(12) UK Patent Application (19) GB (11) 2617735 (13) A

(43) Date of Reproduction by UK Office 18.10.2023

(21) Application No: 2310212.2

(22) Date of Filing: 22.11.2021

Date Lodged: 04.07.2023

(30) Priority Data:
(31) 17114819 (32) 08.12.2020 (33) US

(86) International Application Data:
PCT/IB2021/060808 En 22.11.2021

(87) International Publication Data:
WO2022/123372 En 16.06.2022

(71) Applicant(s):
International Business Machines Corporation
New Orchard Road, Armonk, New York 10504,
United States of America

(72) Inventor(s):
Taesung Lee
Ian Michael Molloy

(74) Agent and/or Address for Service:
Elkington and Fife LLP
Prospect House, 8 Pembroke Road, Sevenoaks, Kent,
TN13 1XR, United Kingdom

(51) INT CL:
G06N 3/08 (2023.01)

(56) Documents Cited:
CN 111667049 A **CN 111295674 A**
US 7409372 B2 **US 20190095629 A1**

(58) Field of Search:
INT CL **G06F, G06N**
Other: **CNPAT, WPI, EPODOC, CNKI**

(54) Title of the Invention: **Dynamic gradient deception against adversarial examples in machine learning models**
Abstract Title: **Dynamic gradient deception against adversarial examples in machine learning models**

(57) Mechanisms are provided for obfuscating a trained configuration of a trained machine learning model. A trained machine learning model processes input data to generate an initial output vector having classification values for each of the plurality of predefined classes. A perturbation insertion engine determines a subset of classification values in the initial output vector into which to insert perturbations. A perturbation insertion engine modifies classification values in the subset of classification values by inserting a perturbation in a function associated with generating the output vector for the classification values in the subset of classification values, to thereby generate a modified output vector. The trained machine learning model outputs the modified output vector. The perturbation modifies the subset of classification values to obfuscate the trained configuration of the trained machine learning model while maintaining accuracy of classification of the input data.

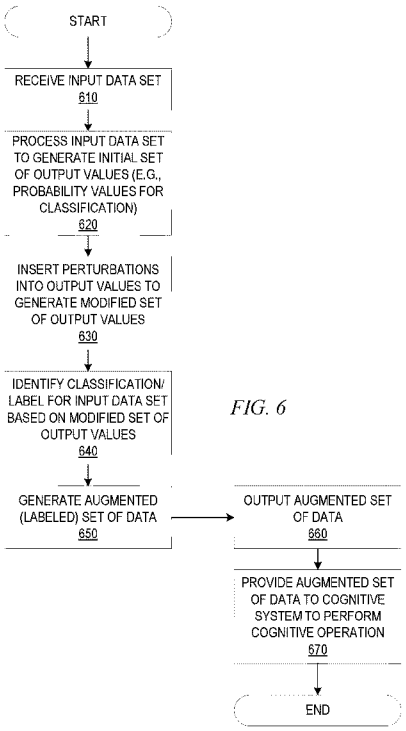

FIG. 6

GB 2617735 A