

Detección temprana de la enfermedad causada por el virus de Hepatitis C usando redes neuronales artificiales

C. Castro^{1*}, B. Benítez-Mata², E. Vargas¹, and D.-L. Flores¹

¹ Universidad Autónoma de Baja California, Baja California, México.

² University of California Irvine, California, USA.

*carlos.castro98@uabc.edu.mx

Resumen

El virus de la Hepatitis C (HCV) representa el 20% de los casos de hepatitis aguda, enfermedad de difícil diagnóstico que afecta a cerca de 170 millones de personas en el mundo. Debido a que solo se detecta en la etapa aguda, es habitual que el tratamiento de la enfermedad sea complicado por lo que es necesario desarrollar herramientas que permitan realizar una detección temprana de la enfermedad. Previamente se ha demostrado la efectividad de las redes neuronales artificiales (RNA) en aplicaciones de predicción y clasificación con biomarcadores. En este trabajo se desarrolló una herramienta basada en RNA que puede hacer la clasificación entre pacientes sanos y con Hepatitis C, por medio de atributos biométricos del paciente. Generando un modelo con un área bajo la curva de 97.22%. Esta herramienta es clave para desarrollar tecnologías que ayuden a los médicos a llevar a cabo diagnósticos tempranos y precisos.

Palabras clave: Detección temprana, Hepatitis C, Redes Neuronales Artificiales

1 Introducción

El virus de la Hepatitis C (HCV) representa aproximadamente un 20% de los casos de hepatitis aguda, enfermedad que daña la función del hígado, afectando a cerca de 170 millones de personas en el mundo según datos de la Organización Mundial de la Salud [1].

El diagnóstico de esta infección permite identificar y tratar adecuadamente a las personas portadoras del virus, evitando así la progresión y esparcimiento de la enfermedad.

Una característica de esta infección es la falta de síntomas que se presentan en una persona portadora, los cuáles aparecen cuando la infección se encuentra en un estadio agudo, por lo tanto, la mayoría de los portadores son asintomáticos y por ende, es importante desarrollar herramientas de diagnóstico que puedan dar una aproximación de la presencia del virus o un diagnóstico de la enfermedad en revisiones médicas rutinarias [2, 3].

El método estándar de diagnóstico es la detección de anticuerpos anti-HCV, y confirmación por presencia de ácidos nucleicos virales por prueba de la reacción en cadena de polimerasa (PCR), la cual amplifica el material genético presente en una muestra hasta tener límites detectables [4, 5], cabe mencionar que este proceso puede tardar desde un par de días hasta una semana para obtener resultados finales. Entre algunas de las variantes de estos métodos de diagnóstico se encuentran la detección de ARN viral por medio de RT-PCR, inmunoensayos enzimáticos, ensayos de *immunoblot* recombinantes, amplificación mediada por transcripción y lo que se conoce como prueba rápida o *point of care test* [2].

La recomendación del Centro de Control de Enfermedades de los Estados Unidos de América (CDC por sus siglas en inglés) es llevar a cabo un cribado constante en personas que se encuentren en riesgo de contagio o que presenten suficientes factores genéticos y demográficos para ser considerados grupos de riesgo [4].

El uso de técnicas computacionales sobre información contenida en un expediente médico electrónico (EME) tiene la capacidad de mejorar el cribado de pacientes [6, 7, 8, 9, 10, 11, 12]. En este caso, el EME contiene información básica de cada paciente que ayuda al algoritmo a determinar si el paciente presenta un riesgo de portar HCV, alertando al personal médico para una pronta actuación en el tratamiento de la enfermedad.

Por lo anterior, tener a disponibilidad herramientas eficaces y confiables que ayuden a realizar esta tarea son prioridad.

Previos grupos han desarrollado algoritmos de *Machine Learning* (ML) e inteligencia artificial (IA), árboles de decisión, máquinas de soporte vectorial, redes neuronales tipo perceptrón multicapa, clasificador bayesiano, entre otros; para analizar conjuntos de datos relacionados al virus de Hepatitis tipo C y B, dichos conjuntos de datos contienen etiquetas binarias (presencia o no de cirrosis) o multiclases (estadios de enfermedad hepática o genotipo del virus) [13, 14, 15, 16, 17].

