Tiempos Determinantes Para la Clasificación de Personas con Depresión Utilizando Algoritmos Genéticos.

Carlos H. Espino-Salinas[1], Pablo C. Rodríguez-Aguayo, Carlos E. Galván-Tejada[1*]

[1] Unidad Académica de Ingeniería Eléctrica, Universidad Autónoma de Zacatecas, Jardín Juarez 147, Centro, Zacatecas, Zac. México; dantecore2@gmail.com, is.pab.ces@gmail.com, ericgalvan@uaz.edu.mx, gatejo@uaz.edu.mx, hamurabigr@uaz.edu.mx, jose.celaya@uaz.edu.mx

Introducción: La depresión afecta a más de 300 millones de personas en el mundo de acuerdo con la Organización Mundial de la Salud (OMS) y su prevalencia oscila entre 3% y 21%. En la mayoría de los países se presenta dos veces más en las mujeres que en los hombres. La depresión se manifiesta como un conjunto de síntomas psicológicos y físicos. El predominio de sentimientos de tristeza, apatía, desesperanza e irritabilidad, así como el deterioro en el aspecto personal, el llanto y el insomnio, son marcadores de la presencia de la enfermedad.[1.] La depresión también puede presentarse junto con otras enfermedades emocionales y/o físicas, por ejemplo, con ansiedad, abuso en el consumo de alcohol y otras sustancias, enfermedades sistemáticas.[2] De acuerdo a la duración de la presencia de los síntomas que van desde transitorios, hasta persistentes (meses o años), se puede clasificar a la depresión como leve, moderada o severa. Dependiendo del grado de depresión, ésta puede interferir con las actividades laborales, escolares, familiares y sociales, teniendo una prevalencia en personas entre los 15 a 45 años.[3.]**Antecedentes:** Cada año se suicidan cerca de 800 000 personas, y el suicidio es la segunda causa de muerte entre las personas de 15 a 29 años (OMS)**.Objetivo:** Determinar los tiempos clave para poder clasificar a personas sanas y personas con depresión utilizando algoritmos genéticos y técnicas de aprendizaje automatico.**Materiales y Métodos:** Utilizando los datos obtenidos del depresjon dataset que es un conjunto de datos que se recopilo originalmente para el estudio de la actividad motriz en la esquizofrenia y la depresión mayor, de este se extraen los tiempos más relevantes o claves usando algoritmos genéticos para después aplicar técnicas de aprendizaje automático como: redes neuronales y regresión logística para clasificar a pacientes entre sanos y con depresión. **Resultados: conclusiones:** Los algoritmos genéticos nos proporcionó una selección de características útiles dentro del conjunto de datos para generar una clasificación satisfactoria con un porcentaje de precisión de **0.747** y un área bajo la curva de **0.751** para el caso de las redes neuronales y en la regresión logística una precisión de **0.756** con área bajo la curva de **0.758**. Esto indica que es posible crear un sistema capaz de clasificar a personas con depresión debido a la posibilidad de aumentar el índice de confiabilidad dadas estas métricas de validación.

Palabras Clave: Aprendizaje automático, algoritmos genéticos, depresión, actividad motriz, clasificación

1. Introducción

La depresión es un trastorno mental frecuente, que se caracteriza por la presencia de tristeza, perdida de interés o placer, sentimientos de culpa o falta de autoestima, trastornos del sueño o del apetito, sensación de cansancio y falta de concentración.

La depresión puede llegar a hacerse crónica i recurrente y dificultar sensiblemente de desempeño en el trabajo o la escuela y la capacidad para afrontar la vida diaria. En su forma más grave, puede conducir al suicidio. Si es leve, se puede tratar sin necesidad de medicamentos, pero cuando tiene carácter moderado o grave se puede necesitar medicamentos y psicoterapia profesional.

La depresión es un trastorno que se puede diagnosticar de forma fiable y que puede ser tratado por no especialistas en el ámbito de la atención primaria (**OMS**).

