



- (51) **International Patent Classification:**
G06N 20/00 (2019.01) *G06F 21/55* (2013.01)
- (21) **International Application Number:**
PCT/US2024/016438
- (22) **International Filing Date:**
20 February 2024 (20.02.2024)
- (25) **Filing Language:** English
- (26) **Publication Language:** English
- (30) **Priority Data:**
63/453,405 20 March 2023 (20.03.2023) US
18/441,918 14 February 2024 (14.02.2024) US
- (71) **Applicant: HIDDENLAYER INC.** [US/US]; 14900 Avery Ranch Blvd, Suite C200, Box 201, Austin, TX 78717 (US).
- (72) **Inventors: BURNS, Tanner;** c/o 14900 Avery Ranch Blvd, Suite C200, Box 201, Austin, TX 78717 (US).

- SESTITO, Chris;** c/o 14900 Avery Ranch Blvd, Suite C200, Box 201, Austin, TX 78717 (US). **BALLARD, James;** c/o 14900 Avery Ranch Blvd, Suite C200, Box 201, Austin, TX 78717 (US).
- (74) **Agent: KUKKONEN, Carl, A., III et al.;** Jones Day, 250 Vesey Street, New York, NY 10281-1047 (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, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MU, 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.

(54) **Title:** MACHINE LEARNING MODEL ADVERSARIAL ATTACK MONITORING

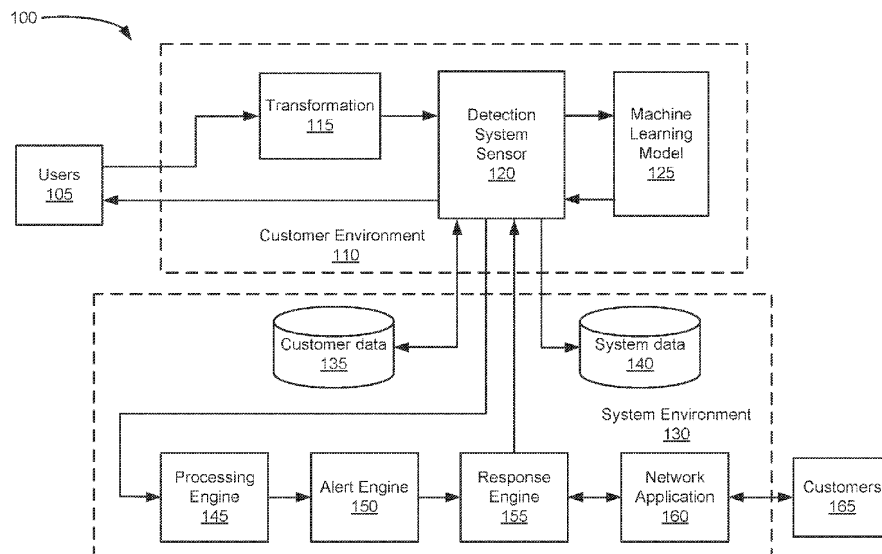


FIG. 1

(57) **Abstract:** Adversarial attacks on a machine learning model are detected by receiving vectorized data input into the machine learning model along with outputs of the machine learning model responsive to the vectorized data. The vectorized data corresponds to a plurality of queries of the machine learning model by a requesting user. A confidence level is determined which characterizes a likelihood of the vectorized data being part of a malicious act directed to the machine learning model by the requesting user. Data providing the determined confidence levels can be provided to a consuming application or process. Multi-tenant architectures are also provided in which multiple machine learning models associated with different customers can be centrally monitored.

[Continued on next page]

(84) Designated States (*unless otherwise indicated, for every kind of regional protection available*): ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, 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, ME, 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))