En los estudios mencionados, cada uno de los conjuntos de datos contiene distintas cantidades de atributos usados para entrenar al algoritmo, sin embargo al tratarse de una patología hepática, comparten por lo menos cinco atributos; edad, género, aspartato amino-transferasa, alanina amino-transferasa y hemoglobina. Además, en algunos casos el conjunto de datos contiene información de secuenciación genética o factores demográficos.

Entre las herramientas con mayor potencial para llevar a cabo las tareas mencionadas anteriormente se encuentran las redes neuronales artificiales (RNA), las cuales tienen la habilidad de realizar procesos de clasificación y predicción, entre otras tareas, que pueden resultar beneficiosas al ser aplicadas en el sector salud de forma eficiente y eficaz [18, 19, 20, 21, 22, 23].

Una RNA tiene como propósito relacionar un dato proporcionado a su entrada con una salida previamente etiquetada correctamente, esta relación es usada para entrenar al algoritmo para cuando un dato nuevo de entrada con etiqueta desconocida sea clasificado a algunas de las etiquetas de salida previamente usadas en el entrenamiento [24].

Una RNA tiene la habilidad de descubrir patrones y clasificar de manera eficiente y eficaz datos que pueden auxiliar a la toma de decisiones por parte del personal médico, ahorrando recursos al personal de salud y ganando tiempo en caso de diagnosticar alguna enfermedad de manera temprana. En este trabajo se realizó un clasificador para la detección temprana de hepatitis C, haciendo uso de redes neuronales artificiales.

2 Metodología

2.1 Conjunto de datos

El conjunto de datos, obtenido del *UCI Machine Learning Repository* [25], disponible a partir del 10 de junio del año 2020, consiste en un total de 582 registros, de los cuales 526 corresponden a pacientes sanos (muestras control) y 56 a pacientes con Hepatitis C, de edades entre 23 y 77 años (con una media de 50, desviación estándar de 10.04). El diagnóstico fue obtenido mediante pruebas serológicas e histopatológicas para Hepatitis C de acuerdo a los autores del conjunto de datos [8].

El conjunto de datos, consiste de 12 atributos: edad, sexo, albúmina (ALB), fosfatasa alcalina (ALP), alanina aminotransferasa (ALT), aspartato aminotransferasa (AST), bilirrubina (BIL), colina esterasa (CHE), colesterol (CHOL), creatinina (CREA), glutamil transferasa (GGT), y proteína total (PROT).

2.2 Selección de atributos

Se realizó un análisis de importancia relativa para reducir el número de atributos a procesar y por lo tanto reducir el tiempo de entrenamiento de las RNA. Esto se hizo con la librería *Scikit-Learn* versión 0.23.2 en Python 3.7. Se utilizó la técnica de bosques aleatorios (*Random Forest*) los cuales usan árboles de decisión con 250 estimadores en submuestras del conjunto de datos para calcular la exactitud de

cada uno y generar una lista de la importancia de cada atributo en el conjunto de datos, este proceso se hizo usando el 75 % de los datos para entrenamiento y el 25 % para prueba.

2.3 Redes neuronales

Para entrenar las RNA y hacer el análisis con hiperparámetros se hizo uso de la librería *Keras* y *Scikit-Learn*. Se crearon RNA de dos capas ocultas, de forma 5- n_1 - n_2 -1, donde n_1 representa el número de neuronas en la primera capa oculta y n_2 representa el número de neuronas en la segunda capa oculta, n_1 y n_2 son dos de los hiperparámetros a evaluar junto con el optimizador del modelo. Para hacer el análisis con hiperparámetros se varió n_1 y n_2 de 10 a 100 con incrementos de 5, y el optimizador entre *Adam* y *Adadelta*. Dando como resultando un total de 234 RNA.

2.4 Métricas de evaluación

Para evaluar las RNA se usó la matriz de confusión. La matriz de confusión esta compuesta por muestras positivas predichas correctamente (TP), muestras negativas predichas correctamente (TN), muestras positivas predichas incorrectamente (FN) y muestras negativas predichas incorrectamente (FP). Además de la matriz de confusión, existen las siguientes métricas para determinar la calidad de un clasificador: exactitud (1), precisión (2), sensibilidad (3), especificidad (4) y *F1-score* (5) [26].

$$Exactitud = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

$$Precisión = \frac{TP}{TP + FP} \quad (2)$$

$$Sensibilidad = \frac{TP}{TP + FN} \quad (3)$$

$$Especificidad = \frac{TN}{FP + TN} \quad (4)$$

$$F1-score = \frac{2 * Precisión * Recall}{Precisión + Recall} \quad (5)$$

Además se utilizó el área bajo la curva característica operativa del receptor (ROC-AUC) que es una medida de diagnóstico para medir el poder de discriminación de un clasificador. Un valor máximo de 1.0, significa que todas las instancias se clasificaron correctamente. La gráfica ROC-AUC contiene una línea recta que representa un valor de 0.5, que representa un 50 % de capacidad de clasificación. Un valor de ROC-AUC mayor de 0.97 tiene un alto valor clínico [27].

