Clustering Proposal Support for the COVID-19 Making Decision Process in a Data Demanding Scenario

A. Orjuela-Cañón, *Senior, IEEE*, and O. Perdomo

***Abstract*—The COVID-19 disease surprised the world in the last months due to the number of infections and deaths have been increased in an exponential way. Since the pandemic was established by the World Health Organization, different strategies have been proposed for dealing diverse problems in cities that the coronavirus affected. This work presents a method to decision making support processes, specifically in environment with few data and variables to be considered. Thus, artificial neural networks architectures were employed to cluster the information available in the Bogota city, and provide a tool that allows generating additional findings in a simultaneous mode, and expressed as a visual map. The present proposal reached sensitivity measures around 75%, obtaining 100% for the best cases.**

***Index Terms*—Artificial Intelligence, Clustering, Artificial Neural Networks, Decision Support Systems, COVID-19**

I. Introducción

Actualmente, la pandemia dada por el COVID-19 es una problemática que el mundo está afrontando debido a su rápida transmisión y tasa de letalidad dada en menos de seis meses. Esta enfermedad es un virus proveniente de la familia de los betacoronavirus que causaron el síndrome respiratorio de medio este (MERS-CoV, del inglés *Middle East Respiratory Sindrome*) y el síndrome respiratorio agudo (SARS-CoV, del inglés *Severe Acute Respiratory Syndrome*). En su versión más reciente éste tipo de coronavirus es llamado SARS-CoV-2, produciendo en los seres humanos la enfermedad conocida como COVID-19 y declarada como emergencia de salud pública por la Organización Mundial de la Salud (OMS) a solo dos meses después de registrado el primer brote [1].

El COVID-19 proviene inicialmente de la ciudad de Wuhan, China, y tuvo su primer brote reportado hacia finales de diciembre de 2019. Es un virus que es transmitido de persona a persona mediante el contacto con fluidos corporales como la saliva y ataca el sistema respiratorio del huésped inicialmente. Sus síntomas van desde fiebre y dolor muscular, hasta tos seca y dificultad para respirar. El tiempo que se demora en aparecer dicha sintomatología es en 14 días aproximadamente y dado que es portador del virus mientras no presente síntomas puede propagarlo sin saber (sujetos asintomáticos), su propagación es bastante acelerada [2]–[4].

This work was supported in part by the Universidad del Rosario under Grant BS123456.

A. Orjuela-Cañón and O. Perdomo are with the School of Medicine and Health Sciences from Universidad del Rosario, Bogota, Colombia (e-mail: alvaro.orjuela@urosario.educ.co and oscarj.perdomo@urosario.edu.co).

Tenemos poca información sobre el COVID-19 y el mundo hace esfuerzos para encontrar soluciones que permitan mitigar los diferentes problemas que ocasiona dicha enfermedad. El brote de China es uno de los más analizados debido a que fue el primer lugar donde se generó el virus y donde ya todo está volviendo a la normalidad. Allí, de los 80904 casos confirmados, se registraron un 80% de cuadros con sintomatología leve o moderada, un 13.8% con cuadro clínico severo y un 6.1% de casos críticos. Los casos que están en hospitalización pueden terminar en estado crítico, ocasionando la muerte a partir de fallas respiratoria y multiorgánica [4], [5].

Desde hace alrededor de una década el uso de tecnologías de la información y las comunicaciones se ha incrementado en la toma de decisiones en ambientes gubernamentales y en el campo de la salud [6], [7]. Esto se debe a la digitalización de la información y el avance de nuevas técnicas de su procesamiento como el *Big Data* y la inteligencia artificial (IA) [6], [8], [9]. Sin embargo, en algunas regiones del mundo con países en vía de desarrollo como en América Latina, los sistemas de información muchas veces son precarios y este tipo de soluciones aún continúan distantes en el corto plazo [10]–[12].

A partir de las ventajas y desventajas que conlleva el análisis de información disponible, es posible procesar los datos que se tienen a disposición con herramientas que proporciona la IA. Es así, donde modelos que aprenden de los datos pueden ser bastante aplicados bajo diferentes escenarios como tamaños de muestra pequeños, disminuido número de variables y diferentes niveles de calidad [13], [14]. Dentro del mundo de la IA, existen modelos de aprendizaje no supervisado, donde no se tienen etiquetas de los datos y donde se tiene como objetivo encontrar patrones a partir de las similitudes en los mismos datos [15]. Estos procesos son conocidos como técnicas de agrupamiento, donde el objetivo es relacionar los datos a través del estudio de las relaciones entre sus variables en un espacio *n*-dimensional dado por las mismas características [16].

Las redes neuronales artificiales son un paradigma de la IA que se basan en sistemas conexionistas, reflejando su aprendizaje a través de pesos sinápticos entre unidades básicas del mismo modelo, con variadas arquitecturas y métodos de entrenamiento. Dentro de ese aprendizaje no supervisado, los mapas auto-organizados (SOM, del inglés *Self Organizing Maps*) se destacan por su amplio uso debido a la visualización de los agrupamientos, los cuales son implementados por una etapa posterior con el uso del algoritmo de *k-Means*, agrupando los pesos de la red [17]–[20]. Otra arquitectura conocida son las

redes ART (del inglés *Adaptive Resonance Theory*) en la que algunas operaciones son reemplazadas por difusas, obteniendo redes Fuzzy-ART [21], teniendo aplicaciones usadas en salud, como apoyo al diagnóstico de enfermedades [22]–[24].

La mayoría de trabajos reportados sobre el uso de redes neuronales como posible solución ante el COVID-19 está orientado hacia el procesamiento de imágenes, empleando aprendizaje profundo [25], [26]. Sin embargo, propuestas que utilizan variables clínicas o epidemiológicas son analizadas desde el aprendizaje automático tradicional, usando etiquetas para el entrenamiento de dichos modelos [27], [28]. Estrategias que tengan aplicaciones en ambientes limitados con pocas variables clínicas han sido poco estudiadas bajo este contexto [29]. A pesar de estos esfuerzos, análisis desde el punto de vista de agrupamiento y apoyo en toma de decisión son escasos, presentándose únicamente para hallar grupos en términos de localización espacial [30], [31], para hacer seguimiento de contagio a través de información epidemiológica [32], o como paso intermedio en la medición de impacto económico en procesos de cuarentena [33].

