de soluciones, pero deben decidir qué herramienta utilizar. Esta elección puede no ser fácil; al menos es necesario estar familiarizado con las suposiciones básicas y la interpretación de los resultados (Riolo et al., 2020).

Hay varios enfoques para desarrollar algoritmos de predicción de blancos de miRNA. Se pueden dividir en dos categorías principales: algoritmos derivados de características de la secuencia de mRNA y/o basados en la interacción miRNA-mRNA, e inferencia estadística basada en aprendizaje automático.
En el primer caso, se tienen en cuenta diferentes características del complejo miRNA-mRNA. Es capaz de analizar y evaluar los apareamientos de la secuencia de semilla de los miRNA y los mRNA. Se puede realizar un análisis termodinámico calculando la energía libre de la formación del apareamiento y su estabilidad termodinámica. Además, evaluar la conservación evolutiva de la secuencia objetivo en especies relacionadas. Por último, es posible evaluar la accesibilidad estructural del 3'-UTR para los miRNA y calcular el número de sitios de blanco de miRNA, ya que los mRNAs soportan ser regulados por la unión de diferentes miRNAs a múltiples sitios de blanco.

En el caso del aprendizaje automático, la idea es identificar los blancos de miRNA que hacen referencia a los dúplex miRNA-mRNA con significado biológico comprobado, en lugar de hacer predicciones "de novo" a partir de características de secuencia. El aprendizaje automático en general es una aplicación de inteligencia artificial que proporciona a los sistemas la capacidad de mejorar automáticamente a través de la experiencia; "aprenden" de los conjuntos de datos de muestra y utilizan la información adquirida para hacer predicciones sobre datos desconocidos (Bishop, 2006).

2.8. Bases de datos de uso público

Durante los últimos años, diversas iniciativas de investigación biomédica han buscado ofrecer acceso libre a sus datos a fin de estimular la innovación. Muchas de estas iniciativas han adoptado el modelo de "código abierto" que ha cobrado relevancia en la industria informática. Cuando se utiliza en el contexto de software, el término "código abierto" se refiere a un proyecto de desarrollo de software para el cual el código fuente del ordenador se hace públicamente

disponible para que los licenciatarios lo utilicen, modifiquen y redistribuyan, siempre y cuando estos licenciatarios pongan sus mejoras a disposición de otros bajo los mismos términos, enfoque conocido como *copyleft*.

Cuando se habla de "biotecnología de código abierto" o "ciencia abierta" en el contexto de datos adquiridos a través de investigaciones biomédicas, se refiere a la liberación rápida de los datos del proyecto en el dominio público, con ciertas condiciones que incluyen la obligación de que los usuarios de los datos no restrinjan el acceso de otros usuarios a los mismos a través de derechos de propiedad intelectual. El término "código abierto" se utiliza para referirse a este enfoque de acceso a los datos. En la investigación biotecnológica, es crucial compartir los datos ya que muchos de los resultados, como secuencias de ADN humano aisladas, no tienen sustitutos (Gitter, 2010a).

El acceso abierto a los datos genómicos para la investigación científica y el progreso médico (Birney et al., s/f; Walport & Brest, 2011; Sharing data from large-scale biological research projects: A system of tripartite responsibility, 2003; *Joint statement by President Clinton and Prime Minister Tony Blair of the*, 2000) ha sido ampliamente reconocido como importante por la comunidad científica, los financiadores de investigación y los gobiernos. Actualmente, el acceso abierto es una práctica bien establecida para proyectos científicos comunitarios de gran escala financiados con fondos públicos, especialmente en el campo de la genómica. Si bien hay un consenso general en favor del acceso abierto, ciertos desarrollos Gitter, (2010) han llevado a científicos y responsables políticos a investigar e implementar restricciones al acceso abierto (Gitter, 2010; Dyke & Hubbard, 2011; Tenopir et al., 2011; Joly et al., 2011; Fortin et al., 2011). Entre ellos, se encuentran las preocupaciones sobre la privacidad dentro de la comunidad genómica y las críticas de algunos investigadores que consideran que el acceso abierto sin regulación podría plantear importantes problemas científicos, éticos y legales, como la calidad de los datos, la adecuada atribución

a los generadores de datos, la relevancia del sistema para proyectos pequeños y medianos (Joly et al., 2012).

2.9. Técnicas de aprendizaje automático.

El cáncer, un trastorno altamente diverso, compuesto por numerosos subtipos, destaca la necesidad de una detección y diagnóstico preciso, lo cual se ha vuelto fundamental en la investigación oncológica. Clasificar a los pacientes en grupos de alto o bajo riesgo es esencial para orientar el tratamiento clínico. Esto ha llevado a investigadores en los campos biomédico y de bioinformática a explorar el uso de técnicas de aprendizaje automático. Estas herramientas se han aplicado para modelar la progresión y el tratamiento del cáncer, y su capacidad para identificar características clave en datos complejos es crucial. Se han empleado diversas técnicas, como Redes Neuronales Artificiales, Redes Bayesianas, Máquinas de Vectores de Soporte y Árboles de Decisión, en investigaciones oncológicas para desarrollar modelos predictivos que mejoren la toma de decisiones clínica.

El objetivo principal de las técnicas de aprendizaje automático es producir un modelo que pueda utilizarse para realizar tareas de clasificación, predicción, estimación u otra tarea similar. La tarea más común en el proceso de aprendizaje es la clasificación. Esta función de aprendizaje clasifica el elemento de datos en una de varias clases predefinidas. Al desarrollar un modelo de clasificación mediante técnicas de aprendizaje automático, pueden producirse errores de entrenamiento y generalización.

Aunque es evidente que el uso de métodos de aprendizaje automático puede mejorar la comprensión de la progresión del cáncer, se necesita un nivel adecuado de validación para que estos métodos sean considerados en la práctica clínica cotidiana (Kourou et al., 2015).

2.10. Algoritmos de predicción de los blancos de miRNA.

Antes del desarrollo de varias herramientas de predicción de blancos de miRNA, los blancos de miRNA eran examinados manualmente y confirmados mediante técnicas que consumían mucho tiempo y trabajo como el ensayo de luciferasa, análisis de expresión génica, identificación rápida de extremos de cDNA-RLM (Rapid Amplification of cDNA Ends), e inmunoprecipitación de los componentes del complejo de silenciamiento inducido por ARN (RISC) (Thomson et al., 2011). Un análisis por Thomson et al. (2011) resume las fortalezas y debilidades de los métodos experimentales utilizados para la identificación de los blancos de miRNA. La idea de desarrollar algoritmos de predicción de blancos *in silico* surgió a partir de la observación de que los miRNAs generalmente tienen un patrón de direccionamiento, lo que llevó al descubrimiento de los primeros blancos para los miRNAs let-7 y lin-4.

El desarrollo de herramientas computacionales para la predicción de miRNAs ha revolucionado la investigación de los miRNAs; no obstante, estas herramientas deben usarse con precaución debido a la alta tasa de falsos positivos (Ab Mutalib et al., 2019).

No obstante, estas herramientas de predicción computacional pueden ayudar a reducir la necesidad de validación experimental. Un solo miRNA puede regular múltiples objetivos; del mismo modo, el mismo objetivo también puede ser regulado por muchos miRNAs (Lewis et al., 2005). Se han identificado experimentalmente grandes cantidades de objetivos validados de miRNAs; sin embargo, la información es bastante dispersa. DIANA-TarBase v7.0, una base de datos que ha catalogado las interacciones miRNA; objetivo, contiene más de 500.000 interacciones validadas mediante curación manual (Vlachos et al., 2015) validadas experimentalmente, curadas a partir de 1165 publicaciones, lo que equivale a 9 a 250 veces más entradas que cualquier otra base de datos relacionada.

Hay muchos principios que se utilizan para la predicción computacional de los blancos de los miRNA. Los algoritmos combinan varias características para aumentar la eficiencia de la predicción, incluyendo (1) complementariedad de la secuencia de semillas (Lewis et al., 2005; Lewis et al., 2003); (2) estado de conservación evolutiva (Lewis et al., 2003); (3) energía libre (Yue et al., 2009); (4) accesibilidad del sitio objetivo (Mahen et al., 2010; Marín & Vaníček, 2011); (5) abundancia del sitio objetivo (Garcia et al., 2011); (6) enfoque basado en patrones (Miranda et al., 2006); (7) contenido de flanqueo AU local (Betel et al., 2010); y (8) wobble GeU (Doench & Sharp, 2004). Sin embargo, estos principios se limitan a la interpretación humana de la interacción entre los miRNAs y sus objetivos (Liu et al., 2008). Por lo tanto, se cree que el algoritmo de aprendizaje automático que no utiliza información de semillas ni estado de conservación tiene el potencial de aumentar aún más la precisión de la predicción.

Aproximadamente el 6% de las herramientas de predicción de miRNA humano incorporan algoritmos de aprendizaje automático (Lukasik et al., 2016) de forma independiente o en combinación con otros principios. Es importante comprender la base de los algoritmos de predicción de blancos antes de decidir qué método aplicar.

Según Li et al. (2020), estos métodos actuales se pueden dividir en dos categorías:
(1) Métodos basados en aprendizaje automático,
(2) Métodos basados en puntuación.
En el primer tipo, las máquinas de vectores de soporte, las redes neuronales, las redes neuronales complejas profundas y la regresión logística se usan comúnmente para crear modelos basados en múltiples características, incluidas las características de topología de los miRNA y las similitudes de la enfermedad.

En este último se utilizan determinadas estrategias o métodos de puntuación basados en diferentes similitudes entre miRNAs y enfermedades.

En general, se acepta que los miRNA regulan la expresión génica mediante la regresión transcripcional de su mRNA blanco específico, mientras que la expresión de un mRNA está regulada por varios miRNA. Para este mecanismo, se tiene un modelo de ecuación lineal, en el que 1 mRNA se ve afectado por varios miRNAs $(x_1, x_2, \ldots, x_m)$ como se muestra en la ec.1:

$$y_i = a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{im}x_m,\ i = 1, \ldots, n \qquad \text{(ec.1)}$$

Donde a_{ij} muestra la influencia del jth miRNA en el ith mRNAs, x_j simboliza el nivel de expresión del jth miRNA y y_i muestra el nivel de expresión del ith mRNAs.

Para estudiar la conexión entre **m** miRNAs y **n** mRNAs simultáneamente, se reescribe el sistema de ecuaciones lineales en forma de matriz ec.2.

$$\begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1m} \\ a_{21} & a_{22} & \ldots & a_{2m} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ a_{11} & a_{12} & \ldots & a_{1m} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ . \\ . \\ . \\ x_n \end{pmatrix} \qquad \text{(ec.2)}$$

En el que las mediciones de la expresión de m diferentes miRNAs se denotan por $(x_1, x_2, \ldots, x_m)$ y de la expresión de n mRNA distintas se denotan por $(y_1, y_2, \ldots, y_n)$. A partir de k veces experimentos ec.3.

$$y_{nxK} = A_{nxM}X_{nxK} \qquad \text{(ec.3)}$$

Al solucionar la ecuación como un problema inverso, y obteniendo a_{ij} como criterio desconocido, a_{ij} representa principalmente el efecto de x_i y y_j (Kim, 2018).

En el momento que se evalúa un clasificador en una serie de pruebas, la estimación de su rendimiento es obtenida en cualquier conjunto de pruebas de la misma distribución (Pereira et al., 2009).

El funcionamiento está muy relacionado con otra tecnología: Big Data (BD). Afortunadamente, muchos sistemas de salud han adoptado la digitalización. Como resultado, el historial del paciente y del tratamiento se guardan en formato digital. De esta forma, los sistemas pueden acogerlos, encontrando datos útiles para realizar tareas médicas analíticas y predictivas (BITAC, 2018).

2.11. Conjunto de datos para entrenamiento, validación y prueba

Los datos de entrenamiento, pruebas y validación son elementos fundamentales para el desarrollo y evaluación de modelos de aprendizaje automático. Cada uno de estos conjuntos de datos cumple distintos propósitos y exhibe características particulares.

Los datos de entrenamiento se usan para ajustar los parámetros del modelo y enseñar al algoritmo a hacer predicciones. Es crucial que los datos de entrenamiento sean representativos del problema y tengan etiquetas relevantes y completas (*Aprendizaje automático y datos de entrenamiento: lo que debes saber*, 2022).

Los datos de prueba se utilizan para validar el rendimiento del modelo y comprobar si puede generalizar bien a datos nuevos y nunca vistos. Deben estar separados de los datos de entrenamiento y no deben usarse para ajustar el modelo.

Los datos de validación se usan para optimizar el modelo y elegir los mejores hiperparámetros, y se pueden obtener a partir de los datos de entrenamiento utilizando técnicas como la validación cruzada o el método *hold-out*.

El uso de estos conjuntos de datos es crucial para crear modelos precisos, robustos y confiables, evitando problemas como el sobreajuste o el subajuste. Además, permite comparar diferentes modelos y elegir el más adecuado para el proyecto en cuestión (*Aprendizaje automático y datos de entrenamiento: lo que debes saber*, 2022; Na, 2020).

En el proceso de entrenamiento de un modelo de aprendizaje automático, es necesario separar el conjunto de datos en dos partes: el conjunto de entrenamiento (train) y el conjunto de pruebas (test), con el fin de evaluar su desempeño. Por lo general, se divide el conjunto de datos inicial en una proporción de 80-20, y se toman muestras aleatorias en lugar de secuenciales para asegurar una mezcla adecuada de los datos (*Aprendizaje automático y datos de entrenamiento: lo que debes saber*, 2022).

3. Antecedentes

Entre los diversos métodos de predicción contenidos en la literatura sobre este tema de investigación, se encuentran los siguientes (Tabla 1):

Tabla 1. Métodos de predicción de blancos de miRNA en cáncer de mama.