3 Resultados y Discusión

En la Tabla 1 se muestran todos los atributos con sus puntajes de importancia obtenidos con los bosques aleatorios descritos en la Sección 2.2. Los cinco atributos con más peso para predecir correctamente resultaron ser: AST, CHE, ALP, BIL, y ALT. De esta forma el conjunto de datos quedó con dimensiones 582x5, lo cual es una ventaja para reducir de manera importante el tiempo de procesamiento y memoria computacional al momento de entrenar todas las RNA con los hiperparámetros.

Al realizar el análisis de hiperparámetros con *GridSearchCV* de la librería *Scikit-Learn*, con una validación cruzada de cinco vías, se encontró que la mejor configuración (con un valor F1 de 0.900015) para este conjunto de datos fue 5-10-30-1 con el optimizador *Adadelta*. Estos resultados son prometedores debido a que un menor número de neuronas en las capas ocultas se traduce a un menor requerimiento de poder computacional.

Tabla 1: Ranking de importancia de los atributos.

Atributo	Puntaje de atributo
AST	0.2262
CHE	0.1123
ALP	0.1092
BIL	0.1004
ALT	0.0982
GGT	0.0959
CREA	0.0642
ALB	0.0491
CHOL	0.0464
Age	0.0456
Prot	0.0370
Sex	0.0152

Después se entrenó una red con esa configuración usando el 75 % de datos para entrenamiento y 25 % para prueba. Se analizaron las predicciones obtenidas con el conjunto de prueba así como con el conjunto total de datos. En la Tabla 2 se muestran los valores obtenidos de las métricas para cada conjunto analizado. Los resultados sobre el conjunto de pruebas permiten saber que no hubo sobre-entrenamiento de la RNA. Por otra parte, los resultados sobre el conjunto total indican cómo se comportará el modelo en general cuando se utilice para generar una predicción con datos de nuevos pacientes.

Tabla 2: Resultados de las métricas de la red 5-10-30-1 con el conjunto de prueba y completo.

Conjunto	AUC	<i>F1-score</i>	Exactitud	Precisión	Sensibilidad	Especificidad
Prueba	0.961538	0.960000	0.993150	1.000000	0.923076	1.000000
Total	0.972263	0.963636	0.993127	0.981481	0.964285	0.971481

De la misma forma se obtuvo la matriz de confusión para cada caso. En la Fig. 1, se observan las predicciones de los casos negativos y positivos para el conjunto de prueba durante el entrenamiento del nuevo modelo (izquierda) y el conjunto total para probar el desempeño en la aplicación del modelo sobre todos los datos disponibles (derecha). En ambos casos el número de casos predichos incorrectamente es bajo, con 1 en el conjunto de prueba y 4 en el total de los datos. Esto es importante porque la red es capaz de discriminar entre los pacientes enfermos y los que no, situación que puede llegar a ser complicada cuando los casos negativos superan a los positivos o viceversa, como es el caso de este conjunto de datos. A pesar de ello, se obtuvieron resultados satisfactorios dando cabida a la aplicación de este tipo de modelos en la ayuda del diagnóstico médico. De la misma manera, en la Fig. 2 se muestra la curva ROC-AUC para cada uno de los conjuntos de prueba y total para la configuración 5-10-30-1.

Al hacer una comparación entre el presente trabajo y el estudio original del cual se obtuvo el conjunto de datos [8], se observa lo siguiente: la idealización del modelo en el trabajo realizado por [8] pretende que el algoritmo de *Machine Learning* emule la toma de decisiones que el personal médico realiza a la hora de diagnosticar a un paciente con síntomas de la enfermedad, por lo que un algoritmo de tipo árboles de decisión se propone como el más óptimo. Esto no implica que dicho algoritmo tendrá el mejor rendimiento de entre otras opciones de ML.