Este documento propone una herramienta computacional que permita apoyar la toma de decisión sobre casos reportados de COVID-19, empleando redes neuronales artificiales. El propósito de la propuesta es plantear soluciones que no comprometan la integridad del profesional de la salud, gestionando el flujo de pacientes que han sido diagnosticados con la enfermedad, y de esta manera hacer remisión a centros hospitalarios, en caso de ser necesario. Al mismo tiempo, gestionar mejor el tiempo que se pueda dedicar al análisis de cada caso. Para esto, se pretende fortalecer la toma de decisión de casos críticos a partir de la visualización de dichos casos en un escenario de tres grupos de riesgo: alto, medio y bajo.

II. Marco Teórico

Este trabajo emplea algoritmos para agrupamiento de datos, principalmente basados en redes neuronales artificiales, que son explicados a continuación. Para el análisis de los agrupamientos también son descritas las métricas empleadas en el presente estudio.

A. Algoritmos de Agrupamiento

El primer algoritmo empleado para el agrupamiento de los datos es el conocido *kMeans*, que utiliza la misma información de los datos para agrupar sus elementos en *k* grupos que mejor describen dicha información [34]. A pesar de ser una estrategia que no es basada en arquitectura de red neuronal, se tomó como base para comparar las estrategias seguidas por las dichas redes.

Dentro de la IA, un campo particular es el de las redes neuronales, presentando ventajas frente a otros métodos de agrupamiento en relación al uso de diferentes tipos de variables en su entrada como categóricas, discretas o continuas [35].

La primera arquitectura de redes neuronales empleada para agrupar los datos fue la Fuzzy-ART, que incorpora operaciones de lógica difusa en redes ART [21]. A diferencia de estas redes, donde los grupos son representados por hiperesferas en el espacio de características, las redes Fuzzy-ART desempeñan esta misma labor a través de hipercubos [36]. De esta forma, los datos de entrada son transformados, empleando un código complementario de la forma:

$$I = (a, a^c) \tag{1}$$

donde *I* representa la nueva entrada que usa la red posterior a la codificación realizada sobre el conjunto original de datos *a*. El número total de variables es duplicado, dada por la operación de complemento realizada en este primer paso. Este procedimiento está bioinspirado en las células en encargadas de la visión en el cerebro, usando respuestas *on-off* que previenen la proliferación de clases [37].

Posterior al proceso de codificación, es realizado un proceso de comparación entre el vector de entrada *I* y los pesos sinápticos de la red que representan cada uno de las neuronas de la arquitectura. La capacidad de plasticidad en estas redes es alta, donde neuronas nuevas se van creando a partir del hallazgo de nuevos patrones o grupos en los datos. Para esto, es medida esa similitud entre el vector de entrada y los pesos sinápticos, si dicha similitud está fuera del radio de vigilancia ρ se actualizan los pesos de dicha neurona, si no, se crea otra neurona de salida. Esto es obtenido a través de la expresión (2), a través de operadores difusos de la forma:

$$\frac{|I \wedge w_j|}{|I|} < \rho \tag{2}$$

donde w_j son los pesos sinápticos de cada *j*-ésimo grupo encontrado en los datos y ρ es el parámetro de vigilancia que contribuye a determinar si los datos corresponden a un grupo representado por las neuronas existentes o si es necesario crear otra neurona.

La segunda alternativa empleada fue dada por los mapas autoorganizados de Kohonen (SOM), que se propusieron a finales de la década del noventa [38]. Este tipo de redes son especialmente usadas debido a que ofrecen la posibilidad de representar un hiperespacio generado por *n* características de los datos en un plano de dos dimensiones, preservando la proximidad del espacio de datos original, haciendo que sea posible visualizar los agrupamientos [39].

Las redes SOM son entrenadas a través de tres grandes etapas: competitiva, cooperativa y adaptativa. La primera etapa realiza una comparación del vector de entrada con los pesos sinápticos que representan cada una de las neuronas del mapa y aquella que mejor represente el vector se denomina como neurona ganadora (BMU del inglés *Best Match Unit*). Esta operación se realiza a través de la expresión (2) de la forma:

$$BMU = min(x(t) - w_i(t)) \tag{2}$$

donde *x* es el vector de entrada y *w* son los pesos sinápticos de la *i*-ésima neurona de la red. El proceso cooperativo se basa en actualizar los pesos sinápticos cercanas a la BMU a través de una función de vecindad unimodal. Esto se da a partir de la expresión (3) de la forma:

$$h_{ij}(t+1) = exp\left(-d_{ij}^2 / 2\sigma^2(t)\right) \tag{3}$$

donde *h(t)* es la función de vecindad, *d* es la distancia entre la entrada y los pesos sinápticos de la BMU y σ es la base de dicha función y que abarca otras neuronas alrededor de la BMU.

Cualquier función unimodal puede emplearse para esta labor, sin embargo, una de las más utilizadas es la función Gaussiana debido a que introduce una representación no lineal entre los datos y lo que aprende la red en sus pesos sinápticos a través del entrenamiento.

Finalmente, la actualización de los pesos sinápticos de las neuronas que comprenden esa región de vecindad de la red se realiza a través de la ecuación (4) de la forma:

$$w_i(t+1) = w_i + \eta(t)(x(t) - w_i(t)) \quad (4)$$

donde $w_i(t+1)$ y w_i son los vectores de pesos sinápticos actuales y previos en el proceso de entrenamiento. La velocidad de esta actualización se determina a través de η que también varía en el tiempo, teniendo un menor cambio hacia el final de la convergencia realizando un ajuste fino de dichos pesos.

B. *Análisis del Agrupamiento*

Posterior al entrenamiento de los modelos usados para el agrupamiento expuestos en la subsección anterior, se emplearon dos diferentes medidas que se emplean para medir la calidad del agrupamiento de los datos.

Una medida empleada para estas tareas es la proporcionada por Davies y Bouldin, donde se mide la dispersión de los grupos normalizada por la separación entre los grupos [40]. Para obtener esta medida, se empleó la expresión (5), representando este índice así:

$$DB = \frac{1}{N}\sum_{i=1, i\neq j}^{N} max\left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)}\right) \quad (5)$$

donde N es el número de grupos, σ_i es la desviación estándar de cada uno de los grupos formados (dispersión intra-grupos) y $d(c_i,c_j)$ es la distancia entre los centros de los grupos (dispersión entre-grupos). Un número pequeño en este índice significa valores con grupos con poca dispersión al interior de cada uno de ellos y alejados entre ellos. Esto quiere decir que entre menor es este índice, mejor será la calidad del agrupamiento obtenido.