Autor, Año	Característica	Fuente	**valor P (**p-value)**, Cambio de pliegue (**Fold-change)	Método de aprendizaje automático	Rendimiento
Naorem, Muthaiyan y Venkatesan. (2019)	Datos de expresión de miRNA	Banco de Datos de Expresión Génica (GEO)	Valor p < 0.05 Cambio de pliegue >= 1.0	* Naïve Bayes-NB * Optimización secuencial mínima [SMO] * Bosque aleatorio [RF])	NB = 96.8447% SMO = 96.966% RF = 96.4806%
Yu et al. (2020)	El análisis de expresión diferencial	Base de Datos TCGA	p ajustado <= 0.01- Cambio de pliegue >= 0.5	* NB * RF * Máquinas de Soporte de Vector Radial l" (SVM con núcleo de	NB = 0.96 RF = 0.98 SMVRadial = 0.97 (tipo basal)

				base radial l)	
Sherafatian (2018)	Datos de expresión de miRNA	Base de Datos TCGA	Valor p > 0.5	Tres algoritmos basados en árboles (RF, Rpart y treebag)	RF = 0.845 (tipo basal)
Qiu et al. (2020)	El perfil de expresión de RNA mensajero y miRNA	Portal de datos Genomic Data Commons (GDC1). Los conjuntos investigados fueron los miRNA diferenciales en la cohorte de la base de TCGA BRCA.	Valor p < 0.05	* SVM	área bajo la curva (AUC) = 0.9633
Sarkar et al. (2021)	Los valores de expresión de miRNA basados en secuenciación de próxima generación (NGS).	El Atlas del Genoma del Cáncer (TCGA).	p-value < 0.05	* Máquinas de Soporte de Vector (SVM) * Red neuronal artificial (ANN) * K Vecinos Más Cercanos (KNN) * Árbol de decisión (DT) * Bosque Aleatorio (RF) * Naive Bayes(NB) y Análisis Discriminante (DISCR)	SVM = 74.9094 % ANN = 74.9094 % KNN = 67.1014 % DT = 64.4565 % RF = 76.5761 % NB = 70.5978 % DISCR = 73.19 %
Andreini et al. (2022)	Descubrir perfiles complejos de expresión de miRNA	El conjunto de datos BRCA del Atlas del Genoma del Cáncer (TCGA).	Cambio de pliegue > 2	* SVM * Bosque Aleatorio (RF) especializado de múltiples clases.	SVM = 0.926 RF = 0.9886

Naorem, Muthaiyan and Venkatesan (2019) diseñaron un estudio centrado en el cáncer de mama triple negativo (TNBC), un subtipo de cáncer de mama con un pobre resultado clínico para el cual no existe un tratamiento aprobado específico. Los miRNA han sido identificados como biomarcadores prometedores con un importante papel en la tumorigénesis del cáncer humano. Debido al creciente conjunto de datos de perfiles de miRNA de TNBC, su investigación requiere un análisis adecuado. La parte interesante de este estudio consiste en la regulación

al alza y a la baja de los miRNA con sus respectivos criterios de corte, como el cambio de pliegue y el valor p. Los miRNA regulados al alza y a la baja se enumeran en diferentes estudios y se priorizan en función del cambio de expresión y las estadísticas de valor p. El valor p o el cambio de pliegue de cada miRNA determina la importancia de la expresión en TNBC. Una metainformación de miRNA liberada de manera significativa (hsa-miR-135b-5p, hsa-miR-18a-5p, hsa-miR-9-5p, hsa-miR-522-3p, hsa-miR-190b, hsa-miR-9a) ha sido identificada en varios estudios y tiene una alta precisión predictiva. Las personas identificadas pueden ser candidatos prometedores para biomarcadores diagnósticos para CMTN. Por lo tanto, se analizaron los miRNA que desempeñan un papel importante en TNBC.

Yu et al. (2020) realizaron un estudio sobre los mecanismos de interacción génica para cada subtipo de cáncer de mama que pueden tener un impacto significativo en el tratamiento personalizado. Integraron la importancia biológica de los genes de las redes de regulación génica en el análisis de expresión diferencial para obtener los genes diferencialmente expresados ponderados (weighted DEGs). Estos contienen la importancia biológica derivada de la red de regulación génica. Basados en los cálculos de SDR ponderados, se aprendieron clasificadores binarios, los cuales mostraron buen rendimiento en términos de métricas como "Sensibilidad", "Especificidad", "Exactitud", "F1" y "AUC", con valores de SDR ponderados para los grupos de control y experimental, proporcionando nuevos resultados de análisis de enriquecimiento de Gene Ontology (GO). Los nuevos términos de GO enriquecidos revelarían funciones biológicas específicas entre todos los subtipos de BRCA.

En el estudio realizado por Sherafatian (2018), se utilizó un conjunto de datos de expresión de miRNA de pacientes con cáncer de mama de la base de datos TCGA para desarrollar modelos predictivos que identificaran biomarcadores de miRNA para el diagnóstico y la subtipificación molecular de BRCA. Se obtuvieron miRNA

de control negativo empíricos *in-silico* y se aplicaron tres algoritmos basados en árboles (bosque aleatorio, Rpart y treebag) al conjunto de datos de entrenamiento equilibrado de secuenciación de miRNA para modelar el estado del cáncer de mama basado en la expresión normalizada de los miRNA filtrados. Los resultados mostraron que hsa-miR-139 y has-miR-96 fueron consistentemente significativos en los tres modelos. Además, los diez mejores miRNA para clasificar los tumores de cáncer de mama del tejido sólido normal en tres algoritmos de aprendizaje automático basados en árboles fueron hsa-miR-139, has-miR-96, 15, 183, 592, 20,125 b.2, 21, 11 y 125b.1.

Qiu et al. (2020) construyó una red de genes objetivo de miRNAs (DMTN) con pares de miRNA-RNAm con puntuaciones significativas de desregulación, basado en relaciones de regulación desreguladas entre miRNAs y genes objetivo. El perfil de expresión de miRNA y RNAm utilizado en el estudio se obtuvo del portal de datos genómicos comunes (GDC1). Todos los análisis estadísticos y gráficos del estudio se realizaron en el entorno R. Además, identificaron 588 miRNAs y 3,146 genes entre fármacos, donde la expresión de miRNAs / genes se asoció significativamente con la respuesta de las células cancerosas a los fármacos anticancerígenos. Sus resultados indican que los niveles de expresión de los miRNAs de riesgo y sus genes monofásicos adyacentes pueden ser indicadores de la sensibilidad de las células cancerosas a los fármacos anticancerígenos. Sugirieron que, con una validación experimental y clínica adicional, estos miRNAs podrían servir como biomarcadores para guiar el tratamiento de pacientes con cáncer de mama.

Sarkar et al. (2021) llevaron a cabo un estudio utilizando datos de NGS de cáncer de mama para identificar los biomarcadores de miRNA más importantes. Los biomarcadores de miRNA seleccionados están fuertemente asociados con múltiples subtipos de cáncer de mama. Para ello, utilizaron datos de The Cancer Genome Atlas (TCGA) y propusieron una técnica en dos pasos llamada métodos

de selección de características incrustados en el aprendizaje automático, seguida de análisis de supervivencia. En la primera fase, para obtener esta lista de miRNA, seleccionaron el mejor entre siete técnicas de aprendizaje automático (Máquina de Vectores de Soporte (SVM), Redes Neuronales Artificiales (ANN), K Vecinos más Cercanos (KNN), Árbol de Decisión (DT), Random Forest (RF), Naive Bayes (NB) y Análisis Discriminante (DISCR)) utilizando el conjunto completo de características, se seleccionó la mejor técnica de aprendizaje automático en este caso RF. En la segunda fase, en función de los valores de precisión de clasificación, se consideran las características más importantes de cada método de selección para hacer un conjunto que proporcione miRNA como 8*, 7*, e incluso 1*. Estos resultados analíticos confirmaron el hecho de que los miRNA seleccionados son biomarcadores potenciales para el diagnóstico de subtipos de cáncer. Además, el análisis de enriquecimiento de GO (Gene Ontology) también reveló procesos biológicos, moleculares y celulares relacionados con el cáncer de mama de los miRNA seleccionados. En general, este estudio identificó 27 miRNA como biomarcadores potenciales y encontró que son responsables de diferentes subtipos de cáncer de mama.

En Andreini et al. (2022) se propuso un enfoque para utilizar fragmentos de miRNA como posibles biomarcadores para la detección del cáncer de mama. Se abordó el problema en dos etapas diferentes. En la primera, se entrenaron dos modelos de aprendizaje automático, el primero fue una máquina de vectores de soporte (SVM) para distinguir entre muestras saludables y células cancerosas, y el segundo fue un bosque aleatorio (RF) para clasificar los subtipos de cáncer. Se seleccionó el modelo más preciso para cada paso, ya sea SVM con 0.9926 o RF con 0.9886, y el conjunto correspondiente de hiperparámetros utilizando una validación cruzada de cuatro vías con un método de búsqueda en cuadrícula. En segundo lugar, se utilizó un enfoque de importancia de características para identificar las características más importantes que el modelo de aprendizaje automático utiliza para hacer sus predicciones, utilizando un enfoque de dos pasos

que utiliza dos clasificadores ad hoc para la clasificación de tumor/salud y, por ejemplo, la detección de subtipos. Una de las principales ventajas de su trabajo es el uso de dos conjuntos de datos completamente independientes para el entrenamiento y la prueba. Se produjeron los materiales con diferentes máquinas de secuenciación, y también se realizaron diferentes procesos bioinformáticos para el preprocesamiento de los datos de origen. En la clasificación saludable/tumoral, se lograron precisiones en línea con los mejores resultados publicados. Curiosamente, ninguna de las muestras tumorales fue clasificada como saludable.

4. Planteamiento Del Problema

En el año 2020, se diagnosticó a 2.3 millones de mujeres en todo el mundo con BRCA, y lamentablemente, 685,000 de ellas perdieron la vida debido a esta enfermedad. Para finales de ese año, alrededor de 7.8 millones de mujeres que habían sido diagnosticadas con BRCA en los últimos cinco años seguían con vida. BRCA es la afección más prevalente a nivel global y se estima que las mujeres con BRCA experimentan una mayor pérdida de años de vida ajustados por discapacidad (AVAD) en comparación con cualquier otro tipo de cáncer. Además, BRCA afecta a mujeres de todas las edades después de la adolescencia. La incidencia en la edad adulta va en aumento a nivel mundial (OMS, 2020).

La identificación de blancos terapéuticos precisos y eficientes es crucial para el desarrollo de tratamientos más efectivos. Se ha comprobado que la alteración en la producción y función de los microRNA está relacionada con el desarrollo de enfermedades en los seres humanos. Por lo tanto, los microRNA han surgido como una nueva herramienta prometedora para la detección de diversas enfermedades entre ellas el cáncer de mama (Rico-Rosillo et al., 2014).

En este estudio se propone abordar la predicción de blancos de miRNAs en el cáncer de mama mediante el uso de técnicas de aprendizaje automático, con el

objetivo de contribuir al avance de la investigación y el desarrollo de tratamientos más personalizados y eficientes para esta enfermedad.

5. Justificación

Los procesos relacionados con la salud generan una gran cantidad de información que es compleja de analizar. Esto se debe principalmente a la cantidad de datos, la velocidad de producción y la variedad, por ejemplo, texto, imágenes, archivos administrativos. Herramientas como el aprendizaje automático u otras técnicas de análisis de datos permiten superar estas dificultades al facilitar el suministro de información rápida y confiable para ayudar a tomar decisiones (Pedrero et al., 2021).

El enfoque básico de este trabajo es realizar predicciones sobre los blancos de miRNA en BRCA. La información será recopilada a través del conjunto de datos, para luego analizar y clasificar mediante un modelo de aprendizaje automático.

6. Hipótesis

Mediante el desarrollo de un modelo basado en aprendizaje automático es posible crear un algoritmo que permita identificar y predecir blancos de miRNAs en cáncer de mama.

7. Objetivos

Objetivo general:

Desarrollar un algoritmo basado en un modelo de aprendizaje automático utilizando un clasificador para predicción de blancos predictivos de miRNAs en cáncer de mama.

Objetivos específicos:

1. Identificar las bases de datos de uso público especializados en información de miRNA en BRCA, para clasificar blancos predictivos.

2. Construir un conjunto de datos de miRNA en BRCA para entrenamiento, validación y prueba que pueda ser utilizado para determinar la técnica de aprendizaje automático en el clasificador.
3. Establecer cuáles son las técnicas de aprendizaje automático que se pueden usar para realizar la clasificación de miRNAs como blancos predictivos en BRCA.
4. Validar matemáticamente el clasificador desarrollado de blancos predictivos de miRNA en BRCA.

CAPITULO II

Metodología.

En la **Figura 3**. se muestra la estrategia bioinformática de la metodología a realizar. El espacio de trabajo en el cual se realizó en la Facultad de Informática de la Universidad Autónoma de Querétaro.

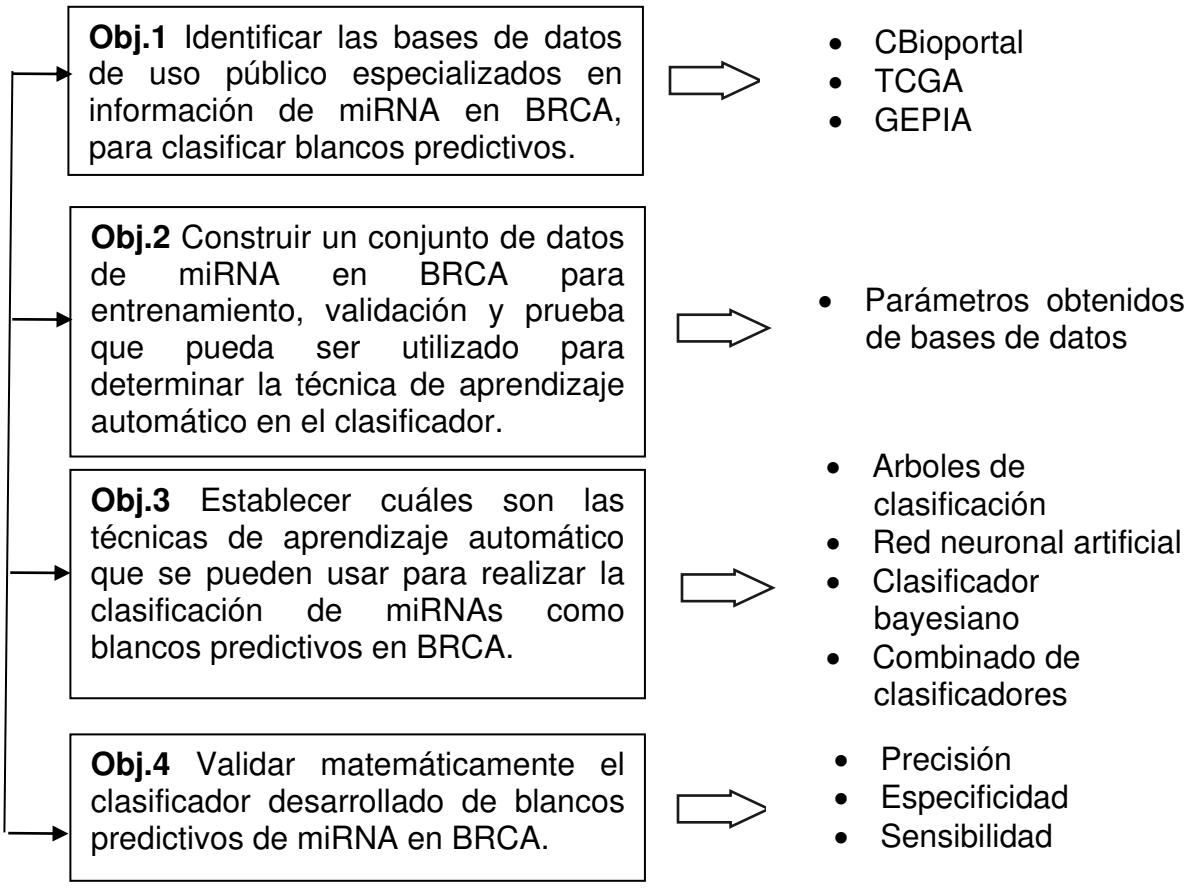


Figura 3

Estrategia bioinformática del modelo en aprendizaje automático para predicción de blancos de miRNAs de BRCA.Fuente: Elaboración propia.