La depresión está asociada con los ritmos biológicos interrumpidos causados por perturbaciones ambientales como el cambio estacional a la luz del día, la alteración de los ritmos sociales debido, por ejemplo, a los turnos de trabajo o los viajes largos; además de estar relacionado con estilos de vida asociados con ritmos diurnos inconscientes con el ciclo natural de la luz del día. La aparición de síntomas depresivos se relaciona además con problemas de salud física, efectos secundarios médicos, eventos de la vida y factores sociales, además del abuso de alcohol y sustancias, y dichos factores también podrían causar síntomas de depresión en todos los humanos. La prevalencia global de depresión en la vida es aproximadamente del 15%, pero la incidencia de episodios con un nivel de gravedad que no cumple con los requisitos para un diagnostico depresivo es mucho más frecuente [1].

Desafortunadamente, el método de evaluación psiquiátrica actual requiere un gran esfuerzo por parte de los médicos para reunir información completa sobre los pacientes. Además, depende en gran medida del esfuerzo de los pacientes para cooperar en la comunicación de sus síntomas y problemas. Por lo tanto, se necesita un mecanismo de detección objetivo basado en señales, biológicas y de comportamiento para mejorar el método de diagnóstico actual [2].

En los últimos años, la investigación sobre el diagnóstico de enfermedades mentales basada en el habla ha ganado popularidad. En general, la metodología implica la recopilación de datos, el procesamiento previo de datos, la extracción de características y la clasificación de esas características [3]. Las lecturas actigraficas de la actividad motriz se considera también un método objetivo para el estudio de esquizofrenia y depresión mayor [1].

Los algoritmos genéticos (AG) son técnicas de búsqueda aleatoria y optimización guiadas por los principios de evolución y genética natural, que tienen una gran cantidad de paralelismo implícito. Los AG realizan búsquedas en ambientes complejos, grandes y multimodales, y brindan soluciones casi óptimas para el objetivo o la función de un problema de optimización. En los AG los parámetros del

espació de búsqueda se codifican en forma de cadenas (llamadas cromosomas). A la colección de tales cadenas se llama población. Inicialmente, se crea una población aleatoria, que representa diferentes puntos en el espacio de búsqueda. Una función de objetivo y de aptitud se asocia con cada cadena que representa el grado de bondad de la cadena. Según el principio de supervivencia del más apto, se seleccionan algunas de las cadenas y a cada una se le asigna una cantidad de copias que van al grupo de apareamiento. En estas cadenas se aplican operadores inspirados biológicamente como el cruce y la mutación para producir una nueva generación. El proceso de selección, cruce y mutación continua durante un número fijo de generaciones o hasta que se cumpla una condición de terminación [4]. Los AG tienen aplicaciones varios campos como el procesamiento de imágenes, redes neuronales, aprendizaje automático, programación de talleres, etc. [5-6].

Esta investigación tiene como objetivo aplicar algoritmos genéticos capaces de determinar los tiempos clave para clasificar a una persona con o sin depresión ya que el conjunto de datos contiene datos para controles (sanos) y otra para el grupo de condición (con depresión). Para cada paciente se tienen los datos del reloj actigraph que mide los niveles de actividad motriz en intervalos de un minuto [1]. Cada minuto se considera como una variable o población para el proceso de evolución y genética de los AG esto quiere decir que el resultado obtenido serán los minutos claves para poder tener una clasificación de los pacientes para esto será necesario la aplicación de técnicas de aprendizaje automático que permitan validar dichos resultados de la selección de características a partir de los algoritmos genéticos.

2. Materiales y Métodos

La metodología propuesta en este trabajo de investigación consta de 4 etapas como lo muestra la **Figura 1**. Inicialmente se adquieren los datos recabados por el reloj actigraph conocidos como **Depresjon Dataset**, Después se aplica el algoritmo genético (GALGO) para seleccionar las características más relevantes, que en este caso son representadas por los minutos que conforman el conjunto de datos para después aplicar regresión logística y redes neuronales para que finalmente estas técnicas sean validadas por medio de la precisión, curvas ROC y área bajo la curva.