Las redes neuronales artificiales en este trabajo, con exactitud del 99.31 % superan por mucho a los algoritmos de árboles de decisión *rpart* y *ctree*, con exactitud 75.3 % y 72.6 % respectivamente, reportados como óptimos en [8]. A la fecha y al conocimiento de los autores, no se han encontrado otros

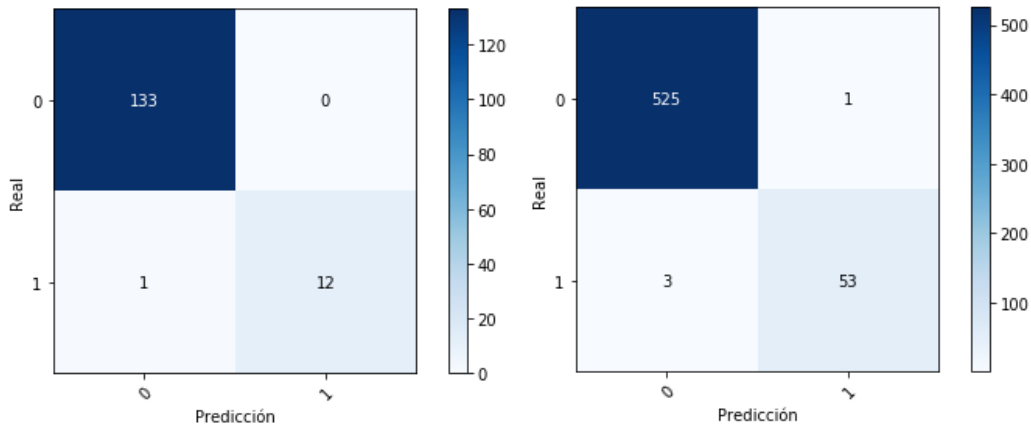


Fig. 1. Matriz de confusión resultante para el conjunto de prueba durante el entrenamiento de la red (izquierda) y el conjunto de datos completo después del entrenamiento (derecha)

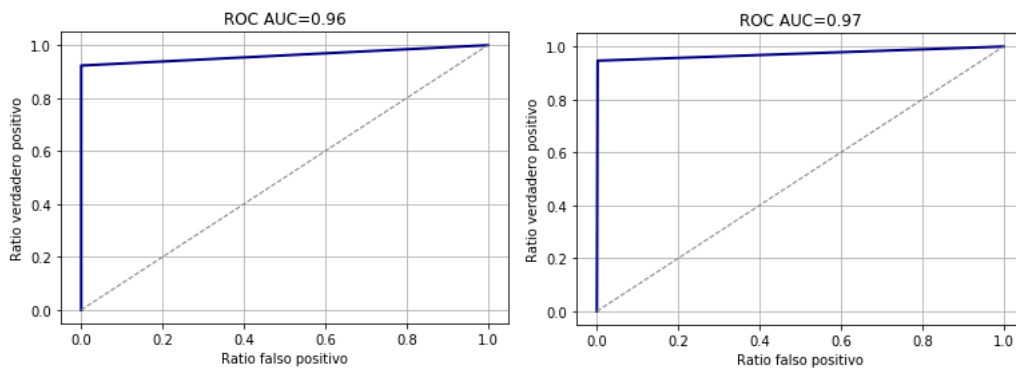


Fig. 2. Curva ROC-AUC resultante para el conjunto de prueba durante el entrenamiento de la red (izquierda) y para el conjunto de datos completo después del entrenamiento (derecha)

trabajos haciendo uso del mismo conjunto de datos.

4 Conclusiones

El mejor modelo tuvo un desempeño de acuerdo a su área bajo la curva de 97.22%, además de una exactitud de 99.31%; mejorando los resultados previamente reportados en el mismo conjunto de datos [8] y demostrando la capacidad de realizar una predicción adecuada de la enfermedad de Hepatitis C con el uso de cinco biomarcadores presentes en el paciente. Esto resulta particularmente importante debido a la dificultad que existe de detectar esta enfermedad en una etapa temprana, en general, los pacientes no presentan síntomas a menos que la enfermedad haya progresado considerablemente, aumentando el riesgo de contagio y por lo tanto, dificultando el tratamiento, prevención y control de la enfermedad [3].