Obj.1 Identificar bases de datos de uso público

Para este punto ya se tienen identificadas algunas bases de datos de uso público entre las cuales están:

a) cBioPortal

b) TCGA
c) GEPIA
d) mirNET

Se obtuvieron estas fuentes de datos realizando una búsqueda bibliográfica en el estado del arte en la cual se tomaron en cuenta:

1. Bases de datos de acceso abierto.
2. Base de datos con información de cáncer de mama.
3. Bases de datos con información de expresión diferencial de miRNAS,

De acuerdo con los criterios de inclusión antes mencionados se optó por seleccionar la plataforma cBioPortal como la fuente de datos para trabajar en el desarrollo de este proyecto.

La siguiente descripción: Secuenciación del exoma completo (510 muestras con normales emparejadas), matrices de número de copias de ADN genómico, metilación del ADN, matrices de RNA mensajero, secuenciación de microRNA y análisis de matrices de proteínas en fase inversa en 825 muestras de cáncer de mama primario. Proyecto de carcinoma invasivo de mama del Atlas del Genoma del Cáncer (TCGA) (Cancer Genome Atlas Network, 2012). Esta relacionada con el tipo del base de datos obtenido.

Para llevar a cabo este objetivo se seleccionó el conjunto de datos nombrado como: brca_tcga_pub_clinical_data y data_mirna, los cuales contienen un total de 825 muestras de cancer primario y 398 miRNA (nombres de miRNAs) y 300 muestras (expresión de miRNA por paciente) respectivamente.

Obj.2 Construir un conjunto de datos de miRNA en BRCA.

Para cumplir con los requerimientos en este tipo de investigaciones se deberán contemplar los siguientes aspectos en nuestro conjunto de datos:

- Que tengan replicas ya que esto puede ayudar en la precisión del modelo.
- Revisar si los datos con los que vamos a trabajar están normalizados.

- Información de la expresión de miRNAs.

El conjunto de datos obtenido cuenta con los criterios ya mencionados, para el caso de que se desee consultar algunas herramientas conocer sobre la expresión diferencial puede revisar la siguiente referencia (Seyednasrollah et al., 2015).

Una vez que se han identificado los datos con los que se trabajará, se procedió a crear una base de datos de microRNA en cáncer de mama (BRCA). Esta base de datos debe estar estructurada de manera que se puedan realizar cálculos de valor p y cambio de pliegue, los cuales requieren que los datos estén divididos en dos grupos. Para cumplir con este requisito, se crearon dos grupos de datos utilizando como referencia el subtipo de cáncer en el archivo brca_tcga_pub_clinical_data y la expresión de cada microRNA en el archivo data_mirna. Los subtipos moleculares de cáncer de mama seleccionados fueron basalike y luminal, lo que resultó en un total de 398 microARNs (filas) y 165 muestras de pacientes (columnas), las cuales se utilizaron como variables dependientes e independientes respectivamente.

Aunado a lo anterior se forma una matriz de expresión de miRNAs donde M es una matriz de tamaño n x m donde n es el número de miRNAs y m es el número de muestras o pacientes. Cada elemento de la matriz, denotado como M(i, j), representa la expresión del miRNA i en la muestra j, y se representa de la siguiente forma:

M = | M(1,1) M(1,2) M(1,3) ··· M(1,166) |
| M(2,1) M(2,2) M(2,3) ··· M(2,166) |
.
.
.
| M(398,1) M(398,2) M(398,3) ··· M(398,166) |

Cada elemento M(i, j) puede contener un valor numérico que indica la expresión del miRNA en la muestra correspondiente. Esta representación permite realizar

operaciones y análisis sobre la matriz de expresión de miRNAs, como cálculos estadísticos, agrupaciones.

Para el entrenamiento del modelo, se utilizó el 80% del conjunto de datos generados, mientras que el 20% restante se usó para validación y prueba. Es crucial asegurar la precisión en la obtención de estos datos, ya que de esta fase depende el funcionamiento y selección de la técnica de aprendizaje automático a emplear.

Obj.3 Establecer cuáles son las técnicas de aprendizaje automático.

En esta etapa se revisaron las siguientes opciones de técnicas de aprendizaje propuestas:

a) Arboles de clasificación.
b) Red neuronal artificial.
c) Clasificador bayesiano.
d) Combinado de clasificadores.

Para el desarrollo de este objetivo, se llevó a cabo una búsqueda exhaustiva en el estado del arte utilizando la técnica de revisión sistemática. Se priorizaron aquellas investigaciones relacionadas con el tema de estudio que incluyeran el uso de métodos de aprendizaje automático. Los resultados obtenidos se presentan en la Tabla 1 de la sección de Antecedentes.

El análisis de los artículos seleccionados, enfocado en el rendimiento y utilizando la precisión como métrica principal, concluyó que los estudios que utilizaron el método de bosques aleatorios obtuvieron una mayor precisión en términos de rendimiento. Además, se observó que el método de máquina de soporte vectorial también tuvo un rendimiento ligeramente inferior al método de bosques aleatorios.

En base a este análisis, se decidió llevar a cabo la experimentación inicial utilizando el método de bosques aleatorios, dadas sus destacadas métricas de rendimiento según la revisión sistemática.

Obj.4 Validar matemáticamente el clasificador

En esta etapa final, se llevó a cabo una rigurosa validación matemática del clasificador que se desarrolló para identificar blancos predictivos de miRNA en BRCA. Además, se implementó un proceso de entrenamiento exhaustivo para evaluar el rendimiento del modelo, teniendo en cuenta los siguientes aspectos clave:

1. Evaluación de Métricas de Desempeño: Se calcularon diversas métricas de desempeño, como la precisión, sensibilidad, especificidad, F1-score y el área bajo la curva ROC. Estas métricas permiten tener una visión completa del rendimiento del modelo y su capacidad para distinguir entre muestras positivas y negativas.
2. Matriz de Confusión: Se calculó la matriz de confusión para evaluar la cantidad de verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos obtenidos por el clasificador. Esto proporciona información detallada sobre el tipo de errores que comete el modelo y ayuda a tomar decisiones informadas para mejorarlo.
3. Comparación con Otros Modelos: Se realizaron comparaciones con otros algoritmos de clasificación, como SVM, para determinar si el clasificador de *Random Forest* es la mejor opción para este problema particular.

Esta fase de evaluación y entrenamiento permitió garantizar la robustez y eficacia del modelo propuesto para la clasificación de miRNA en BRCA, proporcionando una valiosa herramienta para futuros estudios de investigación y aplicaciones clínicas en la detección temprana y el tratamiento personalizado del cáncer de mama.

Materiales.

Software y Herramientas:

- Se genero el código en RStudio, 2022.02.1 Build 461,© 2009-2022 RStudio, PBC, para Windows.
- Las librerías utilizadas para implementar el algoritmo fueron:
  - ✓ library(caret)
  - ✓ library(randomForest)
  - ✓ library(dplyr)
  - ✓ library(pheatmap)
  - ✓ library(RColorBrewer)
  - ✓ library(pROC)
  - ✓ library(grid)

CAPITULO III

En este capítulo, se presentan los resultados obtenidos mediante el análisis y procesamiento de los datos recopilados en el marco de la presente investigación. El objetivo de este estudio ha sido desarrollar y validar un clasificador de blancos predictivos de miRNA en BRCA, con el propósito de contribuir al avance del conocimiento en la identificación de biomarcadores potenciales para el cáncer de mama.

El capítulo inicia con una descripción detallada de los resultados obtenidos en cada etapa del proceso de clasificación, incluyendo la selección de miRNAs relevantes, la configuración del modelo de clasificación y la evaluación del rendimiento del clasificador. A lo largo de esta sección, se presentan tablas, gráficos y figuras que facilitan la visualización y comprensión de los datos obtenidos.

Asimismo, se lleva a cabo una interpretación detallada de los resultados, destacando los hallazgos más relevantes y su trascendencia en el contexto de la investigación. Se comparan los resultados con estudios previos pertinentes y se analizan las implicaciones de los hallazgos en el ámbito de la oncología y la investigación biomédica.

Es importante destacar que este capítulo también expone las limitaciones del estudio, reconociendo las restricciones y posibles sesgos que pudieron haber afectado los resultados obtenidos. Estas consideraciones son fundamentales para una evaluación crítica y objetiva de los hallazgos.

Finalmente, se presentan conclusiones preliminares basadas en los resultados hasta el momento, con la comprensión de que el análisis y la interpretación aún están en curso. Además, se ofrecen perspectivas sobre futuras líneas de investigación que puedan surgir a partir de los resultados presentados.

Objetivos y preguntas de investigación.

Se exponen los objetivos y preguntas propuestos en esta investigación con el fin de recordar el propósito de estudio y tener una visión más clara de los resultados mostrados.

1. Identificar las bases de datos de uso público especializados.
2. Construir un conjunto de datos de miRNA en BRCA para entrenamiento.
3. Establecer cuáles son las técnicas de aprendizaje automático que se pueden usar para realizar la clasificación de miRNAs.
4. Validar matemáticamente el clasificador.
5. Desarrollar un algoritmo basado en un modelo de aprendizaje automático utilizando un clasificador para predicción de blancos predictivos de miRNAs en BRCA.
6. ¿Mediante el desarrollo de un modelo basado en aprendizaje automático es posible crear un algoritmo que permita identificar y predecir blancos de miRNAs en BRCA?

Resultados

Para esta sección se divide la descripción de los resultados en subapartados de acuerdo con los objetivos enlistados en el punto anterior:

1. Se seleccionó la base de datos contenida en la plataforma de cBioPortal (https://www.cbioportal.org/), específicamente, se accedió a los datos del programa TCGA (The Cancer Genome Atlas Program). Esta plataforma proporciona un paquete de datos que cumple con los requerimientos mencionados en el apartado de metodología. El recurso descargado fue un archivo comprimido denominado "brca_tcga_pub.tar", que contiene un total de 52 conjuntos de datos.

   Para nuestro estudio, se tomaron en consideración dos conjuntos de datos cruciales: "brca_tcga_pub_clinical_data" y "data_mirna". El primero consta de 825 registros, mientras que el segundo cuenta con 398 registros.

Estas bases de datos son de vital importancia para la presente investigación, ya que contienen información clínica y datos de expresión de miRNA (microRNA), respectivamente. La combinación de ambos conjuntos de datos permitió un análisis integral y profundo para la identificación de blancos predictivos de miRNA en pacientes con cáncer de mama (BRCA). Es importante destacar que la selección de estos conjuntos de datos se basó en su relevancia y la cantidad de información que proporcionan para los objetivos de investigación. También es importante mencionar que este recurso de datos está sustentado por un artículo de investigación (Cancer Genome Atlas Network, 2012) y su utilización asegura una sólida base de datos para el desarrollo y validación del clasificador propuesto.

2. Para construir una base de datos funcional para este estudio, se llevó a cabo una correlación entre los dos conjuntos de datos principales: "data_mirna" y "brca_tcga_pub_clinical_data". Esta correlación se basó en los registros de los campos denominados "PATIENT_ID" y "PAM50_SUBTYPE". Utilizando un filtro para identificar subtipos específicos, como "basalike" y "luminal", en comparación con los "PATIENT_ID" de "data_mirna", se obtuvo un conjunto de datos con 398 nombres de miRNAs en la columna denominada "Hugo_Symbol" y 165 columnas en total.

De estas 165 columnas, 124 corresponden al subtipo "basalike" y 41 al subtipo "luminal". Cada columna describe la expresión registrada por cada miRNA para cada "PATIENT_ID". Es relevante mencionar que esta correlación entre los dos subtipos permite calcular tanto el valor "p" (p_value) como el cambio de pliegue (fold_change), ya que ambos requieren dos grupos como parámetro en su fórmula.

Este proceso de correlación y selección de datos fue fundamental para obtener un conjunto de datos altamente significativo y relevante para nuestro análisis. La combinación de la información de expresión de miRNA

y los datos clínicos asociados con los subtipos de cáncer de mama (basalike y luminal) enriquecer este estudio y facilitó la identificación de blancos predictivos de miRNA en pacientes con cáncer de mama.

3. En la determinación de la técnica a emplear en el modelo de aprendizaje automático, se llevó a cabo una exhaustiva revisión sistemática de seis estudios altamente relevantes y relacionados con esta investigación. Estos estudios fueron seleccionados de acuerdo con criterios de inclusión y exclusión establecidos previamente, y se encuentran detallados en la Tabla 1 del apartado de Antecedentes.

   El análisis de los resultados reveló que el modelo de aprendizaje automático con mayor precisión fue el denominado "Bosques Aleatorios" (Random Forest, RF). Este modelo fue mencionado en cinco de los seis estudios revisados y mostró un promedio de precisión del 0.9. Por otro lado, otro método ampliamente citado fue la "Máquina de Soporte" (Support Vector Machine, SVM), que obtuvo un promedio de precisión del 0.87 y fue utilizado en tres de los estudios.

   Con base en esta información analizada y los resultados obtenidos, se tomó la decisión de utilizar el modelo de Bosques Aleatorios (RF) para llevar a cabo la experimentación con la base de datos construida previamente. Esta elección se sustenta en la alta precisión reportada por este modelo en diversos estudios relacionados, lo que nos permitirá obtener resultados confiables y significativos en nuestro análisis de blancos predictivos de miRNA en el cáncer de mama.

4. Para abordar el rendimiento y las métricas seleccionadas en este estudio, es importante destacar la relevancia de la revisión sistemática realizada sobre el tema. Esta revisión nos proporcionó valiosa información acerca de la evaluación del rendimiento en los estudios analizados, así como un panorama completo de las métricas utilizadas en este contexto.

Con base en los resultados y hallazgos obtenidos durante la revisión, se determinó que las métricas más adecuadas para nuestra evaluación serían precisión, especificidad, sensibilidad, curva ROC, área bajo la curva y F1-score. La selección de estas métricas fue fundamentada y respaldada por su relevancia en la medición del rendimiento de nuestro clasificador.

Al incluir estas métricas en este análisis, se buscó obtener una evaluación más sólida y fundamentada del desempeño del modelo de aprendizaje automático. Estas métricas permitieron medir tanto la capacidad del clasificador para identificar correctamente las clases de interés como su habilidad para discriminar entre ellas, lo cual es esencial para obtener resultados confiables y significativos en la predicción de blancos de miRNA en el cáncer de mama, para integrar lo mencionado se llevaron a cabo los siguientes procesos:

Detección de miRNAs estadísticamente expresados (MEE).

En la **Figura 4**, se destacan veintitrés miRNAs con expresión estadísticamente significativa, resultado del cálculo estadístico que considera tanto el valor de "p" como el cambio de pliegue (fold change). Estos miRNAs fueron seleccionados mediante la condición MEE, que establece que los miRNAs deben cumplir con la siguiente condición: p-value <= 0.05 & (fold_change >= 1.5 | fold_change <= -1.5).