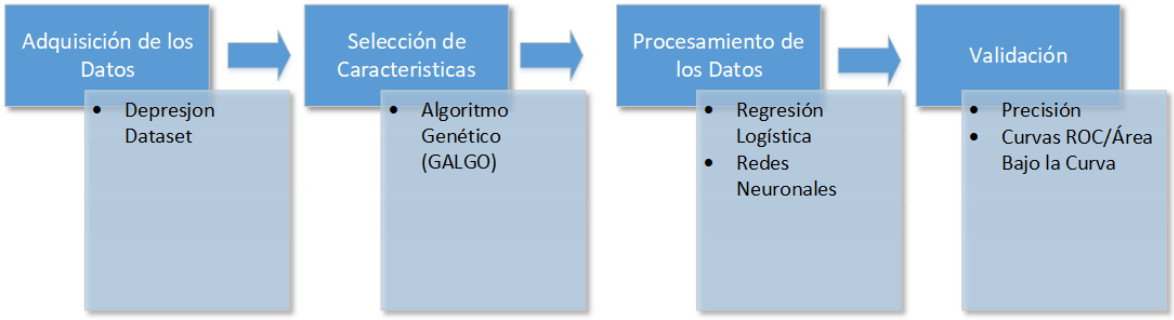

Figura 1. Diagrama de la metodología propuesta.

Todos los procesos y análisis aplicados en esta investigación fueron hechos con el programa con la herramienta RStudio (versión 3.5.3) [7], que es un entorno de software libre para computación estadística y gráficos. Las librerías usadas son las siguientes: GALGO (versión 1.2-01)[8], neuralnet (versión 1.44.2)[9] and Proc (versión 1.11.0)[10].

2.1 Adquisición de los Datos

El conjunto de datos se recopilo originalmente para el estudio de la actividad motriz en la esquizofrenia y la depresión mayor. La activada motriz se controló con un reloj actigraph usado en la muñeca derecha. El reloj actigraph mide los niveles de actividad. La actividad de muestreo es de 32 Hz y se registran movimientos a 0.05 g. Se produce un voltaje correspondiente y se almacena como un recuento de actividad en la unidad de memoria del reloj actigraph. El número de conteos es proporcional a la intensidad del movimiento en intervalos de un minuto.

El dataset contiene registros de actividad motriz de 23 pacientes deprimidos unipolares y bipolares y 32 controles sanos. Para cada paciente se utilizaron datos de un lapso de una semana.

Los sensores portátiles que miden diferentes partes de la actividad de las personas son una tecnología muy común hoy en día. En la investigación, los datos recopilados con estos dispositivos también llaman la atención. Sin embargo, los conjuntos de datos que contienen datos de sensores en el campo de la medicina son raros y a menudo, los datos no son públicos y solo se publican los resultados [11].

2.2 Selección de Características

El objetivo principal de la selección de características o de obtener los minutos mas relevantes para la clasificación de pacientes del lado matemático es evitar el sobreajuste durante el análisis de clasificación, así como mejorar su rendimiento, mediante la reducción de características.

Debido a la gran cantidad de minutos con los que cuenta una semana el procesamiento de estos datos resulta ser muy complejo. En este paso, el algoritmo genético GALGO (GA) realiza una representación esquemática de una selección de variables de dos clases, que implica la combinación de genes o minutos (cromosomas que representan el modelo multivariable) que permite distinguir entre clases en una función de un método de clasificación, donde hay una serie de modelos que realizan este procesamiento varias veces, con distintas combinaciones.

Del lado medico puede permite conocer en que lapsos de tiempo la actividad motriz puede ayudar a entender esta enfermedad y que influencia tienen estos tiempos con los pacientes con depresión.