Es importante recalcar que el modelo mostró resultados prometedores para determinar los pacientes sanos tanto como los que presentan la enfermedad. Aún cuando la cantidad de muestras de pacientes sanos son mayores que las de pacientes enfermos, el modelo fue capaz de diferenciarlos, sugiriendo que

los atributos utilizados son suficientes para realizar dicha discriminación.

Estos resultados demuestran que las redes neuronales artificiales son herramientas adecuadas para generar predicciones de diagnóstico médico que sirvan de apoyo al personal de salud [24, 28], y por lo tanto es importante seguir evaluando estas técnicas en otros objetos de estudio con el fin de perfeccionar su funcionamiento y su posible aplicación en el sector médico.

Al mismo tiempo, es un reto implementar estas herramientas en los ambientes hospitalarios y de ahí la importancia de generar datos de pacientes locales que permitan desarrollar dichas herramientas. Por lo pronto, acudir a un laboratorio clínico donde se realizan estos tipos de análisis a los pacientes y donde dichos parámetros están disponibles para ser analizados por este tipo de algoritmos es una motivación a corto plazo.

Declaración de conflictos de interés

Los autores declaran no tener algún conflicto de interés.

Referencias

- [1] O. M. de la Salud, “Hepatitis c,” 2020.
- [2] H. C. Li and S. Y. Lo, “Hepatitis C virus: Virology, diagnosis and treatment,” *World J. Hepatol.*, vol. 7, no. 10, pp. 1377–1389, 2015.
- [3] M. L. Shiffman, “Universal screening for chronic hepatitis C virus,” *Liver Int.*, vol. 36, pp. 62–66, 2016.
- [4] S. Schillie, C. Wester, M. Osborne, L. Wesolowski, and A. B. Ryerson, “CDC recommendations for hepatitis C screening among adults—United States, 2020,” *MMWR Recommendations and Reports*, vol. 69, no. 2, pp. 1–17, 2020.
- [5] E. Association, “EASL Recommendations on Treatment of Hepatitis C 2016,” *J. Hepatol.*, vol. 66, no. 1, pp. 153–194, 2017.
- [6] V. Shahnazarian, E. Karu, and P. Mehta, “Hepatitis C: improving the quality of screening in a community hospital by implementing an electronic medical record intervention,” *BMJ Qual. Improv. Rep.*, vol. 4, no. 1, p. u208549.w3409, 2015.
- [7] M. A. Konerman, M. Thomson, K. Gray, M. Moore, H. Choxi, E. Seif, and A. S. Lok, “Impact of an electronic health record alert in primary care on increasing hepatitis c screening and curative treatment for baby boomers,” *Hepatology*, vol. 66, no. 6, pp. 1805–1813, 2017.
- [8] G. Hoffmann, A. Bietenbeck, R. Lichtinghagen, and F. Klawonn, “Using machine learning techniques to generate laboratory diagnostic pathways—a case study,” *J. Lab. Precis. Med.*, vol. 3, pp. 58–58, 2018.
- [9] L. Vaughan, “Biomarkers in acute medicine Key points,” *Medicine*, vol. 45, no. 3, pp. 150–156, 2016.
- [10] M. E. Vandenberghe, M. L. Scott, P. W. Scorer, M. Söderberg, D. Balcerzak, and C. Barker, “Relevance of deep learning to facilitate the diagnosis of HER2 status in breast cancer,” *Sci. Rep.*, vol. 7, pp. 1–11, 2017.
- [11] M. Sherafatian, “Tree-based machine learning algorithms identified minimal set of miRNA biomarkers for breast cancer diagnosis and molecular subtyping,” *Gene*, vol. 677, pp. 111–118, 2018.
- [12] S. Y. Loke and A. S. G. Lee, “The future of blood-based biomarkers for the early detection of breast cancer,” *Eur. J. Cancer*, vol. 92, pp. 54–68, 2018.
- [13] S. C. R. Nandipati, C. Xinying, and K. K. Wah, “Hepatitis C Virus (HCV) Prediction by Machine Learning Techniques,” *Applications of Modelling and Simulation*, vol. 4, pp. 89–100, 2020.