La elección de estos miRNAs como base para determinar la eficiencia del clasificador en la detección de blancos de miRNAs es crucial. Estos miRNAs filtrados se consideraron como positivos en la predicción del modelo. Su inclusión en el análisis permitió una evaluación más precisa y enfocada de la capacidad del clasificador para identificar correctamente blancos de miRNAs de relevancia en el contexto del cáncer de mama.

Al resaltar estos miRNAs estadísticamente expresados y utilizarlos como conjunto de referencia, se buscó obtener una visión más precisa del rendimiento y la efectividad del clasificador en la identificación de blancos de miRNAs, lo cual constituye un paso crucial en el desarrollo de un enfoque eficiente y fiable para la detección de miRNAs relevantes en el cáncer de mama.

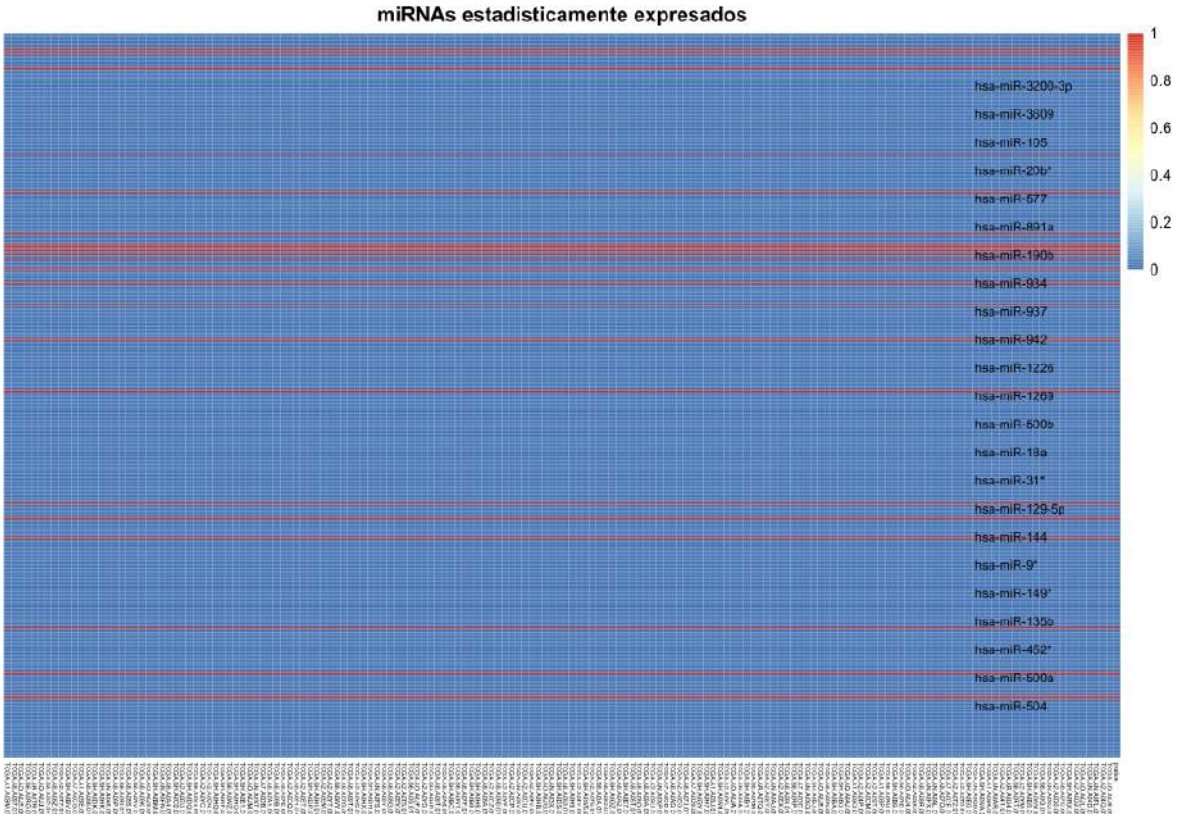


Figura 4. miRNAs estadísticamente expresados. Fuente: Elaboración propia.

Preparación del conjunto de Datos.

Después de obtener los miRNAs con expresión estadísticamente significativa, se procedió a enriquecer el conjunto de datos al introducir una variable objetivo-llamada "Interesante", la cual es binaria en naturaleza. Esta inclusión se efectuó con el propósito de elevar la precisión del modelo. La variable objetivo consta de dos clases: la clase 1 representa los miRNAs

blancos o significativos, mientras que la clase 0 se refiere a los miRNAs que no implican riesgos.

Se llevó a cabo la formación de dos conjuntos de datos, uno destinado al entrenamiento (80%) y el otro a las pruebas (20%). Dentro de los parámetros esenciales para la técnica de Bosques Aleatorios, se encuentra la recopilación de registros de predictores, los cuales abarcan 165 columnas que contienen la información referente a la expresión de 398 miRNAs. Asimismo, otro parámetro relevante es la variable objetivo, que abarca las clases a las que cada miRNA pertenece según el criterio establecido a través del valor de P y el cambio de pliegue.

La configuración del número de árboles, que fue de 100, y la implementación de una función denominada "importance", son otros aspectos determinantes en este proceso. La conjunción de estos parámetros fue fundamental en la fase de entrenamiento del modelo, a partir de la cual se generaron resultados consistentes y descriptivos. En total, 18 elementos proporcionaron información crucial para la validación exhaustiva del modelo.

La etapa subsiguiente implicó la utilización de la función "predict" para generar predicciones, aprovechando el modelo previamente entrenado y el conjunto de datos destinado a las pruebas. Este conjunto de resultados resultó ser de suma importancia en la evaluación del rendimiento integral del modelo.

Al emplear los resultados generados mediante la función "predict" y al contrastarlos con las etiquetas reales del conjunto de pruebas, obtenemos nuestra primera métrica de evaluación de rendimiento: una precisión de **0.9487179.** Este valor denota que nuestro modelo tiene un nivel

significativamente alto de exactitud en sus predicciones. La precisión, cuyo rango oscila entre 0 y 1, encuentra su máximo en 1, que indica una precisión absoluta.

En este contexto, una precisión cercana a **0.95** implica que alrededor del 95% de las predicciones realizadas por el modelo concuerdan con las etiquetas reales de los datos. Este resultado puede considerarse altamente satisfactorio y sugiere que el modelo está efectuando una clasificación con gran precisión en el conjunto de pruebas.

Otro elemento que nos permite conocer el rendimiento de nuestro modelo es la **matriz de confusión** la cual representa las predicciones realizadas y los datos obtenidos fueron:

	0	**1**
0	73	4
1	0	1

Las filas hacen referencia a las categorías verdaderas, mientras que las columnas hacen referencia a las categorías anticipadas por el modelo. Esto se puede entender de la siguiente manera:

Verdaderos Negativos (TN): El valor **73** representa la cantidad de instancias que el modelo clasificó correctamente como clase 0 (negativo) y la clase real también era clase 0.

Falsos Positivos (FP): El valor **0** representa la cantidad de instancias que el modelo clasificó incorrectamente como clase 1 (positivo) cuando la clase real era clase 0 (negativo).

Falsos Negativos (FN): El valor **4** representa la cantidad de instancias que el modelo clasificó incorrectamente como clase 0 (negativo) cuando la clase real era clase 1 (positivo).

Verdaderos Positivos (TP): El valor **1** representa la cantidad de instancias que el modelo clasificó correctamente como clase 1 (positivo) y la clase real también era clase 1.
La evaluación del rendimiento continuó con el cálculo de la métrica de **Especificidad.** El resultado obtenido fue de **1**, lo cual refleja que el modelo está logrando clasificar de manera precisa la totalidad de las instancias pertenecientes a la clase negativa, representando un acierto del 100% en la identificación de casos verdaderamente negativos.

En paralelo, se procedió al cálculo de la **Sensibilidad**. En este caso, el resultado obtenido fue de **0.2**. Este valor indica que el modelo enfrenta dificultades en la correcta clasificación de todas las instancias pertenecientes a la clase positiva. Es decir, se están presentando errores en la identificación de casos positivos. Esta observación es crucial, ya que una baja sensibilidad puede impactar negativamente en la detección de casos positivos, lo cual es especialmente relevante si la clase positiva tiene un valor crítico en el contexto del estudio.

En la continuación de la evaluación de rendimiento, se optó por calcular la métrica **F1_score**, arrojando un resultado de **0.3303571**. Esta métrica es particularmente útil en situaciones de desbalance entre clases, es decir, cuando una clase tiene significativamente más muestras que la otra. Un valor cercano a 1 indica un alto rendimiento del modelo, donde tanto la precisión como la sensibilidad son altas. En nuestro caso, esta puntuación dista de 1, lo que sugiere que el modelo presenta un desequilibrio entre precisión y sensibilidad. Esto señala que el modelo realiza predicciones precisas, pero no identifica correctamente la mayoría de las muestras positivas.

También se empleó la curva **ROC (Receiver Operating Characteristic curve, en inglés) Figura 5** con tres propósitos específicos: determinar el punto de corte en una escala continua donde se alcanza la máxima sensibilidad y especificidad, evaluar la capacidad discriminativa de la prueba diagnóstico al diferenciar entre individuos sanos y enfermos, y comparar la capacidad discriminativa de dos o más pruebas diagnósticos que expresan sus resultados en escalas continuas (Cerda & Cifuentes, 2012).

.

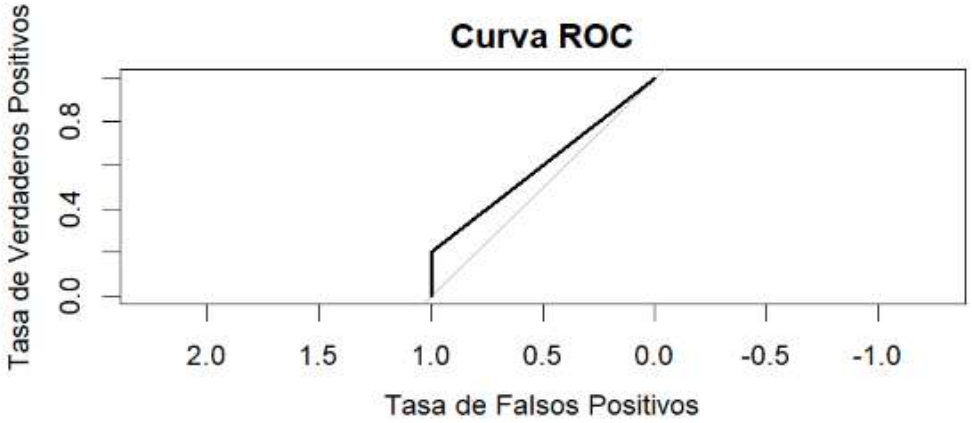


Figura 5. Curva ROC. Fuente: Elaboración propia.

El resultado obtenido en **AUC (Area Under Curve)** fue de **0.6**. lo que indica que el modelo exhibe cierta capacidad discriminativa, aunque no alcanza la perfección. Un AUC de 0.6 indica que las predicciones del modelo superan las expectativas al azar, pero aún existe espacio para mejoras. En otras palabras, se podría buscar optimizar el rendimiento del modelo con la aspiración de alcanzar un AUC de **1**, lo que representaría un desempeño óptimo en las predicciones.

En la predicción realizada mediante el modelo, se identificó el miRNA Mir-500b como un blanco predictivo relevante para el estudio del cáncer de mama, como se muestra en la **Figura 6**.

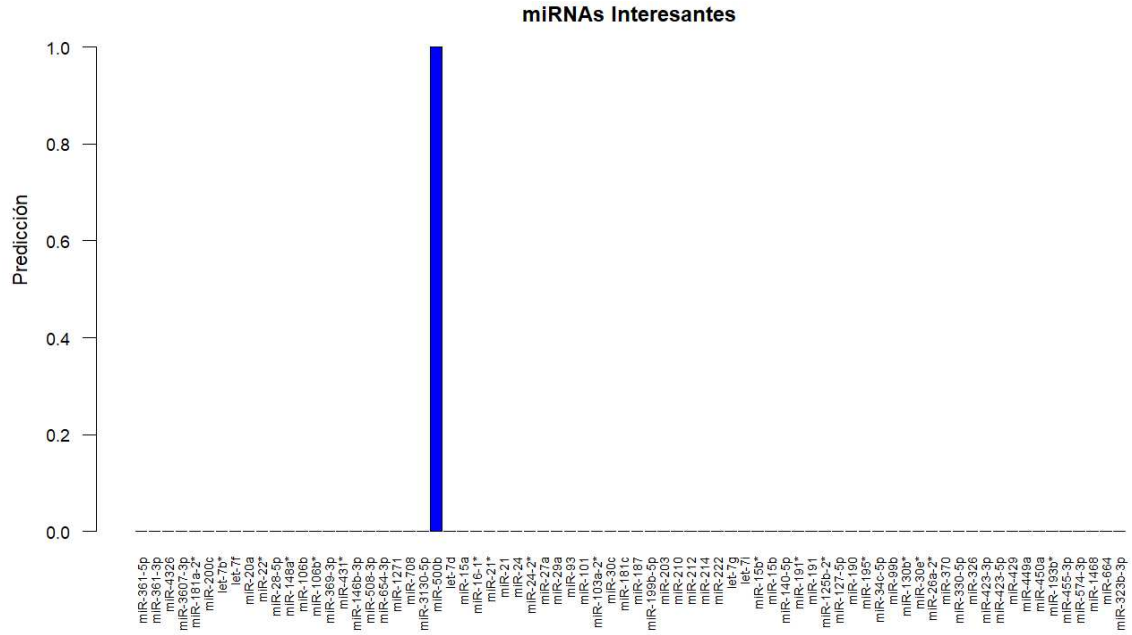


Figura 6. miRNA blancos predictivos/Interesantes. Fuente: Elaboración propia.

Además, se generó una representación gráfica de los resultados del conjunto de datos clasificados como correctos e incorrectos por el modelo. Esto se realizó con el propósito de confirmar que la clasificación de los miRNAs, tanto los de clase negativa como los de clase positiva según nuestra variable objetivo, se ajustara a los valores de p (valor P) y cambio de pliegue (fold_change). Estos resultados se ilustran en la Figura 7.