2.3 Procesamiento de los Datos.

Para el procesamiento de los datos fue necesario dividir los datos en dos conjuntos de datos: entrenamiento y pruebas. El conjunto de entrenamiento consiste en el 70% del total de los datos y el conjunto de pruebas en el 30%. La distribución de los datos es generada de forma aleatoria para después aplicar las técnicas de aprendizaje automático.

Afortunadamente, el campo del aprendizaje automático depende de la generosidad de la naturaleza tanto para la inspiración del mecanismo. Muchos sistemas de aprendizaje automático ahora se basan en gran medida en el pensamiento actual de la ciencia cognitiva otro ejemplo donde se ha aprovechado el ejemplo natural es el caso de los algoritmos genéticos [12].

2.4 Validación

Para finalizar es necesario tener una fuente confiable para determinar si las técnicas de aprendizaje automático aplicadas pueden ser utilizadas para clasificar pacientes y en qué medida o porcentaje son capaces de realizar esta acción de manera satisfactoria, para esto se utilizaron métricas de validación como lo son: la precisión, curvas ROC y área bajo la curva.

El análisis en base a curvas ROC (Receiver Operating Chracteristic Curve) constituyen un método estadístico para determinar la exactitud diagnostica de pruebas que utilizan escalas continuas, siendo utilizadas con tres propósitos específicos: determinar el punto de corte en el que se alcanza la sensibilidad y especificidad más alta, evaluar la capacidad discriminativa de la prueba diagnóstico, es decir, su capacidad de diferenciar sujetos sanos versus enfermos, y comparar la capacidad discriminativa de dos o más pruebas diagnósticos que expresan sus resultados como escalas continuas[13-14].

El parámetro a estimar es el área bajo la curva ROC (AUC, área under the curve), medida única e independiente de la prevalencia de la enfermedad. El AUC refleja que tan bueno es el test para discriminar pacientes con y sin la enfermedad a lo largo de todo el rango de puntos de corte posible [15].

3. *Resultados y Discusión*

El conjunto de datos utilizados para este trabajo está conformado de 1440 variables, después de la selección de características este número es reducido en solo 7 variables clave para la clasificación de pacientes entre ellas se encuentran los siguientes minutos: 15:43, 15:41,12:11, 7:25, 15:40, 7:29, 12:15 como lo muestra la Figura 2. Donde se muestra la frecuencia de aparición de cada tiempo, siendo las características en negro las que tienen el rango más alto.