- [14] R. Wei, J. Wang, X. Wang, G. Xie, Y. Wang, H. Zhang, C. Y. Peng, C. Rajani, S. Kwee, P. Liu, and W. Jia, "Clinical prediction of HBV and HCV related hepatic fibrosis using machine learning," *EBioMedicine*, vol. 35, pp. 124–132, 2018.
- [15] A. J. Mueller-Breckenridge, F. Garcia-Alcalde, S. Wildum, S. L. Smits, R. A. de Man, M. J. van Campenhout, W. P. Brouwer, J. Niu, J. A. Young, I. Najera, L. Zhu, D. Wu, T. Racek, G. B. Hundie, Y. Lin, C. A. Boucher, D. van de Vijver, and B. L. Haagmans, "Machine-learning based patient classification using Hepatitis B virus full-length genome quasispecies from Asian and European cohorts," *Sci. Rep.*, vol. 9, no. 1, pp. 1–12, 2019.
- [16] A. A. A. Radwan and H. Mamdouh, "An Analysis of Hepatitis C Virus Prediction Using Different Data Mining Techniques," *International Journal of Computer Science Engineering and Information Technology Research (IJCEITR)*, vol. 3, no. 4, pp. 209–220, 2013.
- [17] S. M. Abd El-Salam, M. M. Ezz, S. Hashem, W. Elakel, R. Salama, H. ElMakhzangy, and M. ElHefnawi, "Performance of machine learning approaches on prediction of esophageal varices for Egyptian chronic hepatitis C patients," *Inform. Med. Unlocked*, vol. 17, p. 100267, 2019.
- [18] H. Wang, B. Zheng, S. W. Yoon, and H. S. Ko, "A support vector machine-based ensemble algorithm for breast cancer diagnosis," *Eur. J. Oper. Res.*, vol. 267, no. 2, pp. 687–699, 2018.
- [19] L. Tapak, N. Shirmohammadi-Khorram, P. Amini, B. Alafchi, O. Hamidi, and J. Poorolajal, "Prediction of survival and metastasis in breast cancer patients using machine learning classifiers," *Clin. Epidemiol. Glob. Health.*, vol. 7, pp. 1–7, 2018.
- [20] D. M. Camacho, K. M. Collins, R. K. Powers, J. C. Costello, and J. J. Collins, "Next-Generation Machine Learning for Biological Networks," *Cell*, vol. 173, no. 7, pp. 1581–1592, 2018.
- [21] C. Cao, F. Liu, H. Tan, D. Song, W. Shu, W. Li, Y. Zhou, X. Bo, and Z. Xie, "Deep Learning and Its Applications in Biomedicine," *GENOM. PROTEOM. BIOINF.*, vol. 16, no. 1, pp. 17–32, 2018.
- [22] D. L. Flores, C. Gómez, D. Cervantes, A. Abaroa, C. Castro, and R. A. Castañeda-Martínez, "Predicting the physiological response of Tivela stultorum hearts with digoxin from cardiac parameters using artificial neural networks," *BioSystems*, vol. 151, pp. 1–7, 2017.
- [23] B. Ehteshami Bejnordi, M. Mullooly, R. M. Pfeiffer, S. Fan, P. M. Vacek, D. L. Weaver, S. Hershorn, L. A. Brinton, B. van Ginneken, N. Karssemeijer, A. H. Beck, G. L. Gierach, J. A. van der Laak, and M. E. Sherman, "Using deep convolutional neural networks to identify and classify tumor-associated stroma in diagnostic breast biopsies," *Mod. Pathol.*, vol. 31, no. 10, pp. 1502–1512, 2018.
- [24] E. M. El Houby, "A survey on applying machine learning techniques for management of diseases," *J. Appl. Biomed.*, vol. 16, no. 3, pp. 165–174, 2018.
- [25] D. Dua and C. Graff, "UCI machine learning repository," 2017.
- [26] B. Boukenze, H. Mousannif, and A. Haqiq, "Performance of data mining techniques to predict in healthcare case study : Chronic kidney failure disease," *Int. J. Database Manag. Syst.*, vol. 8, pp. 1–9, 2016.
- [27] J. Fan, S. Upadhye, and A. Worster, "Understanding receiver operating characteristic (ROC) curves," *Can J Emerg Med*, vol. 8, no. 1, pp. 19–20, 2006.
- [28] C. Argyropoulou, A. Karioti, and H. Skaltsa, "Disease prediction with different types of neural network classifier," *Chem. Biodivers.*, vol. 8, no. 10, pp. 1880–1890, 2011.