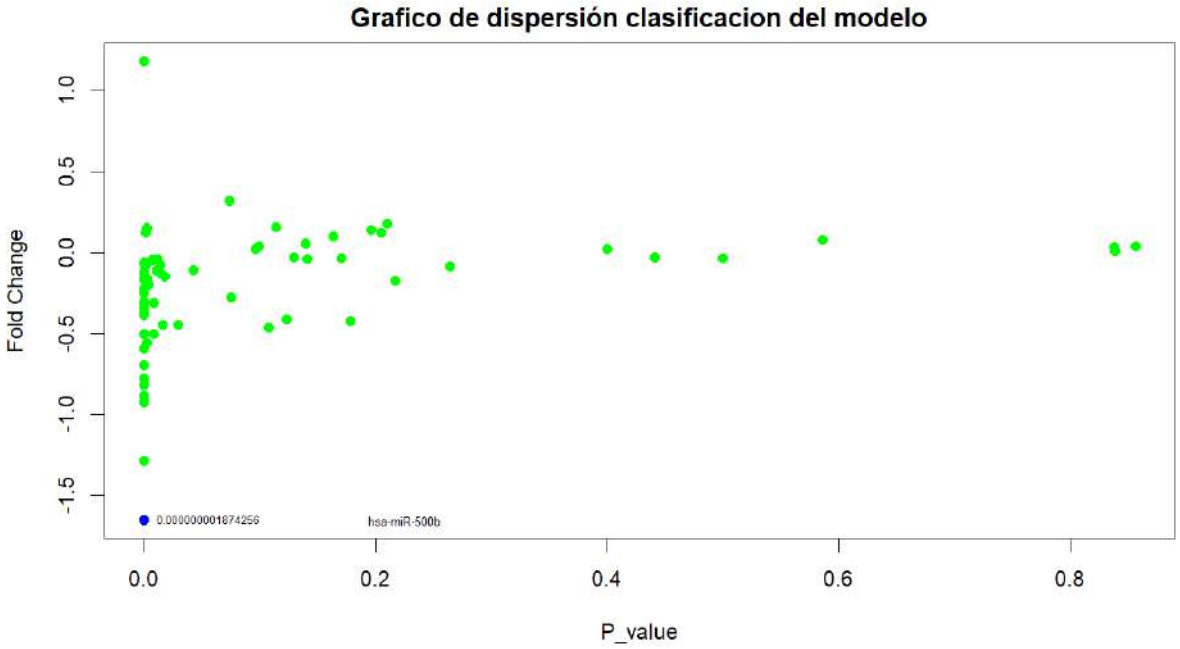


Figura 7. Clasificación del modelo. Fuente: Elaboración propia

En la Figura 7, los puntos de color azul oscuro representan las predicciones del modelo y validan la clasificación de la clase positiva, identificando el miRNA considerado como un blanco predictivo relevante para el estudio del cáncer de mama.

CAPITULO IV

Discusión

En este capítulo, se describe la fase de análisis y reflexión profunda sobre los resultados obtenidos a lo largo de esta investigación. Durante las secciones previas, se ha presentado una descripción detallada de la metodología utilizada, la recopilación y análisis de datos, así como los resultados obtenidos. Ahora, es el momento de explorar el significado de estos resultados, evaluar su relevancia y contribución al campo de estudio y, finalmente, abordar las implicaciones prácticas y las posibles direcciones futuras de investigación.

Para guiar esta discusión de manera efectiva, se abordarán los siguientes puntos clave:

a) **Interpretación de los Resultados** Los resultados obtenidos en relación con la hipótesis de investigación y los objetivos planteados fueron positivos, ya que se logró desarrollar un algoritmo basado en un modelo de aprendizaje automático capaz de predecir blancos de miRNAs asociados al cáncer de mama. No obstante, aún hay margen para mejorar la métrica de precisión del modelo, especialmente en la predicción de la clase positiva.

   Es importante destacar que estos resultados están estrechamente relacionados con la hipótesis original de esta investigación. Además, ofrecen la posibilidad de ser validados mediante pruebas de laboratorio, lo que podría acelerar y simplificar significativamente el proceso de identificación de miRNAs como blancos predictivos. En consecuencia, este tipo de investigaciones tiene el potencial de contribuir de manera significativa al estudio de miRNAs con implicaciones en el cáncer de mama.

   Además, al aprovechar las capacidades de la inteligencia artificial, se pueden abrir nuevas oportunidades para el desarrollo de investigaciones en este campo. La combinación de la bioinformática y el aprendizaje automático puede

llevar a avances significativos en la identificación y comprensión de los miRNAs relacionados con esta enfermedad.

b) **Comparación con la Literatura Existente** Con el fin de evaluar exhaustivamente el rendimiento del modelo en comparación con otros enfoques, se realizaron pruebas utilizando la técnica de Máquina de Soporte Vectorial (SVM). La elección de esta técnica se basó en los resultados previamente reportados en la Tabla 1, que destacaba tanto los Bosques Aleatorios como las Máquinas de Soporte Vectorial como enfoques comunes y efectivos en estudios similares.

   Sin embargo, al comparar el rendimiento de nuestros Bosques Aleatorios con SVM, se obtuvieron resultados notables. El modelo basado en de Bosques Aleatorios logró una impresionante métrica de precisión del 95%, representada por un valor de 0.9487179, lo que significa que el 95% de las predicciones coincidieron con las etiquetas reales. En contraste, el SVM alcanzó una precisión del 93.59%, ligeramente inferior, y no clasificó ninguna instancia en la clase positiva.

   Es importante señalar que nuestras métricas fueron valor de $p \leq 0.05$ y cambio de pliegue (Fold-change) de ≥ 1.5 y los estudios que se cotejaron se basaron en criterios estadísticos diferentes en comparación con el propuesto. Por ejemplo, en un estudio similar realizado por Yu et al. (2020), utilizaron un valor de p ajustado de ≤ 0.01 y Fold-change de ≥ 0.5 en un modelo de Bosques Aleatorios, logrando una precisión impresionante del 98%. Esta comparación resalta la influencia de los criterios estadísticos en el rendimiento de los modelos, y los resultados obtenidos, aunque ligeramente inferiores, se obtuvieron con parámetros diferentes.

Adicionalmente, el estudio de Sarkar et al. (2021) se centró en un solo parámetro estadístico, p-value < 0.05, y aplicó tanto Bosques Aleatorios como Máquinas de Soporte Vectorial. Lograron una precisión de aproximadamente el 76.58% con Bosques Aleatorios y el 74.91% con SVM. Aquí, el modelo superó significativamente estas cifras, alcanzando una precisión del 95% mediante Bosques Aleatorios. Estas comparaciones subrayan la efectividad de nuestro enfoque y la contribución de nuestros criterios de selección estadística en el rendimiento superior observado.

c) **Limitaciones y Consideraciones** Una de las limitaciones más destacadas en esta investigación se relaciona con la cantidad de muestras correspondientes a la clase positiva en el conjunto de datos utilizado. Esta limitación impacta en la capacidad de entrenamiento del modelo, lo que resulta en una menor capacidad para predecir todas las instancias de la clase positiva en el conjunto de prueba. Es crucial reconocer la importancia de mantener un equilibrio entre las muestras positivas y negativas para garantizar que el rendimiento del modelo alcance el nivel de precisión deseado.

   El desafío de contar con un número reducido de muestras positivas en comparación con las negativas es una dificultad común en estudios de clasificación, y es un aspecto crítico por considerar en investigaciones futuras. Abordar esta limitación podría implicar la adquisición de conjuntos de datos más grandes con una representación más equitativa de clases o la aplicación de técnicas de balanceo de clases durante el proceso de entrenamiento para mejorar la capacidad de generalización y precisión del modelo.

   Además de la desigualdad de muestras, es importante señalar que cualquier conjunto de datos puede contener ruido o características irrelevantes que pueden afectar el rendimiento del modelo. La identificación y mitigación de

estas características no triviales también representan un área importante para futuras investigaciones y mejoras en la metodología de nuestro estudio.

d) **Aportación académica en Ciencias de la Computación** Después de analizar los resultados de la revisión sistemática Tabla 1, se observó que los estudios revisados no hacen mención del uso de criterios estadísticos específicos, como los parámetros P value <= 0.05 y Fold change >= 1.5, <= -1.5, para determinar la expresión diferencial de miRNA y/o predicción de blancos predictivos de cáncer de mama.

   En este sentido, el presente estudio propone una metodología basada en los parámetros antes mencionados para la determinación de la expresión diferencial de miRNA, empleando un algoritmo con un enfoque basado en aprendizaje automático para la predicción de blancos predictivos de cáncer de mama. Esta aproximación ofrece una alternativa para el análisis de datos de expresión génica en el contexto de miRNA.

   Una recomendación específica que puede derivarse en esta investigación es que la base de datos con la cual se abordara un estudio, ya este previamente avalada y analizada.

e) **Direcciones Futuras de Investigación** Es importante mencionar que los resultados obtenidos no son definitivos es decir que se someterán a estudios que podrán determinar la importancia del hallazgo, así como procesos biológicos específicos que podrían estar involucrados. También es conveniente considerar que existen nuevas tecnologías o enfoques experimentales que podrían aplicarse para mejorar el resultado. Es sustancial considerar la posibilidad de llevar a cabo estudios de validación independientes para confirmar nuestros hallazgos actuales. Esto puede incluir colaboraciones con otros investigadores o la aplicación del método desarrollado a diferentes conjuntos de datos.

Conclusiones.

Según la implementación del algoritmo de predicción de blancos que se desarrolló y en función de los resultados obtenidos, se obtuvieron a diversas conclusiones significativas. En primer lugar, se identifica que existe margen para mejorar las métricas de rendimiento del modelo. Una de las estrategias para lograrlo podría ser la inclusión de un mayor número de registros en el proceso de entrenamiento de la clase positiva. Esta adición de datos contribuiría a un aprendizaje más eficaz del modelo, ya que actualmente existe un porcentaje de mejora en la clasificación de datos verdaderos positivos.

Asimismo, las métricas de precisión, sensibilidad y la curva de ROC todavía tienen un espacio considerable para el mejoramiento. La exploración de técnicas de validación cruzada podría fortalecer aún más el modelo, otorgándole una mayor robustez.

A pesar de los desafíos mencionados, el modelo de aprendizaje automático, que emplea la técnica de bosques aleatorios, logró clasificar y predecir miRNAs de la clase positiva. Esta capacidad es de suma importancia en la investigación del cáncer de mama, ya que identifica blancos predictivos cruciales. Además, los resultados obtenidos respaldan la hipótesis planteada en este estudio.

Es relevante destacar que, tras realizar pruebas con la técnica de máquina de soporte vectorial, mencionada en la literatura como otra opción en este tipo de investigaciones (como se muestra en la Tabla 1), no se obtuvieron predicciones para la clase positiva. Esto subraya aún más la efectividad de la elección de utilizar bosques aleatorios.

Si bien nuestros resultados son prometedores, también es conveniente reconocer que esta investigación presenta retos y áreas que requieren mejoras. Estos desafíos abren oportunidades para investigaciones futuras que podrían culminar en un modelo más robusto y eficiente.

Productividad Académica

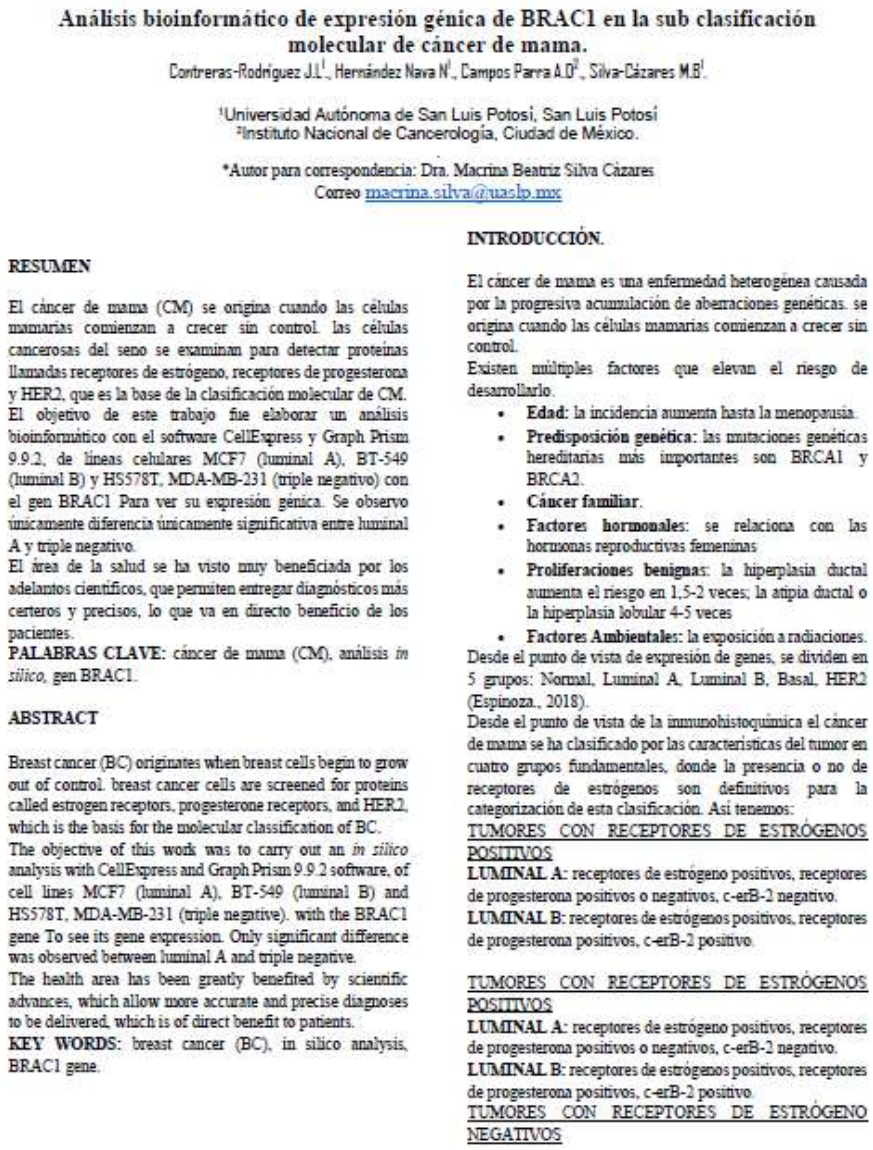
a) Publicaciones científicas

Revista Medicina de Torreón ISSN 1405 5422

Articulo " ANÁLISIS BIOINFORMÁTICO DE EXPRESIÓN GÉNICA DE BRAC 1 EN LA SUB-CLASIFICACIÓN MOLECULAR DE CÁNCER DE MAMA "

Autores

LI Jorge Alberto Contreras Rodríguez, Dra. Nereyda Hernández Nava, Dra. Alma Delia Campos Parra, Dra. Macrina Beatriz Silva Cázares



Análisis bioinformático de expresión génica de BRAC1 en la sub clasificación molecular de cáncer de mama.

Contreras-Rodríguez J.L[1], Hernández Nava N[1], Campos Parra A.D[2], Silva-Cázares M.B[1].

[1]Universidad Autónoma de San Luis Potosí, San Luis Potosí
[2]Instituto Nacional de Cancerología, Ciudad de México.