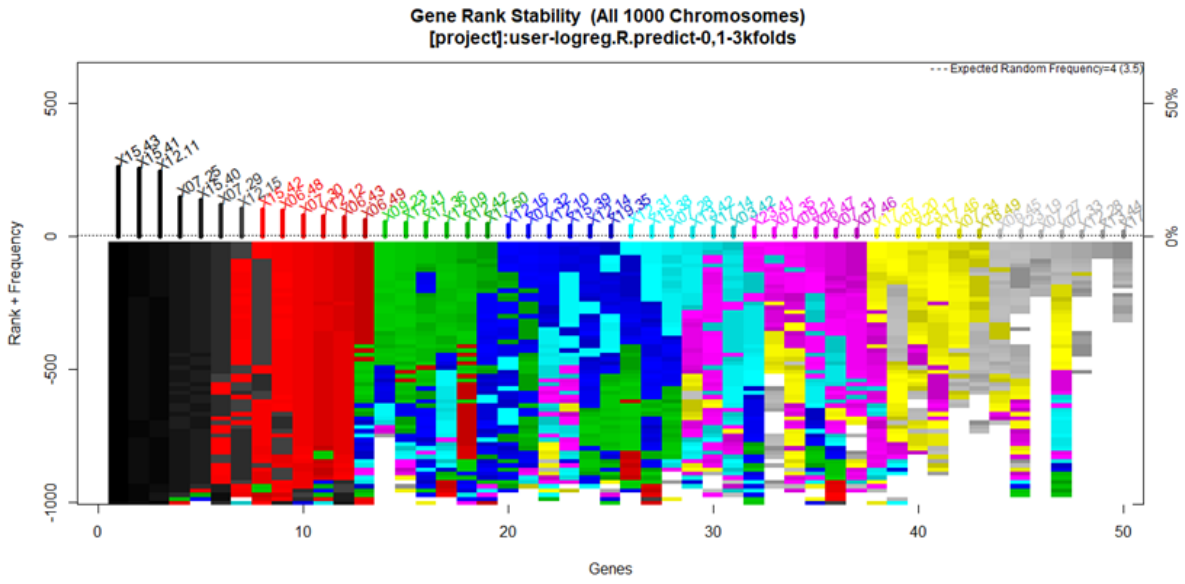

Figura 2. Características más relevantes obtenidas del algoritmo genético.

A partir de estos resultados y en base a estos datos se aplicaron dos técnicas de aprendizaje automático redes neuronales y regresión logística, como se puede observar en la Tabla 1. Muestra la precisión y el área bajo la curva como métricas de validación obtenidas de cada una de estas técnicas.

Métrica	**Regresión Logística**	**Redes Neuronales**
Precisión	0.756	0.747
Área Bajo la Curva	0.758	0.751

Tabla 1. Resultados obtenidos en las métricas de validación por cada de técnica de aprendizaje automático.

Como se puede observar los valores obtenidos son muy similares y su variación es muy pequeña, pero en todo caso la regresión logística tiene mejor desempeño para poder clasificar pacientes.

Otro detalle importante es que los algoritmos genéticos presentan previamente un aproximado de los resultados que se obtendrán en la precisión para clasificar, como lo muestra la Figura 2. Que son los modelos utilizados con selección hacia adelante.

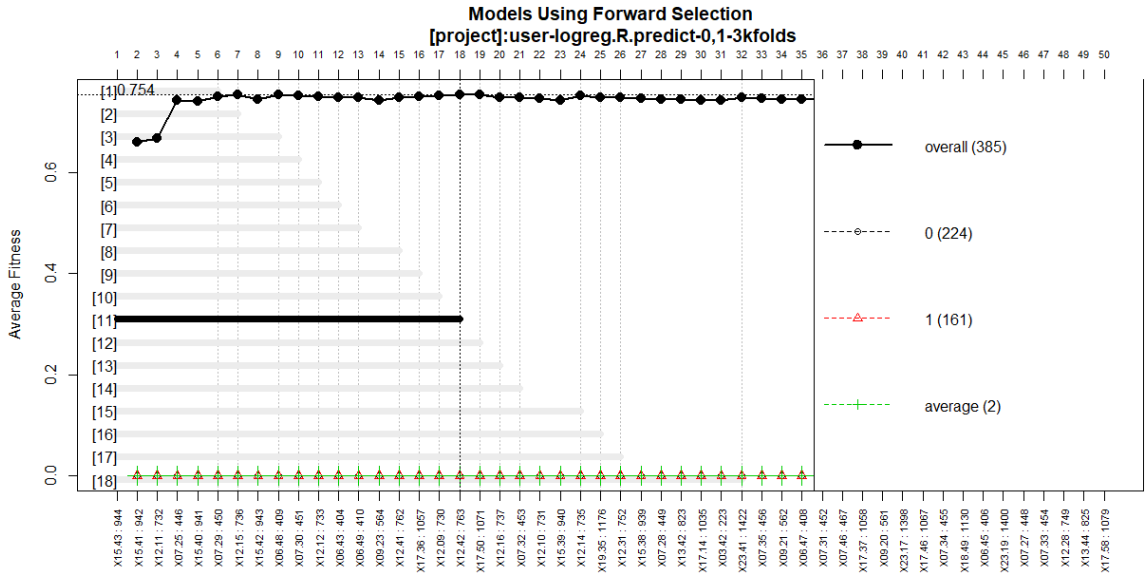

Figura 2. Modelos usando selección hacia adelante para la predicción de pacientes con depresión (1) y sin depresión (0).

