



- (51) **International Patent Classification:**
G06N 20/00 (2019.01)
- (21) **International Application Number:**
PCT/US2022/022298
- (22) **International Filing Date:**
29 March 2022 (29.03.2022)
- (25) **Filing Language:** English
- (26) **Publication Language:** English
- (30) **Priority Data:**
63/167,479 29 March 2021 (29.03.2021) US
- (71) **Applicant: THE REGENTS OF THE UNIVERSITY OF CALIFORNIA** [US/US]; 1111 Franklin Street, Twelfth Floor, Oakland, California 94607 (US).
- (72) **Inventors: BARANZINI, Sergio E.**; c/o The Regents of the University of California, 1111 Franklin Street, 12th Floor, Oakland, California 94607 (US). **NELSON, Charlotte A.**; c/o The Regents of the University of California, 1111 Franklin Street, 12th Floor, Oakland, California 94607 (US).
- (74) **Agent: BAUMGARTNER, John Adam et al.**; Kilpatrick Townsend & Stockton LLP, Mailstop: IP Docketing 22, 1100 Peachtree Street, Suite 2800, Atlanta, Georgia 30309 (US).
- (81) **Designated States** (*unless otherwise indicated, for every kind of national protection available*): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA,

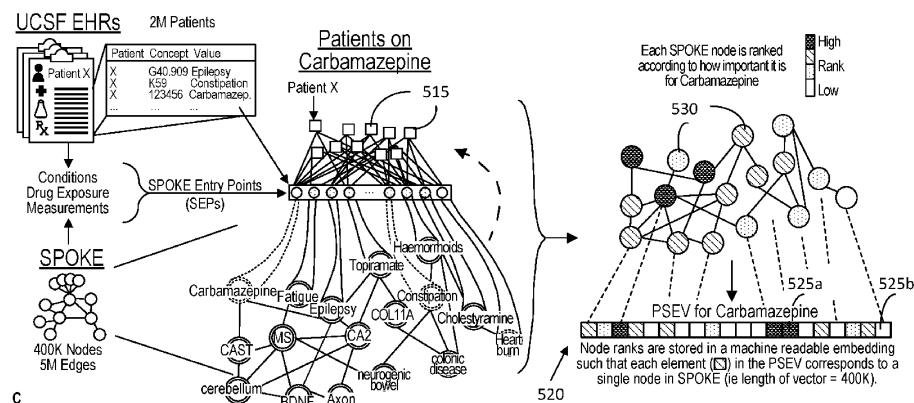
(54) **Title:** GRAPH DATABASE TECHNIQUES FOR MACHINE LEARNING

FIG. 5

(57) **Abstract:** A process is provided for using a graph database (e.g., SPOKE) to generate training vectors (SPOKEsigs) and train a machine learning model to classify biological entities. A cohort's input data records (EHRs) are compared to graph database nodes to identify overlapping concepts. Entry nodes (SEPs) associated with these overlapping concepts are used to generate propagated entry vectors (PSEVs) that encode the importance of each database node for a particular cohort, which helps train the model with only relevant information. Further, the propagated entry vectors for a given entity with a known classification can be aggregated to create training vectors. The training vectors are used as inputs to train a machine learning model. Biological entities with an unknown classification can be classified with a trained machine learning model. Entity signature vectors are generated for entities without a classification and input into the trained machine learning model to obtain a classification.

[Continued on next page]

MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI,
NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU,
RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM,
TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM,
ZW.

(84) Designated States (*unless otherwise indicated, for every kind of regional protection available*): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

- *with international search report (Art. 21(3))*
 - *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*
-