También, el índice silueta se utilizó para realizar una medición de la calidad del agrupamiento de manera tal que permita obtener información de la relación de los datos de un grupo con respecto a los demás grupos encontrados [41]. Este índice se obtiene de (8) a partir de (6) y (7) de la forma:

$$a(i) = \frac{1}{|C_i|-1}\sum_{j=1, i\neq j}^{C_i} d(i,j) \quad (6)$$

$$b(i) = min_{k\neq i}\frac{1}{|C_k|}\sum_{j=1}^{C_k} d(i,j) \quad (7)$$

$$Silueta = \frac{b(i)-a(i)}{max\{a(i),b(i)\}} \quad (8)$$

donde $a(i)$ es la medida del correcto agrupamiento del dato i a su grupo o clúster C_i, y $b(i)$ es la medida entre el mismo dato i con respecto al clúster más cercano C_k, para cuando el número de grupos es mayor a 1. Un valor elevado en este índice representa una mejor calidad de agrupamiento de los datos. Para encontrar este valor se trabajó sobre el valor máximo de la media de los valores de este índice, similar a la propuesta de Kaufman et al. [42].

III. Metodología

La metodología empleada está basada en tres etapas: obtención y tratamiento de los datos, análisis de agrupamiento no supervisado y análisis agrupamiento semi-supervisado que determina tres grupos de riesgo. La Figura 1 muestra estas etapas de forma gráfica.

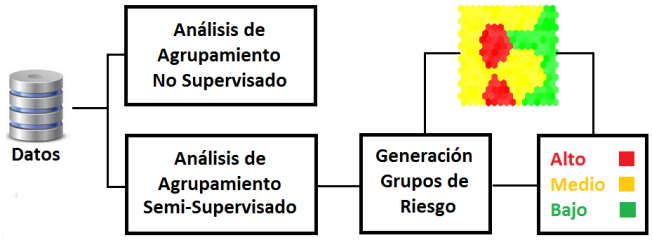

Fig. 1. Metodología empleada en el presente trabajo.

A. *Obtención y Tratamiento de Datos*

Colombia, es un país que tiene política de datos abierta [43], esto quiere decir que los datos empleados en este trabajo fueron obtenidos de la Alcaldía Mayor de Bogotá, que hace un seguimiento a los casos reportados, recuperados y fallecidos del COVID-19 [44]. Esta ciudad es la sexta de entre las ciudades más pobladas de América Latina con 9.2 millones de habitantes, una extensión de 1.775 km^2, con una densidad poblacional de 26643 habitantes por km^2 y más casos reportados de COVID-19 en el país a la fecha [44].

El sistema actualiza los datos cada 24 horas, reportando información de ocho diferentes variables. Para el presente estudio tomamos seis de ellas: sexo, edad, localidad (21 en total), tipo de caso, en dónde se está llevando su tratamiento y su estado. Las dos variables restantes no fueron incluidas debido a que corresponden a fecha de diagnóstico y ciudad, que para todos los casos resulta ser la misma.

Los datos fueron codificados usando la estrategia *one-hot encoder* y de esta forma no generar una categorización ordinal, evitando involucrar un orden a las variables [45]. Así, cada caso de COVID-19 registrado en el sistema de base de datos oficial del distrito, es representado por un vector de ceros y unos que significan los valores asociados a cada una de las variables. La variable edad se utilizó numérica, normalizada por el valor máximo encontrado para esa misma variable. Finalmente, se tiene un vector de 36 valores de acuerdo a las características mencionadas (Tabla I).

En total fueron 3821 casos reportados al 7 de mayo de 2020, que se dividieron en un conjunto de desarrollo para el ajuste y entrenamiento de los modelos empleados, considerando casos hasta el 30 de abril, es decir, 2795 casos (73.15%). Esta porción de datos es usada para el entrenamiento y generación de modelos, dejando el 26.85% restante para validar los modelos obtenidos. Esta división de tipo *hold-out* se realiza manteniendo el contexto de la actual emergencia, donde no se tiene acceso al total de datos para dividirlos, sino que nuevos datos van llegando. De esta forma, se genera información a partir de los datos disponibles en esa ventana de tiempo que sirvan para la toma de decisiones para ventanas posteriores.

B. *Agrupamiento No Supervisado*

Una exploración de los datos es necesaria como primer paso, donde se emplearon los algoritmos de *kMeans*, SOM y Fuzzy-ART para observar de qué manera dichos datos se relacionan

entre sí. Este proceso se repitió para cada algoritmo 100 veces para examinar la influencia de parámetros de inicialización. En cada caso fue determinado el número de grupos con el mejor índice de agrupamiento de acuerdo a las métricas Davies Bouldin y Silueta explicadas en la sección anterior.

TABLA I
VARIABLES UTILIZADAS
EN EL ESTUDIO DE AGRUPAMIENTO

Variable	Valores
Localidad	Antonio Nariño Barrios Unidos Bosa Chapinero Ciudad Bolívar Engativá Fontibón Fuera de Bogotá Kennedy La Candelaria Los Mártires Puente Aranda Rafael Uribe Uribe San Cristóbal Santa Fe Sin Dato Suba Teusaquillo Tunjuelito Usaquén Usme
Edad	Valores entre 1 y 103 años
Sexo	Masculino Femenino Desconocido
Tipo de Caso	En Estudio Importado Relacionado
Ubicación	Casa Hospital Hospital UCI
Estado	Crítico Fallecido Moderado Recuperado Severo

C. Agrupamiento Semi-Supervisado

De forma simultánea a la exploración realizada por el agrupamiento no supervisado, otra alternativa es planteada con el agrupamiento semi-supervisado, donde de antemano se determinan los grupos que debe tener el proceso. En este caso, los datos que se tenían para el análisis de agrupamiento no supervisado se analizaron a manera de *triage*, formando tres grupos de riesgo de acuerdo a su estado final: crítico, severo, fallecido, moderado y recuperado (ver Tabla I). La Tabla II muestra cómo se determinaron los grupos de riesgo: alto, medio y bajo. Tres grupos de riesgo fueron implementados de acuerdo a un *triage* y debido a que la cantidad de elementos en los grupos tienen mayor proporcionalidad al comparar las agrupaciones.