Por ultimo tenemos los gráficos obtenidos en base a la especificidad y sensibilidad la primera nos indica la capacidad de nuestro estimador para dar como casos positivos los casos realmente enfermos; proporción de enfermos correctamente identificados y el segundo nos indica la capacidad de nuestro estimador para dar como casos negativos los casos realmente sanos; proporción de sanos correctamente identificados. Teniendo así el área bajo la curva y grafico correspondiente como lo muestra la Figura 3.

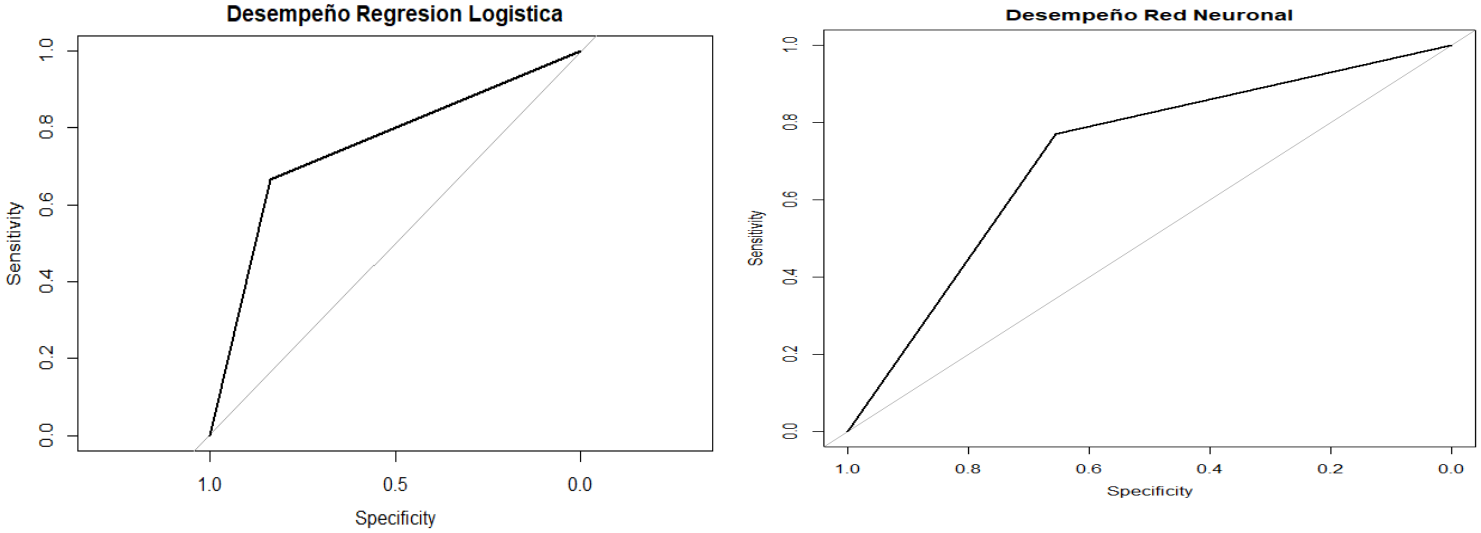

Figura 3. Curvas de regresión logística y redes neuronales utilizadas en esta investigación.

1. *Conclusiones y Trabajos Futuros.*

Como conclusión se puede decir que la selección de características por medio de algoritmos genéticos también puede ser útil como método de reconocimiento de patrones capaces de mejorar la comprensión de enfermedades, como es el caso de la depresión. Una de las desventajas de este tipo de técnicas es el tiempo y costo computacional ya que es muy alto pero que en contra parte facilita el uso del aprendizaje automático reduciendo el número de variables, tiempo y costo computacional del procesamiento de los datos. Otra observación interesante es el hecho de que es posible crear un sistema de clasificación confiable utilizando solo tiempos claves y no todo el conjunto de datos sin un pre-procesamiento.

Como trabajos futuros se propone utilizar redes neuronal profundas que permitan aumentar el nivel de confianza al momento de clasificar pacientes para así poder implementarlas en algún dispositivo inteligente como un Smartphone o Smartwatch y prevenir este padecimiento tan frecuente hoy en día en las personas.