*Autor para correspondencia: Dra. Macrina Beatriz Silva Cázares
Correo macrina.silva@uaslp.mx

RESUMEN

El cáncer de mama (CM) se origina cuando las células mamarias comienzan a crecer sin control. las células cancerosas del seno se examinan para detectar proteínas llamadas receptores de estrógeno, receptores de progesterona y HER2, que es la base de la clasificación molecular de CM.
El objetivo de este trabajo fue elaborar un análisis bioinformático con el software CellExpress y Graph Prism 9.9.2, de líneas celulares MCF7 (luminal A), BT-549 (luminal B) y HS578T, MDA-MB-231 (triple negativo) con el gen BRAC1 Para ver su expresión génica. Se observo únicamente diferencia únicamente significativa entre luminal A y triple negativo.
El área de la salud se ha visto muy beneficiada por los adelantos científicos, que permiten entregar diagnósticos más certeros y precisos, lo que va en directo beneficio de los pacientes.
PALABRAS CLAVE: cáncer de mama (CM), análisis *in silico*, gen BRAC1.

ABSTRACT

Breast cancer (BC) originates when breast cells begin to grow out of control. breast cancer cells are screened for proteins called estrogen receptors, progesterone receptors, and HER2, which is the basis for the molecular classification of BC.
The objective of this work was to carry out an *in silico* analysis with CellExpress and Graph Prism 9.9.2 software, of cell lines MCF7 (luminal A), BT-549 (luminal B) and HS578T, MDA-MB-231 (triple negative). with the BRAC1 gene To see its gene expression. Only significant difference was observed between luminal A and triple negative.
The health area has been greatly benefited by scientific advances, which allow more accurate and precise diagnoses to be delivered, which is of direct benefit to patients.
KEY WORDS: breast cancer (BC), in silico analysis, BRAC1 gene.

INTRODUCCIÓN.

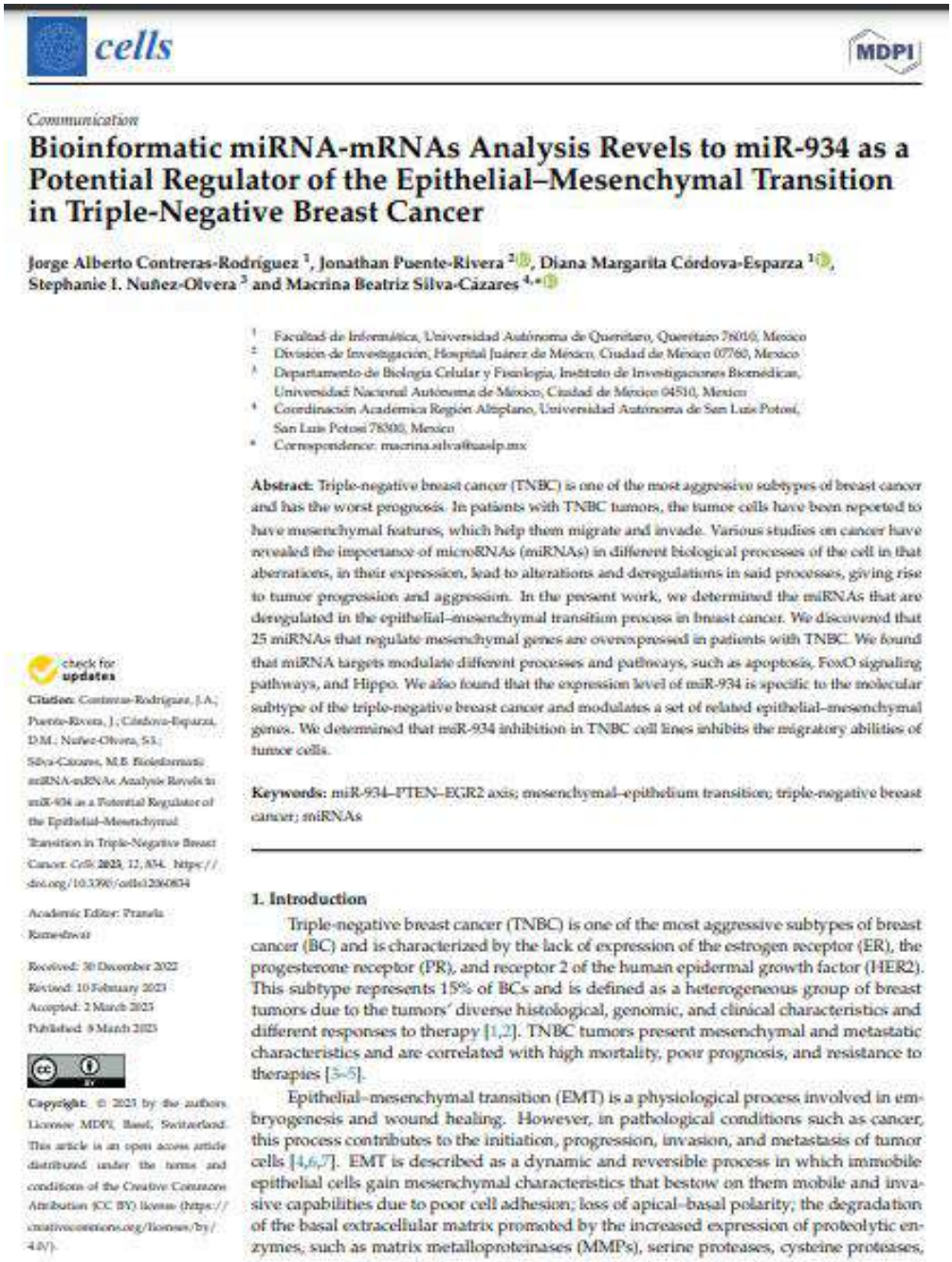
El cáncer de mama es una enfermedad heterogénea causada por la progresiva acumulación de aberraciones genéticas. se origina cuando las células mamarias comienzan a crecer sin control.
Existen múltiples factores que elevan el riesgo de desarrollarlo.

- **Edad:** la incidencia aumenta hasta la menopausia.
- **Predisposición genética:** las mutaciones genéticas hereditarias más importantes son BRCA1 y BRCA2.
- **Cáncer familiar.**
- **Factores hormonales:** se relaciona con las hormonas reproductivas femeninas
- **Proliferaciones benignas:** la hiperplasia ductal aumenta el riesgo en 1,5-2 veces; la atipia ductal o la hiperplasia lobular 4-5 veces
- **Factores Ambientales:** la exposición a radiaciones.

Desde el punto de vista de expresión de genes, se dividen en 5 grupos: Normal, Luminal A, Luminal B, Basal, HER2 (Espinoza., 2018).
Desde el punto de vista de la inmunohistoquímica el cáncer de mama se ha clasificado por las características del tumor en cuatro grupos fundamentales, donde la presencia o no de receptores de estrógenos son definitivos para la categorización de esta clasificación. Así tenemos:
TUMORES CON RECEPTORES DE ESTRÓGENOS POSITIVOS
LUMINAL A: receptores de estrógeno positivos, receptores de progesterona positivos o negativos, c-erB-2 negativo.
LUMINAL B: receptores de estrógenos positivos, receptores de progesterona positivos, c-erB-2 positivo.

TUMORES CON RECEPTORES DE ESTRÓGENOS POSITIVOS
LUMINAL A: receptores de estrógeno positivos, receptores de progesterona positivos o negativos, c-erB-2 negativo.
LUMINAL B: receptores de estrógenos positivos, receptores de progesterona positivos, c-erB-2 positivo.
TUMORES CON RECEPTORES DE ESTRÓGENO NEGATIVOS

Contreras-Rodríguez, J. A., Puente-Rivera, J., Córdova-Esparza, D. M., Nuñez-Olvera, S. I., & Silva-Cázares, M. B. (2023). **Bioinformatic miRNA-mRNAs analysis revels to miR-934 as a potential regulator of the epithelial–mesenchymal transition in triple-negative breast cancer**. *Cells (Basel, Switzerland)*, *12*(6), 834. https://doi.org/10.3390/cells12060834



cells

MDPI

Communication

Bioinformatic miRNA-mRNAs Analysis Revels to miR-934 as a Potential Regulator of the Epithelial–Mesenchymal Transition in Triple-Negative Breast Cancer

Jorge Alberto Contreras-Rodríguez [1], Jonathan Puente-Rivera [2], Diana Margarita Córdova-Esparza [1], Stephanie I. Nuñez-Olvera [3] and Macrina Beatriz Silva-Cázares [4,*]

1 Facultad de Informática, Universidad Autónoma de Querétaro, Querétaro 76010, Mexico
2 División de Investigación, Hospital Juárez de México, Ciudad de México 07760, Mexico
3 Departamento de Biología Celular y Fisiología, Instituto de Investigaciones Biomédicas, Universidad Nacional Autónoma de México, Ciudad de México 04510, Mexico
4 Coordinación Académica Región Altiplano, Universidad Autónoma de San Luis Potosí, San Luis Potosí 78300, Mexico
* Correspondence: macrina.silva@uaslp.mx

Abstract: Triple-negative breast cancer (TNBC) is one of the most aggressive subtypes of breast cancer and has the worst prognosis. In patients with TNBC tumors, the tumor cells have been reported to have mesenchymal features, which help them migrate and invade. Various studies on cancer have revealed the importance of microRNAs (miRNAs) in different biological processes of the cell in that aberrations, in their expression, lead to alterations and deregulations in said processes, giving rise to tumor progression and aggression. In the present work, we determined the miRNAs that are deregulated in the epithelial–mesenchymal transition process in breast cancer. We discovered that 25 miRNAs that regulate mesenchymal genes are overexpressed in patients with TNBC. We found that miRNA targets modulate different processes and pathways, such as apoptosis, FoxO signaling pathways, and Hippo. We also found that the expression level of miR-934 is specific to the molecular subtype of the triple-negative breast cancer and modulates a set of related epithelial–mesenchymal genes. We determined that miR-934 inhibition in TNBC cell lines inhibits the migratory abilities of tumor cells.

Keywords: miR-934–PTEN–EGR2 axis; mesenchymal–epithelium transition; triple-negative breast cancer; miRNAs

check for updates

Citation: Contreras-Rodríguez, J.A.; Puente-Rivera, J.; Córdova-Esparza, D.M.; Nuñez-Olvera, S.I.; Silva-Cázares, M.B. Bioinformatic miRNA-mRNAs Analysis Revels to miR-934 as a Potential Regulator of the Epithelial–Mesenchymal Transition in Triple-Negative Breast Cancer. *Cells* **2023**, *12*, 834. https://doi.org/10.3390/cells12060834

Academic Editor: Pranela Rameshwar

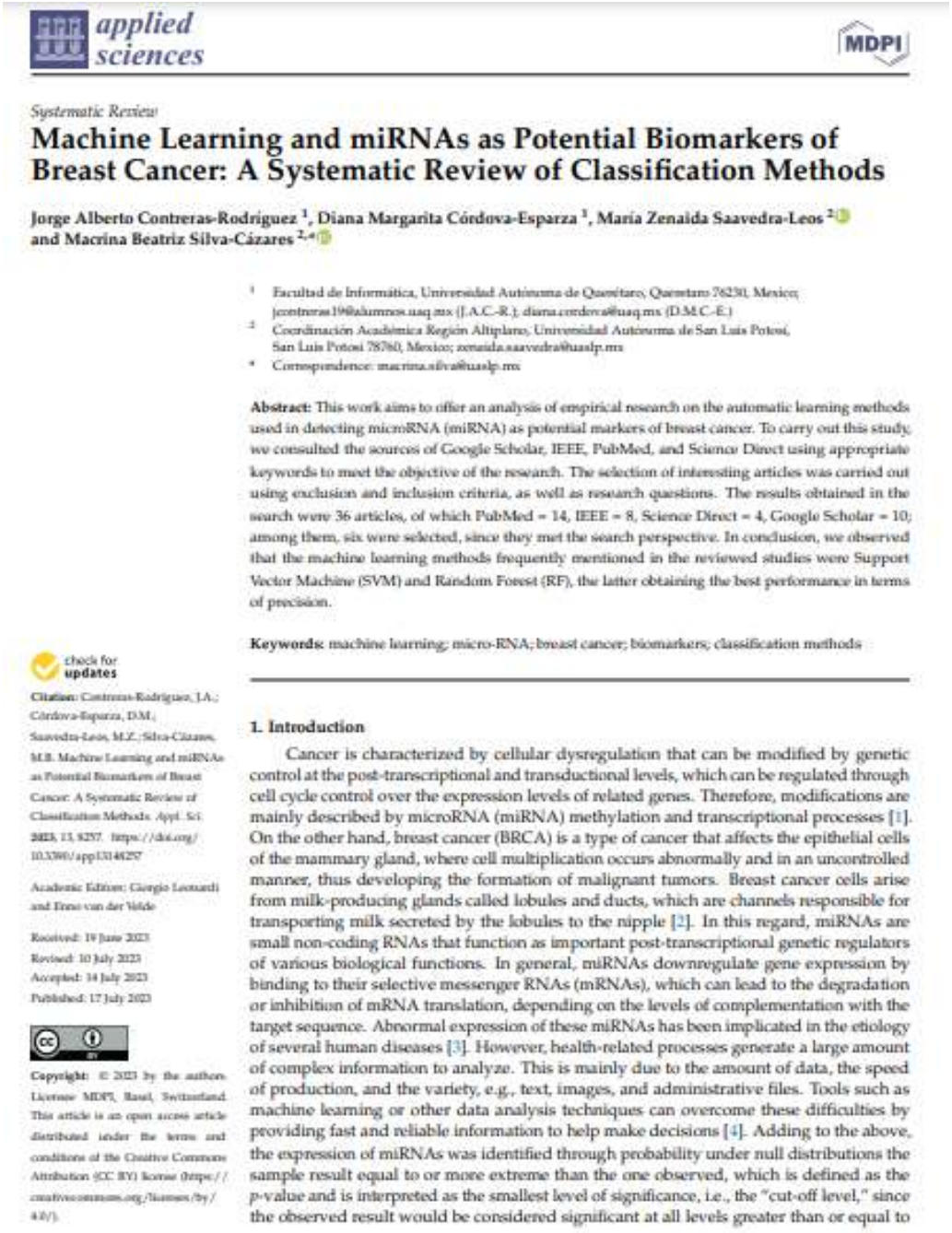
Received: 30 December 2022
Revised: 10 February 2023
Accepted: 2 March 2023
Published: 8 March 2023

1. Introduction

Triple-negative breast cancer (TNBC) is one of the most aggressive subtypes of breast cancer (BC) and is characterized by the lack of expression of the estrogen receptor (ER), the progesterone receptor (PR), and receptor 2 of the human epidermal growth factor (HER2). This subtype represents 15% of BCs and is defined as a heterogeneous group of breast tumors due to the tumors' diverse histological, genomic, and clinical characteristics and different responses to therapy [1,2]. TNBC tumors present mesenchymal and metastatic characteristics and are correlated with high mortality, poor prognosis, and resistance to therapies [3–5].