La arquitectura Fuzzy-ART fue forzada entonces a un agrupamiento de tres clústeres dado por tres neuronas en la salida, haciendo que cada uno represente un grupo de riesgo: alto, medio y bajo. También, otra arquitectura fue empleada basada en el mapa SOM, agrupando los pesos sinápticos de la red a través de la estrategia SOM+*kMeans,* obteniendo tres grupos de riesgo. Este modelo es también aprovechada, debido a las ventajas que presenta en cuanto a visualización que se tienen con el mapa proporcionado por la red SOM, en la cual es posible ilustrar la representación del espacio de 36 dimensiones en un arreglo bidimensional que preserva las distancias entre los datos.

TABLA II
GRUPOS DE RIESGO ANALIZADOS

Grupos de Riesgo (color)	Valores
Alto (Rojo)	Fallecido Crítico Severo
Medio (Amarillo)	Moderado
Bajo (Verde)	Recuperado

Posterior a esto, se utilizó el conjunto de datos que se dejó para validar el agrupamiento semi-supervisado con 1026 registros que no fueron usados en el desarrollo del agrupamiento no supervisado. Debido a la naturaleza de los datos, solo se tienen registros para el grupo de riesgo alto, con 105 registros, y 921 para medio riesgo. Esto, porque no se han registrado los casos recuperados de las últimas semanas. Esto no representaría problema debido a que se realizó el modelo semi-supervisado habilitado para encontrar los casos de alto riesgo, permitiendo obtener información adicional que permita la gestión de dichos sujetos.

De esta forma, tomando información del estado de los sujetos en el grupo de desarrollo, se catalogaron los grupos. Así mismo, para etiquetar los grupos se tomó en cuenta el número de activaciones del número de neuronas entre sí. También, aprovechando la información visual de los tres grupos de riesgo formados, se proyecta la información de las variables de interés para encontrar su relación con cada grupo de riesgo.

Finalmente, se evalúan medidas de sensibilidad y especificidad con los datos del grupo de validación de los modelos obtenidos, analizando únicamente información de los grupos de riesgo alto y medio. Es importante notar que para el proceso de validación la información correspondiente a estado y ubicación fue convertida a cero, simulando un escenario real donde no se tienen estos datos y se quiere saber en qué grupo de riesgo estaría el sujeto diagnosticado de COVID-19 y saber que tratamiento debería recibir el paciente y que recursos hospitalarios (disponibilidad de camas en hospitales y UCI) deberían estar listos a proveerse.

IV. RESULTADOS

Los resultados son mostrados de acuerdo a los dos escenarios descritos en la metodología. Primero el entrenamiento no supervisado es presentado para poder entender el agrupamiento con las propuestas detalladas y posteriormente el ajuste semi-supervisado es mostrado apoyado de los resultados visuales que se obtuvieron.

A. Agrupamiento No Supervisado

La Tabla III muestra los resultados para el número de grupos obtenido con las diferentes estrategias usadas para el agrupamiento y las correspondientes medidas de la calidad de agrupación. Para evitar sesgo en la obtención de los resultados, cada proceso de agrupamiento tuvo 100 inicializaciones diferentes y poder evaluar su estadística. En cada inicialización se identificaron el número de grupos dado por los índices de calidad de agrupación descritos. Debido a ello, y que los resultados no siguen un patrón, valores de media (μ), desviación estándar (σ) y mediana se presenta en la Tabla III.

TABLA III
NÚMEROS DE GRUPOS ENCONTRADOS A PARTIR DE LOS ÍNDICES DE CALIDAD DE AGRUPAMIENTO

Propuesta	Davies Bouldin			Silueta		
	μ	σ	mediana	μ	σ	mediana
kMeans	16.2	2.6	17	8.5	4.9	8
SOM + *kMeans*	17	2	17	2.6	1.6	2
Fuzzy ART	7.2	7.1	2.5	3.4	2.7	3

B. Agrupamiento Semi-Supervisado

De forma similar al caso no supervisado, se realizaron cien inicializaciones para ver la fluctuación estadística de los resultados. La Tabla IV muestra estos resultados para sensibilidad, especificidad y exactitud para el conjunto de datos usados en la validación. Para la obtención de dichas medidas, los valores del grupo de riesgo alto fueron usados para obtener la sensibilidad, donde se procura encontrar esos patrones de sujetos reportados como de estado crítico, severo o fallecido. La especificidad se obtuvo a través de los valores del grupo de riesgo moderado, debido a que para este conjunto de datos de validación no estaban confirmados los casos recuperados.

TABLA IV
RESULTADOS PARA LOS MODELOS SEM-SUPERVISADOS

Modelo	Medidas Empleadas en Conjunto de Validación		
	Sensibilidad (μ±σ)	Especificidad (μ±σ)	Exactitud (μ±σ)
Fuzzy-ART	70+12	27±11	48±12
SOM+*kMeans*	75+17	47±19	61±18

Empleando el mejor modelo obtenido en el conjunto de desarrollo de los modelos y los colores usados fueron los descritos en la Tabla II, son graficados los grupos de riesgo en el mapa SOM entrenado y agrupado con *kMeans*. Las Figuras 2 y 3 muestran las tres regiones que describen cada grupo de riesgo para el mejor resultado. Los puntos negros son los datos del conjunto de desarrollo que activan diferentes neuronas en los grupos de riesgo alto, medio y bajo (ver Figura 2).

De igual forma, empleando el mismo modelo y los datos del conjunto de validación, se proyectan estos datos activando los grupos de riesgo alto y moderado (Figura 3). Es importante notar aquí que para obtención de estas proyecciones, la información de ubicación y estado (ver Tabla I) fue puesta toda como cero para simular el escenario real, donde no se conoce dicha información al momento del diagnóstico de los pacientes de COVID-19. Es posible ver una sensibilidad del 100 % y una sensibilidad del 86% (ver Figura 3-b puntos fuera de la zona roja). En esta última figura es posible ver como dos neuronas (126 activaciones) del grupo de alto riesgo se activaron con datos del grupo de riesgo medio, siendo calculados como errores.