1. Bibliografia

[1] Enrique Garcia-Ceja et al. “Depresjon: A Motor Activity Database of De-pression Episodes in Unipolar and Bipolar Patients”. In:Proceedings ofthe 9th ACM on Multimedia Systems Conference. MMSys’18. Amsterdam,The Netherlands: ACM, 2018.doi:10.1145/3204949.3208125.url:http://doi.acm.org/10.1145/3204949.3208125.

[2] Safiah Khairuddin, Salmiah Ahmad, Abdul Halim Embong, Nik Nur Wahidah Nik Hashim, Tareq M.K. Altamas, Syarifah Nuratikah Syd Badaruddin, Surul Shahbudin Hassan, "Classification of the Correct Quranic Letters Pronunciation of Male and Female Reciters", *IOP Conference Series: Materials Science and Engineering*, vol. 260, pp. 012004, 2017.

[3] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, “A review of depression and suicide risk assessment using speech analysis,” Speech Commun., vol. 71, pp. 10–49, 2015.

[4] J.L.R. Filho, P.C. Treleaven, C. Alippi, Genetic algorithm programming environments, IEEE Comput. 27 (1994) 28-43.

[5] S.K. Pal, D. Bhandari, Selection of optimal set of weights in a layered network using genetic algorithms, Inform. Sci. 80 (1994) 213-234

[6] L.J. Eshelman (Ed.), Proceedings of the Sixth International Conference Genetic Algorithms, Morgan Kaufmann, San Mateo, 1995.

[7] R Core Team.The R Project for Statistical Computing. 2019.url: https://www.r-project.org/.

[8] Victor Trevino and Francesco Falciani.GALGO An R package for Ge-netic Algorithm Searches (Customized for Variable Selection in FunctionalGenomics). Tech. rep. 2005.url: http://www.wag-inc.org.

[9] Stefan Fritsch, Frauke Guenther, and Marvin N. Wright.neuralnet: Train-ing of Neural Networks. R package version 1.44.2. 2019.url: https : / /CRAN.R-project.org/package=neuralnet.

[10] Xavier Robin et al. “pROC: an open-source package for R and S+ toanalyze and compare ROC curves”. In: BMC Bioinformatics12.1 (2011),p. 77.doi: 10.1186/1471-2105-12-77.url: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-12-77.

[11] Garcia-Ceja, E.; Riegler, M.; Jakobsen, P.; Torresen, J.; Nordgreen, T.; Oedegaard, K.J.; Fasmer, O.B.217Motor Activity Based Classification of Depression in Unipolar and Bipolar Patients. Proceedings - IEEE Symposium on Computer-Based Medical Systems. Institute of Electrical and Electronics Engineers Inc.,2192018, Vol. 2018-June, pp. 316–321. doi:10.1109/CBMS.2018.00062.

[12] Goldberg, D. E. (1989). Genetic algorithms in search, optimization, and machine learning. Reading, MA: Addison-Wesley.

[13] Akobeng A. Understanding diagnostic tests 3: receiver characteristic characteristic curves. Acta Paediatr 2007; 96: 644-7.doi: 10.1111/j.1651-2227.2006.00178.x url:https://doi.org/10.1111/j.1651-2227.2006.00178.x

[14] Altman DG, Bland JM. Diagnostic tests 3: receiver operating characteristic plots. BMJ. 1994;309(6948):188. doi:10.1136/bmj.309.6948.188

[15] Cerda, Jaime, & Cifuentes, Lorena. (2012). Using ROC curves in clinical investigation: Theoretical and practical issues. Revista chilena de infectología, 29(2), 138-141. https://dx.doi.org/10.4067/S0716-10182012000200003

Desarrollar un modelo basado redes neuronales profundas que mejoren el grado de especificidad y sensibilidad, así como de precisión y área bajo la curva para la clasificación de pacientes con depresión