Epithelial–mesenchymal transition (EMT) is a physiological process involved in embryogenesis and wound healing. However, in pathological conditions such as cancer, this process contributes to the initiation, progression, invasion, and metastasis of tumor cells [4,6,7]. EMT is described as a dynamic and reversible process in which immobile epithelial cells gain mesenchymal characteristics that bestow on them mobile and invasive capabilities due to poor cell adhesion; loss of apical–basal polarity; the degradation of the basal extracellular matrix promoted by the increased expression of proteolytic enzymes, such as matrix metalloproteinases (MMPs), serine proteases, cysteine proteases,

Contreras-Rodríguez, J. A., Córdova-Esparza, D. M., Saavedra-Leos, M. Z., & Silva-Cázares, M. B. (2023). **Machine Learning and miRNAs as Potential Biomarkers of Breast Cancer: A Systematic Review of Classification Methods**. *Applied Sciences*, *13*(14).



applied sciences

MDPI

Systematic Review

Machine Learning and miRNAs as Potential Biomarkers of Breast Cancer: A Systematic Review of Classification Methods

Jorge Alberto Contreras-Rodríguez [1], Diana Margarita Córdova-Esparza [1], María Zenaida Saavedra-Leos [2] and Macrina Beatriz Silva-Cázares [2,*]

[1] Facultad de Informática, Universidad Autónoma de Querétaro, Querétaro 76230, Mexico; jcontreras19@alumnos.uaq.mx (J.A.C.-R.); diana.cordova@uaq.mx (D.M.C.-E.)
[2] Coordinación Académica Región Altiplano, Universidad Autónoma de San Luis Potosí, San Luis Potosí 78700, Mexico; zenaida.saavedra@uaslp.mx
* Correspondence: macrina.silva@uaslp.mx

Abstract: This work aims to offer an analysis of empirical research on the automatic learning methods used in detecting microRNA (miRNA) as potential markers of breast cancer. To carry out this study, we consulted the sources of Google Scholar, IEEE, PubMed, and Science Direct using appropriate keywords to meet the objective of the research. The selection of interesting articles was carried out using exclusion and inclusion criteria, as well as research questions. The results obtained in the search were 36 articles, of which PubMed = 14, IEEE = 8, Science Direct = 4, Google Scholar = 10; among them, six were selected, since they met the search perspective. In conclusion, we observed that the machine learning methods frequently mentioned in the reviewed studies were Support Vector Machine (SVM) and Random Forest (RF), the latter obtaining the best performance in terms of precision.

Keywords: machine learning; micro-RNA; breast cancer; biomarkers; classification methods

Citation: Contreras-Rodríguez, J.A.; Córdova-Esparza, D.M.; Saavedra-Leos, M.Z.; Silva-Cázares, M.B. Machine Learning and miRNAs as Potential Biomarkers of Breast Cancer: A Systematic Review of Classification Methods. *Appl. Sci.* **2023**, *13*, 8257. https://doi.org/10.3390/app13148257

Academic Editor: Giorgio Leonardi and Enno van der Velde

Received: 19 June 2023
Revised: 10 July 2023
Accepted: 14 July 2023
Published: 17 July 2023

1. Introduction

Cancer is characterized by cellular dysregulation that can be modified by genetic control at the post-transcriptional and transductional levels, which can be regulated through cell cycle control over the expression levels of related genes. Therefore, modifications are mainly described by microRNA (miRNA) methylation and transcriptional processes [1]. On the other hand, breast cancer (BRCA) is a type of cancer that affects the epithelial cells of the mammary gland, where cell multiplication occurs abnormally and in an uncontrolled manner, thus developing the formation of malignant tumors. Breast cancer cells arise from milk-producing glands called lobules and ducts, which are channels responsible for transporting milk secreted by the lobules to the nipple [2]. In this regard, miRNAs are small non-coding RNAs that function as important post-transcriptional genetic regulators of various biological functions. In general, miRNAs downregulate gene expression by binding to their selective messenger RNAs (mRNAs), which can lead to the degradation or inhibition of mRNA translation, depending on the levels of complementation with the target sequence. Abnormal expression of these miRNAs has been implicated in the etiology of several human diseases [3]. However, health-related processes generate a large amount of complex information to analyze. This is mainly due to the amount of data, the speed of production, and the variety, e.g., text, images, and administrative files. Tools such as machine learning or other data analysis techniques can overcome these difficulties by providing fast and reliable information to help make decisions [4]. Adding to the above, the expression of miRNAs was identified through probability under null distributions the sample result equal to or more extreme than the one observed, which is defined as the *p*-value and is interpreted as the smallest level of significance, i.e., the "cut-off level," since the observed result would be considered significant at all levels greater than or equal to

b) <u>Participación en congreso</u>

Jorge Alberto Contreras Rodríguez, Macrina Beatriz Silva Cázares, Daniel Cantón Enríquez. "***Análisis bioinformático de genes VEGF/HER2 en cáncer de mama como factor pronóstico la supervivencia en la supervivencia clínica***" 2do Congreso Internacional de computación y tecnología educativa. Universidad Autónoma de Querétaro. Juriquilla, Querétaro. 20 de octubre del 2021.

Dra. Macrina Beatriz Silva Cazares, **Jorge Alberto Contreras Rodriguez**, Guillermo Eduardo Loera Bautista, Nereyda Hernandez Nava, Meile Martha Adriana Maza Calviño, Gabriela Alvarado Macias. "***In Silico* Identification of Potential Therapeutic Target of miRNAs in Triple-Negative Breast Cancer**" Congreso Internacional de Investigación de Academia Journals Puebla TecNM 2023.

c) Participación en coloquios

1ER COLOQUIO NACIONAL EN COMPUTACIÓN Y APLICACIONES TECNOLÓGICAS, "**Modelo en machine learning para predicción de blancos de micro-RNAs de cáncer de mama**" Universidad Autónoma de Querétaro, Facultad de informática, **Jorge Alberto Contreras Rodriguez.** Diciembre 2021.

2DO. COLOQUIO NACIONAL DE INVESTIGACION EN COMPUTACIÓN Y APLICACIONES TECNOLÓGICAS, “**Modelo en machine learning para predicción de blancos de micro-RNAs de cáncer de mama**” Universidad Autónoma de Querétaro, Facultad de informática, **Jorge Alberto Contreras Rodriguez.** Junio 2022.

3ER. COLOQUIO NACIONAL DE INVESTIGACION EN COMPUTACIÓN Y APLICACIONES TECNOLÓGICAS, “**Modelo en machine learning para predicción de blancos de micro-RNAs de cáncer de mama**” Universidad Autónoma de Querétaro, Facultad de informática, **Jorge Alberto Contreras Rodriguez.** Junio 2023.


UNIVERSIDAD AUTÓNOMA DE QUERÉTARO
FACULTAD DE INFORMÁTICA

CNICyAT
Coloquio Nacional de Investigación en Computación y Aplicaciones Tecnológicas
12 y 13 de Junio

UNIVERSIDAD AUTÓNOMA DE QUERÉTARO
FACULTAD DE INFORMÁTICA
OTORGA LA PRESENTE
CONSTANCIA
A
Jorge Alberto Contreras Rodríguez
Por su participación como Ponente en el Coloquio Nacional de Investigación en Computación y Aplicaciones Tecnológicas, llevado a cabo los días 12 y 13 de junio de 2023, Juriquilla, Querétaro, México

DRA. GABRIELA XICOTÉNCATL RAMÍREZ
DIRECTORA DE LA FACULTAD DE INFORMÁTICA

DRA. ANA MARCELA HERRERA NAVARRO
JEFA DE POSGRADO DE LA FACULTAD DE INFORMÁTICA

DRA. VERÓNICA LÓPEZ MARTÍNEZ
COORDINADORA CNICYAT 2023

Referencias

Ab Mutalib, N.-S., Sulaiman, S. A., & Jamal, R. (2019). Computational tools for microRNA target prediction. En *Computational Epigenetics and Diseases* (pp. 79–105). Elsevier.

Adams, J. M., & Cory, S. (2007). The Bcl-2 apoptotic switch in cancer development and therapy. *Oncogene*, *26*(9), 1324–1337. https://doi.org/10.1038/sj.onc.1210220

Albíztegui, R. E. O., Balboa, P. G., Lozada, L. E. G., Chávez, S. A. G., Corona, R. E. F., & Patraca, D. L. B. (2014). El papel de los microRNAs (miRNAs) en el cáncer de mama. *An Med* (Mex), 59(4), 267-270.

Almeida, M. I., Reis, R. M., & Calin, G. A. (2011). MicroRNA history: discovery, recent applications, and next frontiers. *Mutation Research*, *717*(1–2), 1–8. https://doi.org/10.1016/j.mrfmmm.2011.03.009

Ambros, V., & Horvitz, H. R. (1987). The lin-14 locus of Caenorhabditis elegans controls the time of expression of specific postembryonic developmental events. *Genes & Development*, *1*(4), 398–414. https://doi.org/10.1101/gad.1.4.398

Andreini, P., Bonechi, S., Bianchini, M., & Geraci, F. (2022). MicroRNA signature for interpretable breast cancer classification with subtype clue. *Journal of Computational Mathematics and Data Science*, *3*(100042), 100042. https://doi.org/10.1016/j.jcmds.2022.100042

Aprendizaje automático y datos de entrenamiento: lo que debes saber. (2022, junio 29). Ciberseguridad. https://ciberseguridad.com/guias/nuevas-tecnologias/machine-learning/datos-entrenamiento/

Berdasco, M., & Esteller, M. (2010). Aberrant epigenetic landscape in cancer: how cellular identity goes awry. *Developmental Cell*, *19*(5), 698–711. https://doi.org/10.1016/j.devcel.2010.10.005

Betel, D., Koppal, A., Agius, P., Sander, C., & Leslie, C. (2010). Comprehensive modeling of microRNA targets predicts functional non-conserved and non-

canonical sites. *Genome Biology*, *11*(8), R90. https://doi.org/10.1186/gb-2010-11-8-r90

Birney, E., Hudson, T. J., Green, E. D., & Gunter, C. (s/f). Toronto International Data Release Workshop Authors (2009) Prepublication data sharing. *Nature*, *461*, 168–170.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Breastcancer.org. (2021). [internet] Disponible en: https://www.breastcancer.org/es

Broughton, J. P., Lovci, M. T., Huang, J. L., Yeo, G. W., & Pasquinelli, A. E. (2016). Pairing beyond the seed supports MicroRNA targeting specificity. *Molecular Cell*, *64*(2), 320–333. https://doi.org/10.1016/j.molcel.2016.09.004

Cancer Genome Atlas Network. (2012). Comprehensive molecular portraits of human breast tumours. *Nature*, *490*(7418), 61–70. https://doi.org/10.1038/nature11412.

Cáncer de mama. (s/f). Who.int. Recuperado el 24 de abril de 2023, de https://www.who.int/es/news-room/fact-sheets/detail/breast-cancer

Catalanotto, C., Cogoni, C., Zardo, G. (2016). MicroRNA in control of gene expression: An overview of nuclear functions. *Int J Mol Sci*,17(10), 1712.

Centro Nacional de Equidad de Género y Salud Reproductiva. (s/f). *Información Estadística Cáncer de Mama*. gob.mx. Recuperado el 24 de abril de 2023, de https://www.gob.mx/salud%7ccnegsr/acciones-y-programas/informacion-estadistica-cancer-de-mama.

Cerda, J., & Cifuentes, L. (2012). Uso de curvas ROC en investigación clínica: Aspectos teórico-prácticos. *Revista Chilena de Infectologia: Organo Oficial de La Sociedad Chilena de Infectologia*, *29*(2), 138–141. https://doi.org/10.4067/s0716-10182012000200003

Chalfie, M., Horvitz, H. R., & Sulston, J. E. (1981). Mutations that lead to reiterations in the cell lineages of C. elegans. *Cell*, *24*(1), 59–69. https://doi.org/10.1016/0092-8674(81)90501-8

Colotta, F., Allavena, P., Sica, A., Garlanda, C., & Mantovani, A. (2009). Cancer-related inflammation, the seventh hallmark of cancer: links to genetic

instability. *Carcinogenesis*, *30*(7), 1073–1081. https://doi.org/10.1093/carcin/bgp127

Cómo el Machine learning puede ayudar en la medicina predictiva. (2018, mayo 7). BITAC. https://www.bitac.com/2018/05/07/machine-learning-medicina-predictiva/

Cruz, J. A., & Wishart, D. S. (2007). Applications of machine learning in cancer prediction and prognosis. *Cancer Informatics*, *2*, 59–77. https://doi.org/10.1177/117693510600200030

Davis-Dusenbery, B. N., & Hata, A. (2010). Mechanisms of control of microRNA biogenesis. *The Journal of Biochemistry*, *148*(4), 381–392. https://doi.org/10.1093/jb/mvq096

Doench, J. G., & Sharp, P. A. (2004). Specificity of microRNA target selection in translational repression. *Genes & Development*, *18*(5), 504–511. https://doi.org/10.1101/gad.1184404

Dyke, S. O., & Hubbard, T. J. (2011). Developing and implementing an institute-wide data sharing policy. *Genome Medicine*, *3*(9), 60. https://doi.org/10.1186/gm276

Esteller, M. (2007). Cancer epigenomics: DNA methylomes and histone-modification maps. *Nature Reviews. Genetics*, *8*(4), 286–298. https://doi.org/10.1038/nrg2005

Evan, G., & Littlewood, T. (1998). A matter of life and cell death. *Science (New York, N.Y.)*, *281*(5381), 1317–1322. https://doi.org/10.1126/science.281.5381.1317.

Fan, X., & Kurgan, L. (2015). Comprehensive overview and assessment of computational prediction of microRNA targets in animals. *Briefings in bioinformatics*, 16(5), 780–794. https://doi.org/10.1093/bib/bbu044

Ferguson, E. L., Sternberg, P. W., & Horvitz, R. (1987). Corrigendum: A genetic pathway for the specification of the vulval cell lineages of Caenorhabditis elegans. *Nature*, *327*(6117), 82–82. https://doi.org/10.1038/327082b0

Flier, J. S., Underhill, L. H., & Dvorak, H. F. (1986). Tumors: Wounds that do not heal. *The New England Journal of Medicine*, *315*(26), 1650–1659. https://doi.org/10.1056/nejm198612253152606

Fortin, S., Pathmasiri, S., Grintuch, R., & Deschênes, M. (2011). Access arrangements' for biobanks: a fine line between facilitating and hindering collaboration. *Public Health Genomics*, *14*, 104–114.