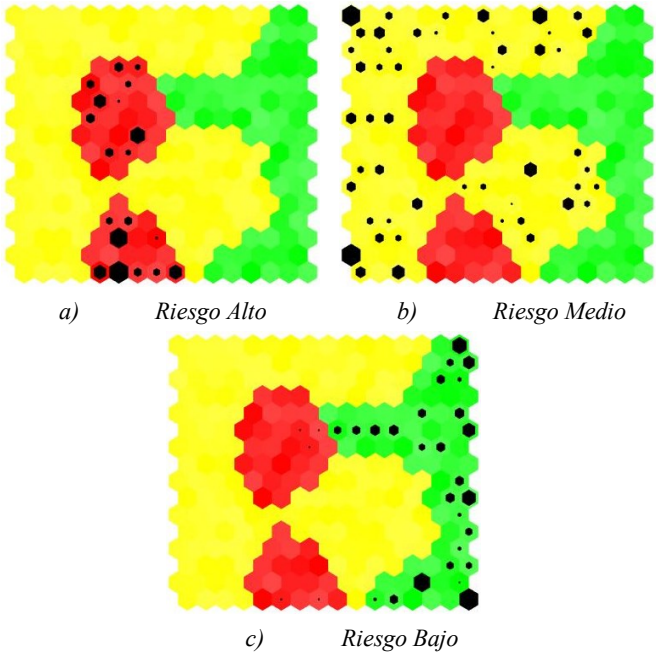

Fig. 2. Datos de los sujetos del conjunto de desarrollo y del grupo de riesgo: a) alto (rojo); b) medio (amarillo); c) bajo (verde)

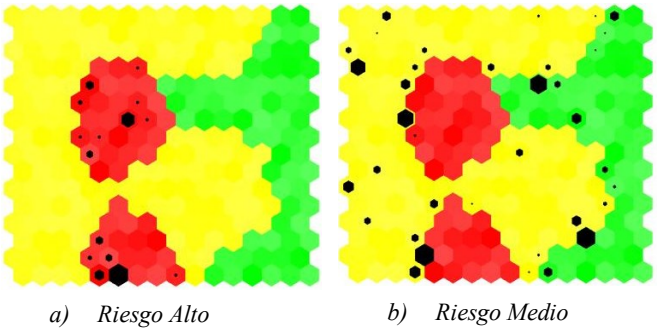

Fig. 3. Datos de los sujetos del conjunto de validación y del grupo de riesgo a) alto (rojo) y b) medio (amarillo).

Aprovechando la información del mapa agrupado en tres regiones conocidas y relacionadas con el grupo de riesgo, se analizaron las proyecciones de algunas variables de interés como: estado (Figura 4), edad (Figura 5) y localidad con tres valores, correspondientes a las de mayor población en la ciudad como lo son: Kennedy, Engativá y Bosa (Figura 6). De esta forma, es posible ver la relación de los valores de dichas variables con respecto a cada uno de los grupos en el mapa de riesgo dado por los tres colores.

V. Discusión

Una primera observación sobre los resultados está relacionada con el uso de variables categóricas ordinales. Esto, como fue observado, amplía el hiperespacio de las características de forma disyuntiva debido a la naturaleza de los datos. De igual forma, algunos algoritmos de agrupamiento pueden tener problemas al realizar grupos bajo este escenario [34], [46]. Existen algunas adaptaciones al respecto, principalmente para el algoritmo de *kMeans*, donde ajustes son realizados para tratar este problema [47], [48]. Sin embargo, el algoritmo de *kMeans* se usó como punto de comparación para las técnicas basadas en redes neuronales, teniendo resultados comparables para las medidas de calidad usadas (ver Tabla III). Así mismo, a partir de dicho resultado en la técnica supervisada, la misma Tabla III evidencia como el uso de tres grupos puede ser respaldada por dichas medidas de calidad empleadas.

En cuanto a resultados de sensibilidad, especificidad y exactitud para la estrategia realizada en el presente trabajo, la adaptación que se implementó en la unión de la estrategia SOM+*kMeans* permitió mejores resultados frente a los modelos con redes Fuzzy-ART. Comparada con aplicaciones similares, donde se busca relacionar las variables estudiadas y grupos de riesgo se han tratado en estudios anteriores ante enfermedades trasmisibles como la tuberculosis [20], [22], [24]. Esto demuestra la potencialidad de estas herramientas con enfermedades transmisibles como en este caso lo es el COVID-19, donde hay pocos desarrollos cercanos a los abordados en el presente trabajo. Esto puede verse en [29], donde al emplear variables similares y técnicas de aprendizaje automático, los autores alcanzaron exactitudes entre el 61% y el 71%. Otro estudio basado en regresión logística, usando pocas variables como en el presente caso pero incluyendo variables clínicas alcanzó una sensibilidad entre el 62% y 69% y una especificidad entre el 82% y el 85% [49]. Modelos similares para diagnóstico y prognosis, alcanzaron rangos entre el 65% al 99% con el uso de más variables que incluían datos de laboratorio y comorbilidades. Sin embargo, sus propuestas fueron variadas, empleando modelos estadísticos y de aprendizaje automático clásico diferente a agrupamiento [28]. Más trabajos que afronten el problema desde ésta perspectiva con fines similares son escasos actualmente [25], [49].

Otra ventaja mostrada por la alternativa dada por la estrategia SOM+*kMeans* es la visualización de los grupos de riesgo. Esto permite ver patrones a partir de los datos de los sujetos que son aprendidos por el mapa, siendo una herramienta importante para toma de decisión debido a que permite planear acciones y demanda logística a partir del perfil aprendido de los sujetos que ya presentan condiciones de fallecido, estado crítico o severo. Esto, es evidenciado por los resultados, donde se tuvo una sensibilidad del 100% en el mejor caso, siendo bastante útil para encontrar sujetos que pueda desarrollar consecuencias fatales, posterior a la detección de la enfermedad. Complementario a este resultado, al realizar un análisis sobre la especificidad, con resultados más conservadores, es posible interpretar dichos valores de falsos positivos en zonas colindantes entre los grupos de riesgo alto y medio (ver Figura 3-b). Allí se muestra como 126 (13.7%) datos de sujetos con riesgo medio activan la región de alto riesgo, evidenciando como dos neuronas del grupo de riesgo alto (zona roja) están bastante próximas a la región de medio riego (zona amarilla). En este caso, el mapa proporciona a los profesionales encargados de la toma de decisión una mayor información que en este caso significa ahorro de tiempo, debido a que no es necesario analizar todos los casos positivos de COVID-19 de ese grupo (921 sujetos), usando patrones de sujetos de riesgo alto. Al mismo tiempo, esta información puede ser empleada en proyecciones de futuro uso de recursos hospitalarios que deberán estar disponibles para atender dicha demanda.

