

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property
Organization
International Bureau

(43) International Publication Date
29 August 2024 (29.08.2024)



(10) International Publication Number
WO 2024/177954 A1

(51) International Patent Classification:

G06F 21/57 (2013.01) G06N 20/00 (2019.01)
G06F 21/56 (2013.01) H04L 9/40 (2022.01)
G06F 21/53 (2013.01)

(21) International Application Number:

PCT/US2024/016430

(22) International Filing Date:

20 February 2024 (20.02.2024)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

18/113,444 23 February 2023 (23.02.2023) US

(71) Applicant: **HIDDENLAYER INC.** [US/US]; 14900 Avery Ranch Blvd, Suite C200, Box 201, Austin, TX 78717 (US).

(72) Inventors: **BURNS, Tanner**; 14900 Avery Ranch Blvd, Suite C200, Box 201, Austin, TX 78717 (US). **SESTITO,**

Chris; 14900 Avery Ranch Blvd, Suite C200, Box 201, Austin, TX 78717 (US). **BALLARD, James**; 14900 Avery Ranch Blvd, Suite C200, Box 201, Austin, TX 78717 (US). **BONNER, Thomas**; 14900 Avery Ranch Blvd, Suite C200, Box 201, Austin, TX 78717 (US). **JANUS, Marta**; 14900 Avery Ranch Blvd, Suite C200, Box 201, Austin, TX 78717 (US). **WICKENS, Eoin**; 14900 Avery Ranch Blvd, Suite C200, Box 201, Austin, TX 78717 (US).

(74) Agent: **KUKKONEN, Carl, A.** 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,

(54) Title: SCANNING AND DETECTING THREATS IN MACHINE LEARNING MODELS

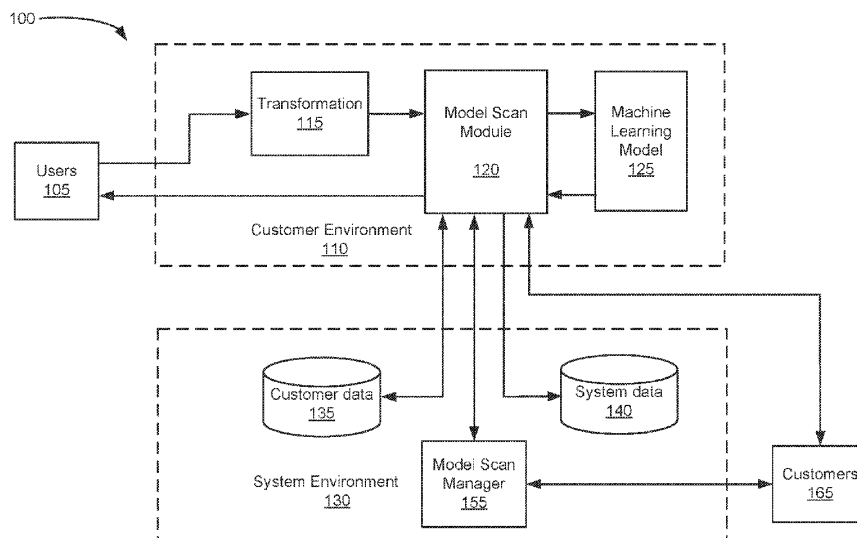


FIGURE 1

(57) Abstract: A machine learning model is scanned to detect actual or potential threats. The threats can be detected before execution of the machine learning model or during an isolated execution environment. The threat detection may include performing a machine learning file format check, vulnerability check, tamper check, and stenography check. The machine learning model may also be monitored in an isolated environment during an execution or runtime session. After performing a scan, the system can generate a signature based on actual, potential, or absence of detected threats.

[Continued on next page]

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, 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))*
- *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))*