Friedländer, M. R., Lizano, E., Houben, A. J. S., Bezdan, D., Báñez-Coronel, M., Kudla, G., Mateu-Huertas, E., Kagerbauer, B., González, J., Chen, K. C., LeProust, E. M., Martí, E., & Estivill, X. (2014). Evidence for the biogenesis of more than 1,000 novel human microRNAs. *Genome Biology*, *15*(4), R57. https://doi.org/10.1186/gb-2014-15-4-r57

Fu, G., Brkić, J., Hayder, H., & Peng, C. (2013). MicroRNAs in human placental development and pregnancy complications. *International Journal of Molecular Sciences*, *14*(3), 5519–5544. https://doi.org/10.3390/ijms14035519

Garcia, D. M., Baek, D., Shin, C., Bell, G. W., Grimson, A., & Bartel, D. P. (2011). Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs. *Nature Structural & Molecular Biology*, *18*(10), 1139–1146. https://doi.org/10.1038/nsmb.2115

Gitter, D. M. (2010). The challenges of achieving open-source sharing of biobank data. *Biotechnology law report*, *29*(6), 623–635. https://doi.org/10.1089/blr.2010.9909

Gitter, D. M. (2010a). The challenges of achieving open-source sharing of biobank data. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1598400

Ha, M., & Kim, V. N. (2014). Regulation of microRNA biogenesis. *Nature Reviews. Molecular Cell Biology*, *15*(8), 509–524. https://doi.org/10.1038/nrm3838

Hanahan, D., & Folkman, J. (1996). Patterns and emerging mechanisms of the angiogenic switch during tumorigenesis. *Cell*, *86*(3), 353–364. https://doi.org/10.1016/s0092-8674(00)80108-7

Hanahan, D., & Weinberg, R. A. (2000). The hallmarks of cancer. *Cell*, *100*(1), 57–70. https://doi.org/10.1016/s0092-8674(00)81683-9

Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell*, 144(5),646–674. https://doi.org/10.1016/j.cell.2011.02.013

Hayes, J., Peruzzi, P. P., & Lawler, S. (2014). MicroRNAs in cancer: biomarkers, functions and therapy. *Trends in Molecular Medicine*, *20*(8), 460–469. https://doi.org/10.1016/j.molmed.2014.06.005

Horvitz, H. R., & Sulston, J. E. (1980). Isolation and genetic characterization of cell-lineage mutants of the nematode Caenorhabditis elegans. *Genetics*, *96*(2), 435–454. https://doi.org/10.1093/genetics/96.2.435

Joly, Y., Dove, E. S., Knoppers, B. M., Bobrow, M., & Chalmers, D. (2012). Data sharing in the post-genomic world: the experience of the International Cancer Genome Consortium (ICGC) Data Access Compliance Office (DACO). *PLoS Computational Biology*, *8*(7), e1002549. https://doi.org/10.1371/journal.pcbi.1002549

Joly, Y., Zeps, N., & Knoppers, B. M. (2011). Genomic databases access agreements: legal validity and possible sanctions. *Human Genetics*, *130*(3), 441–449. https://doi.org/10.1007/s00439-011-1044-3

Jones, P. A., & Baylin, S. B. (2007). The epigenomics of cancer. *Cell*, *128*(4), 683–692. https://doi.org/10.1016/j.cell.2007.01.029.

Kim, S. (2018). Computational Model for Predicting the Relationship Between Micro-RNAs and Their Target Messenger RNAs in Breast and Colon Cancers. *Cancer informatics*, 17. https://doi.org/10.1177/1176935118785145.

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, *13*, 8–17. https://doi.org/10.1016/j.csbj.2014.11.005

Lee, R. C., Feinbaum, R. L., & Ambros, V. (1993). The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell, 75(5), 843–854. https://doi.org/10.1016/0092-8674(93)90529-y

Lee, R., Feinbaum, R., & Ambros, V. (2004). A short history of a short RNA. *Cell*, *116*(2 Suppl), S89-92, 1 p following S96. https://doi.org/10.1016/s0092-8674(04)00035-2

Lewis, B. P., Shih, I.-H., Jones-Rhoades, M. W., Bartel, D. P., & Burge, C. B. (2003). Prediction of mammalian microRNA targets. *Cell*, *115*(7), 787–798. https://doi.org/10.1016/s0092-8674(03)01018-3

Lewis, B. P., Burge, C. B., & Bartel, D. P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, *120*(1), 15–20. https://doi.org/10.1016/j.cell.2004.12.035

Li, J., Chen, X., Huang, Q., Wang, Y., Xie, Y., Dai, Z., Zou, X., & Li, Z. (2020). Seq-SymRF: a random forest model predicts potential miRNA-disease associations based on information of sequences and clinical symptoms. *Scientific reports*, 10(1), 17901. https://doi.org/10.1038/s41598-020-75005-9

Li, S.-C., Chan, W.-C., Hu, L.-Y., Lai, C.-H., Hsu, C.-N., & Lin, W.-C. (2010). Identification of homologous microRNAs in 56 animal genomes. *Genomics*, *96*(1), 1–9. https://doi.org/10.1016/j.ygeno.2010.03.009

Liu, H., Yue, D., Zhang, L., Bai, Z., Lei, X., Gao, S.-J., & Huang, Y. (2008). A MACHINE LEARNING APPROACH FOR miRNA TARGET PREDICTION. *IEEE International Workshop on Genomic Signal Processing and Statistics*, *2008*, 1–3. https://doi.org/10.1109/GENSIPS.2008.4555655

Loh, H. Y., Norman, B. P., Lai, K. S., Rahman, N., Alitheen, N., & Osman, M. A. (2019). The Regulatory Role of MicroRNAs in Breast Cancer. *International journal of molecular sciences*, *20*(19), 4940. https://doi.org/10.3390/ijms20194940

Lowe, S. W., Cepero, E., & Evan, G. (2004). Intrinsic tumour suppression. *Nature*, *432*(7015), 307–315. https://doi.org/10.1038/nature03098

Lukasik, A., Wójcikowski, M., & Zielenkiewicz, P. (2016). Tools4miRs – one place to gather all the tools for miRNA analysis. *Bioinformatics (Oxford, England)*, *32*(17), 2722–2724. https://doi.org/10.1093/bioinformatics/btw189

Luo, J., Solimini, N. L., & Elledge, S. J. (2009). Principles of cancer therapy: Oncogene and non-oncogene addiction. *Cell*, *138*(4), 807. https://doi.org/10.1016/j.cell.2009.08.006

Mahen, E. M., Watson, P. Y., Cottrell, J. W., & Fedor, M. J. (2010). mRNA secondary structures fold sequentially but exchange rapidly in vivo. *PLoS Biology*, *8*(2), e1000307. https://doi.org/10.1371/journal.pbio.1000307

Makarova, J. A., Shkurnikov, M. U., Wicklein, D., Lange, T., Samatov, T. R., Turchinovich, A. A., & Tonevitsky, A. G. (2016). Intracellular and extracellular microRNA: An update on localization and biological role. *Progress in Histochemistry and Cytochemistry*, *51*(3–4), 33–49. https://doi.org/10.1016/j.proghi.2016.06.001

Marín, R. M., & Vaníček, J. (2011). Efficient use of accessibility in microRNA target prediction. *Nucleic Acids Research*, *39*(1), 19–29. https://doi.org/10.1093/nar/gkq768

Miranda, K. C., Huynh, T., Tay, Y., Ang, Y.-S., Tam, W.-L., Thomson, A. M., Lim, B., & Rigoutsos, I. (2006). A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. *Cell*, *126*(6), 1203–1217. https://doi.org/10.1016/j.cell.2006.07.031

Naorem, L. D., Muthaiyan, M., & Venkatesan, A. (2019). Identification of dysregulated miRNAs in triple negative breast cancer: A meta-analysis approach: NAOREM et al. *Journal of Cellular Physiology*, *234*(7), 11768–11779. https://doi.org/10.1002/jcp.27839

Navarro Ponz, A (2008). Análisis comparativo de la expresión de miRNAs en el desarrollo embrionario del colon, el cáncer colorectal y el linfoma de Hodgkin (Doctor). facultad de medicina. https://1library.co/document/z3ln137z-tesis-doctoral-facultad-de-medicina.html/

Negrini, S., Gorgoulis, V. G., & Halazonetis, T. D. (2010). Genomic instability-an evolving hallmark of cancer. *Nat. Rev. Mol. Cell Biol*, *11*, 220–228.

Netter, F. H. (2019). Atlas de Anatomía Humana (7a ed.). Elsevier. https://www.elsevier.com/books/atlas-de-anatomia-humana/netter/978-84-9113-468-8

Organización Mundial de la Salud. (2020). Un reporte sobre la salud. Recuperado de: https://www.who.int/es/news-room/fact-sheets/detail/breast-cancer#

Page S, F., Galon, J., Dieu-Nosjean, M. C., Tartour, E., Saute S-Fridman, C., & Fridman, W. H. (2010). Immune infiltration in human tumors: a prognostic factor that should not be ignored. *Oncogene*, *29*, 1093–1102.

Palsson, B. (2000). The challenges of *in silico* biology. *Nature Biotechnology*, *18*(11), 1147–1150. https://doi.org/10.1038/81125

Pasquinelli, A. E., Reinhart, B. J., Slack, F., Martindale, M. Q., Kuroda, M. I., Maller, B., Hayward, D. C., Ball, E. E., Degnan, B., Müller, P., Spring, J., Srinivasan, A., Fishman, M., Finnerty, J., Corbo, J., Levine, M., Leahy, P., Davidson, E., & Ruvkun, G. (2000). Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature*, *408*(6808), 86–89. https://doi.org/10.1038/35040556

Paul, P., Chakraborty, A., Sarkar, D., Langthasa, M., Rahman, M., Bari, M., Singha, R. K. S., Malakar, A. K., & Chakraborty, S. (2018). Interplay between miRNAs and human diseases. *Journal of Cellular Physiology*, *233*(3), 2007–2018. https://doi.org/10.1002/jcp.25854

Pedrero, V., Reynaldos-Grandón, K., Ureta-Achurra, J., & Cortez-Pinto, E. (2021). Generalidades del Machine Learning y su aplicación en la gestión sanitaria en Servicios de Urgencia. *Revista médica de Chile*, *149*(2), 248-254.

Pereira, F., Mitchell, T., Botvinick, M. (2009). Machine learning classifiers and fMRI: A tutorial overview, *NeuroImage. El sevier*, 45(1), 199-209.

Qiu, M., Fu, Q., Jiang, C., & Liu, D. (2020). Machine learning based network analysis determined clinically relevant miRNAs in breast cancer. *Frontiers in Genetics*, *11*, 615864. https://doi.org/10.3389/fgene.2020.615864

Rico-Rosillo, M. G., Vega-Robledo, G. B., & Oliva-Rico, D. (2014). Importancia de los microARN en el diagnóstico y desarrollo de enfermedades. *Revista Médica del Instituto Mexicano del Seguro Social*, *52*(3), 302–307.

Riolo, G., Cantara, S., Marzocchi, C., & Ricci, C. (2020). MiRNA targets: From prediction tools to experimental validation. *Methods and Protocols*, *4*(1), 1. https://doi.org/10.3390/mps4010001

Sánchez, P., & Valencia Orozco, S. (2020). *Análisis crítico de las técnicas de inteligencia artificial más utilizadas en la predicción del cáncer de mama a partir de mamografías*.

Sarkar, J. P., Saha, I., Sarkar, A., & Maulik, U. (2021). Machine learning integrated ensemble of feature selection methods followed by survival analysis for predicting breast cancer subtype specific miRNA biomarkers. *Computers in Biology and Medicine*, *131*(104244), 104244. https://doi.org/10.1016/j.compbiomed.2021.104244

Seyednasrollah, F., Laiho, A., & Elo, L. L. (2015). Comparison of software packages for detecting differential expression in RNA-seq studies. *Briefings in Bioinformatics*, *16*(1), 59–70. https://doi.org/10.1093/bib/bbt086

Sherafatian, M. (2018). Tree-based machine learning algorithms identified minimal set of miRNA biomarkers for breast cancer diagnosis and molecular subtyping. *Gene*, *677*, 111–118. https://doi.org/10.1016/j.gene.2018.07.057

Ta, T., Nguyen, Q., Chu, H., Long, V. (2019). RAS/RAF mutations and their associations with epigenetic alterations for distinct pathways in Vietnamese colorectal cancer. *Pathol Res Pract*.

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., Manoff, M., & Frame, M. (2011). Data sharing by scientists: practices and

perceptions. *PloS One*, *6*(6), e21101. https://doi.org/10.1371/journal.pone.0021101

Thomson, D. W., Bracken, C. P., & Goodall, G. J. (2011). Experimental strategies for microRNA target identification. *Nucleic Acids Research*, *39*(16), 6845–6853. https://doi.org/10.1093/nar/gkr330

The Wellcome Trust (2003) Sharing data from large-scale biological research projects: A system of tripartite responsibility. Report of a meeting organized by the Wellcome Trust and held on 14–15 January 2003 at Fort Lauderdale, USA.
Available: http://www.genome.gov/pages/research/wellcomereport0303.pdf. Accessed 3 May 2012.

The White House Office of the Press Secretary (2000) Joint statement by President Clinton and Prime Minister Tony Blair of the U.K. Available: http://clinton4.nara.gov/WH/EOP/OSTP/html/00314.html. Accessed 3 May 2012.

Tüfekci, K. U., Oner, M. G., Meuwissen, R. L. J., & Genç, S. (2014). The role of microRNAs in human diseases. *Methods in Molecular Biology (Clifton, N.J.)*, *1107*, 33–50. https://doi.org/10.1007/978-1-62703-748-8_3

Vlachos, I. S., Paraskevopoulou, M. D., Karagkouni, D., Georgakilas, G., Vergoulis, T., Kanellos, I., Anastasopoulos, I.-L., Maniou, S., Karathanou, K., Kalfakakou, D., Fevgas, A., Dalamagas, T., & Hatzigeorgiou, A. G. (2015). DIANA-TarBase v7.0: indexing more than half a million experimentally supported miRNA:mRNA interactions. *Nucleic Acids Research*, *43*(Database issue), D153-9. https://doi.org/10.1093/nar/gku1215

Walport, M., & Brest, P. (2011). Sharing research data to improve public health. *Lancet*, *377*(9765), 537–539. https://doi.org/10.1016/s0140-6736(10)62234-9

Warburg, O. (1956). Origin of cancer cells. *Oncologia (Basel)*, *9*(2), 75–83.

Wightman, B., Ha, I., & Ruvkun, G. (1993). Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C.

elegans. *Cell*, *75*(5), 855–862. https://doi.org/10.1016/0092-8674(93)90530-4

Yin, W., Wang, J., Jiang, L., & James Kang, Y. (2021). Cancer and stem cells. *Experimental Biology and Medicine (Maywood, N.J.)*, *246*(16), 1791–1801. https://doi.org/10.1177/15353702211005390

Yue, D., Liu, H., & Huang, Y. (2009). Survey of computational algorithms for MicroRNA target prediction. *Current Genomics*, *10*(7), 478–492. https://doi.org/10.2174/138920209789208219

Yu, Z., Wang, Z., Yu, X., & Zhang, Z. (2020). RNA-seq-based breast cancer subtypes classification using machine learning approaches. *Computational Intelligence and Neuroscience*, *2020*, 4737969. https://doi.org/10.1155/2020/4737969

Zhi, H., Huang, H., Wu, C., Jung, M. (2011). Omics based molecular target and biomarker identification. *Meth Mol Biol,* 719, 547-571.