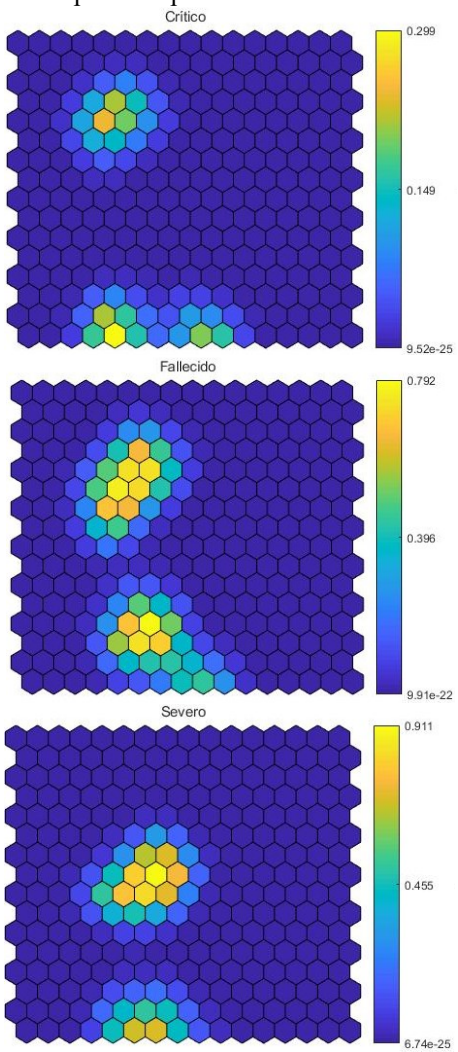

Fig. 4. Variable Estado (Crítico, Fallecido y Severo) y su proyección en el mapa entrenado y agrupado.

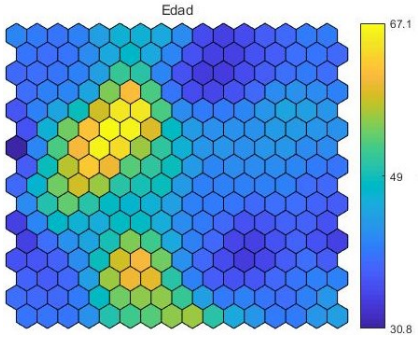

Fig. 5. Variable Edad y su proyección en el mapa entrenado y agrupado.

Otro tipo de análisis es realizado al cruzar la visualización de las variables involucradas y grupos de riesgo, donde se puede apreciar que dichas activaciones de falsos positivos estarían relacionados con patrones de sujetos que fallecieron o que tienen estado severo (ver Figura 4 para variable Estado). De igual forma, las neuronas activas en la región de bajo riesgo (zona verde, Figura 3) mostrarían los sujetos con tendencia a recuperación debido a que esta zona del mapa representa dichos individuos dado el aprendizaje que se tuvo en el entrenamiento.

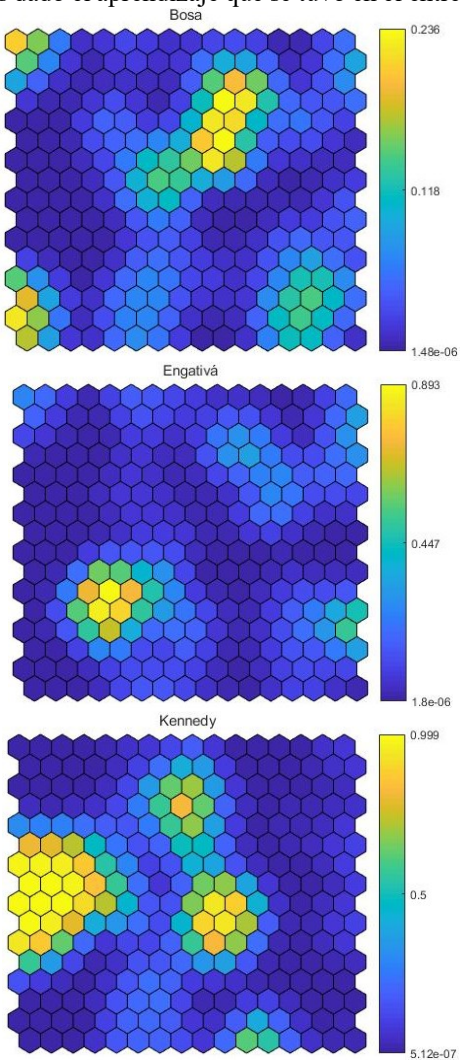

Fig. 6. Variable Localidad (Bosa, Engativá y Kennedy) y su proyección en el mapa entrenado y agrupado.

Para la variable localidad, es posible ver como el grupo de alto riesgo está relacionado con los estados: crítico, severo y fallecido, haciendo énfasis cada uno de ellos en zonas diferentes de dicho grupo de riesgo (ver Figura 6). También, de las localidades analizadas, la que presenta más riesgo es la de Engativá, donde sus más altos valores (zonas amarillas) están en la región de alto riesgo. De acuerdo a los resultados en otros países y como característica del COVID-19 se sabe que tiene efectos más letales en la población adulta mayor. Esto se puede ver en la Figura 5, donde las más altas edades (zonas amarillas) están en la región de alto riesgo y se solapan con el estado fallecido y severo del grupo de riesgo alto.

VI. Conclusiones

El presente trabajo mostró como una estrategia semi-supervisada puede contribuir a la toma de decisión para el gerenciamiento de sujetos diagnósticos con COVID-19 en un entorno de poca información. Para esto, se utilizaron estrategias basadas en el entrenamiento de redes neuronales artificiales con datos disponibles para determinar tres grupos de riesgo: alto, medio y bajo. Los resultados mostraron, en el mejor de los casos, una alta sensibilidad, obteniendo los casos de riesgo alto a los que se les debería realizar un seguimiento más detallado para evitar consecuencias fatales.

Dentro de las limitaciones del presente estudio, se encuentran los pocos datos y la forma de validar los modelos. Sin embargo, la validación se realizó basado en como los datos son obtenidos. Esto quiere decir que los datos empleados para realizar la experimentación nunca va a ser un tema cerrado, y debido a esto se presentaron de esta forma. Como van llegando se va analizando el conjunto de datos disponible, planteando como trabajo futuro una ampliación de la ventana de tiempo para involucrar más datos, y validar una vez más los resultados obtenidos hasta el momento.

Reconocimientos

Los autores quieren agradecer a la Escuela de Medicina y Ciencias de la Salud de la Universidad del Rosario por su apoyo en la realización de este trabajo en el marco del proyecto titulado "*Prevención, diagnóstico y asistencial virtual para COVID 19*", dada la dedicación de los investigadores.

Referencias

[1] W. H. Organization and others, "WHO statement regarding cluster of pneumonia cases in Wuhan, China," *Beijing WHO*, vol. 9, 2020.

[2] N. Chen *et al.*, "Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study," *Lancet*, vol. 395, no. 10223, pp. 507–513, 2020.

[3] A. Patel, D. B. Jernigan, and others, "Initial public health response and interim clinical guidance for the 2019 novel coronavirus outbreak--United States, December 31, 2019--February 4, 2020," *Morb. Mortal. Wkly. Rep.*, vol. 69, no. 5, p. 140, 2020.

[4] K. J. Smereka, Jacek and Szarpak, Lukasz and Filipiak, "Modern medicine in COVID-19 era," *Disaster Emerg. Med. J.*, vol. 5, no. 2, pp. 103–105, 2020.

[5] S. Zaim, J. H. Chong, V. Sankaranarayanan, and A. Harky, "COVID-19 and Multiorgan Response," *Curr. Probl. Cardiol.*, vol. 45, no. 8, p. 100618, 2020.

[6] B. N. Silva, M. Khan, and K. Han, "Towards sustainable smart cities: A review of trends, architectures, components, and open challenges in smart cities," *Sustain. Cities Soc.*, vol. 38, pp. 697–713, 2018.

[7] M. I. Pramanik, R. Y. K. Lau, H. Demirkan, and M. A. K. Azad, "Smart health: Big data enabled health paradigm within smart cities," *Expert Syst. Appl.*, vol. 87, pp. 370–383, 2017.

[8] Y. Wang *et al.*, "Clinical information extraction applications: a literature review," *J. Biomed. Inform.*, vol. 77, pp. 34–49, 2018.

[9] C. Xiao, E. Choi, and J. Sun, "Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review," *J. Am. Med. Informatics Assoc.*, vol. 25, no. 10, pp. 1419–1428, 2018.

[10] A. Manya and P. Nielsen, "Reporting practices and data quality in health information systems in developing countries: an exploratory case study in Kenya," *J. Health Inform. Dev. Ctries.*, vol. 10, no. 1, 2016.

[11] Y. Glèlè Ahanhanzo, E.-M. Ouendo, A. Kpozèhouen, A. Levêque, M. Makoutodé, and M. Dramaix-Wilmet, "Data quality assessment in the

routine health information system: an application of the lot quality assurance sampling in Benin," *Health Policy Plan.*, vol. 30, no. 7, pp. 837–843, 2015.

[12] J. Macinko, F. C. Guanais, P. Mullachery, and G. Jimenez, "Gaps in primary care and health system performance in six Latin American and Caribbean countries," *Health Aff.*, vol. 35, no. 8, pp. 1513–1521, 2016.

[13] N. Peek, C. Combi, R. Marin, and R. Bellazzi, "Thirty years of artificial intelligence in medicine (AIME) conferences: A review of research themes," *Artif. Intell. Med.*, vol. 65, no. 1, pp. 61–73, 2015.

[14] F. Jiang *et al.*, "Artificial intelligence in healthcare: past, present and future," *Stroke Vasc. Neurol.*, vol. 2, no. 4, pp. 230–243, 2017.

[15] J. A. Hartigan, *Clustering algorithms*. 1975.

[16] D. Xu and Y. Tian, "A comprehensive survey of clustering algorithms," *Ann. Data Sci.*, vol. 2, no. 2, pp. 165–193, 2015.

[17] E. Elveren and N. Yumuvak, "Tuberculosis disease diagnosis using artificial neural network trained with genetic algorithm," *J. Med. Syst.*, vol. 35, no. 3, pp. 329–332, 2011.

[18] P. Venkatesan and M. Mullai, "Clustering of Disease Data base using Self Organizing Maps and Logical Inferences," *Indian J. Autom. Artif. Intell.*, vol. 1, no. 1, pp. 2–6, 2013.

[19] S.-L. Shieh and I.-E. Liao, "A new approach for data clustering and visualization using self-organizing maps," *Expert Syst. Appl.*, vol. 39, no. 15, pp. 11924–11933, 2012.

[20] F. S. Aguiar, R. C. Torres, J. V. F. Pinto, A. L. Kritski, J. M. Seixas, and F. C. Q. Mello, "Development of two artificial neural network models to support the diagnosis of pulmonary tuberculosis in hospitalized patients in Rio de Janeiro, Brazil," *Med. Biol. Eng. Comput.*, vol. 54, no. 11, pp. 1751–1759, 2016.

[21] G. A. Carpenter, S. Grossberg, and D. B. Rosen, "Fuzzy ART: Fast stable learning and categorization of analog patterns by an adaptive resonance system," *Neural networks*, vol. 4, no. 6, pp. 759–771, 1991.

[22] A. D. Orjuela-Cañón, J. E. C. Mendoza, C. E. A. García, and E. P. V. Vela, "Tuberculosis diagnosis support analysis for precarious health information systems," *Comput. Methods Programs Biomed.*, 2018.

[23] A. D. Orjuela-Cañón and J. de Seixas, "Fuzzy-ART neural networks for triage in pleural tuberculosis," in *Health Care Exchanges (PAHCE), 2013 Pan American*, 2013, pp. 1–4.

[24] A. D. Orjuela-Cañón, J. M. de Seixas, and A. Trajman, "SOM Neural Networks as a Tool in Pleural Tuberculosis Diagnostic," in *Annals of the 11th Brazilian Congress on Computational Intelligence*, 2013, pp. 1–5.

[25] Y. Mohamadou, A. Halidou, and P. T. Kapen, "A review of mathematical modeling, artificial intelligence and datasets used in the study, prediction and management of COVID-19," *Appl. Intell.*, pp. 1–13, 2020.

[26] A. Kumar, P. K. Gupta, and A. Srivastava, "A review of modern technologies for tackling COVID-19 pandemic," *Diabetes Metab. Syndr. Clin. Res. Rev.*, vol. 14, no. 4, pp. 569–573, 2020.

[27] S. Debnath *et al.*, "Machine learning to assist clinical decision-making during the COVID-19 pandemic," *Bioelectron. Med.*, vol. 6, no. 1, pp. 1–8, 2020.

[28] L. Wynants *et al.*, "Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal," *bmj*, vol. 369, 2020.

[29] M. Nemati, J. Ansary, and N. Nemati, "Machine Learning Approaches in COVID-19 Survival Analysis and Discharge Time Likelihood Prediction using Clinical Data," *Patterns*, p. 100074, 2020.

[30] R. Chen *et al.*, "Risk factors of fatal outcome in hospitalized subjects with coronavirus disease 2019 from a nationwide analysis in China," *Chest*, 2020.

[31] M. R. Desjardins, A. Hohl, and E. M. Delmelle, "Rapid surveillance of COVID-19 in the United States using a prospective space-time scan statistic: Detecting and evaluating emerging clusters," *Appl. Geogr.*, p. 102202, 2020.

[32] S. E. F. Yong *et al.*, "Connecting clusters of COVID-19: an epidemiological and serological investigation," *Lancet Infect. Dis.*, 2020.

[33] M. A. Rahman, "Data-driven dynamic clustering framework for mitigating the adverse economic impact of Covid-19 lockdown practices," *Sustain. Cities Soc.*, p. 102372, 2020.

[34] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient k-means clustering algorithm: Analysis and implementation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 881–892, 2002.

[35] S. Haykin, *Neural Networks and Learning Machines*, 3ra ed. Pearson, 2009.

[36] C. Budayan, I. Dikmen, and M. T. Birgonul, "Comparing the performance of traditional cluster analysis, self-organizing maps and fuzzy C-means method for strategic grouping," *Expert Syst. Appl.*, vol. 36, no. 9, pp. 11772–11781, 2009.

[37] J. Huang, M. Georgiopoulos, and G. L. Heileman, "Fuzzy ART properties," *Neural Networks*, vol. 8, no. 2, pp. 203–213, 1995.

[38] T. Kohonen, "Self-organizing maps, ser," *Inf. Sci. Berlin Springer*, vol. 30, 2001.

[39] M. Zribi, Y. Boujelbene, I. Abdelkafi, and R. Feki, "The self-organizing maps of Kohonen in the medical classification," in *Sciences of Electronics, Technologies of Information and Telecommunications (SETIT), 2012 6th International Conference on*, 2012, pp. 852–856.

[40] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Trans. Pattern Anal. Mach. Intell.*, no. 2, pp. 224–227, 1979.

[41] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, 1987.

[42] L. Kaufman and P. J. Rousseeuw, *Finding groups in data: an introduction to cluster analysis*, vol. 344. John Wiley & Sons, 2009.

[43] Ministerio de Tecnologías de la Información y las Comunicaciones, "Guía para el uso y aprovechamiento de Datos Abiertos en Colombia." 2016.

[44] Alcaldía Mayor de Bogotá, "COVID-19 en Bogotá." 2020.

[45] A. Agresti, *An introduction to categorical data analysis*, vol. 135. Wiley New York, 1996.

[46] A. Ahmad and L. Dey, "A k-mean clustering algorithm for mixed numeric and categorical data," *Data Knowl. Eng.*, vol. 63, no. 2, pp. 503–527, 2007.

[47] T.-H. T. Nguyen, D.-T. Dinh, S. Sriboonchitta, and V.-N. Huynh, "A method for k-means-like clustering of categorical data," *J. Ambient Intell. Humaniz. Comput.*, pp. 1–11, 2019.

[48] S. Khanmohammadi, N. Adibeig, and S. Shanehbandy, "An improved overlapping k-means clustering method for medical applications," *Expert Syst. Appl.*, vol. 67, pp. 12–18, 2017.

[49] G. Cherry *et al.*, "Loss of smell and taste: a new marker of COVID-19? Tracking reduced sense of smell during the coronavirus pandemic using search trends," *Expert Rev. Anti. Infect. Ther.*, vol. 18, no. 11, pp. 1165–1170, 2020

Alvaro D. Orjuela-Cañón (StM' 00-M'06–SM'17) nació en Bogotá D.C., Colombia en 1981. Recibió su grado de ingeniería electrónica de la Universidad Distrital Francisco José de Caldas in Bogotá D.C., en el año 2006. Realizó su maestría y doctorado en la Universidad e Federal do Rio de Janeiro, RJ, Brasil en 2009 y 2015, respectivamente. Actualmente hace parte del programa de ingeniería biomédica de la Escuela de Medicina y Ciencias de la Salud de la Universidad del Rosario en la misma ciudad. Tiene intereses en áreas como el procesamiento digital de señales biomédicas, inteligencia computacional en salud, así como energías alternativas. Dr. Orjuela-Cañón es miembro de IEEE en los últimos 18 años. Participando activamente en el capítulo profesional de inteligencia computacional IEEE-CIS.

Oscar J. Perdomo nació en Neiva, Huila, Colombia en 1986. Recibió su título de ingeniero electrónico de la Universidad Sur colombiana en el año 2009 en Neiva - Huila, Colombia. Tiene una maestría en ingeniería biomédica de la Universidad e Federal de Santa Catarina, en Florianópolis, Brasil en 2012, doctorado en ingeniería de sistemas y computación de la Universidad Nacional de Colombia en Bogotá DC. Es profesor asistente del programa de ingeniería biomédica hace parte del programa de ingeniería biomédica de la Escuela de Medicina y Ciencias de la Salud de la Universidad del Rosario en la misma ciudad. Dentro de sus áreas de interés están la ingeniería biomédica, la bioinstrumentación, big data, aprendizaje de máquina y aprendizaje profundo en salud.