(19) Deutsches Patent- und Markenamt

(10) **DE 10 2018 220 711 A1** 2020.06.04

(12) # Offenlegungsschrift

(21) Aktenzeichen: **10 2018 220 711.9**
(22) Anmeldetag: **30.11.2018**
(43) Offenlegungstag: **04.06.2020**

(51) Int Cl.: ***G06N 3/02*** (2006.01)
G06K 9/62 (2006.01)
G06N 20/00 (2019.01)
G07C 9/00 (2020.01)

(71) Anmelder:
Robert Bosch GmbH, 70469 Stuttgart, DE

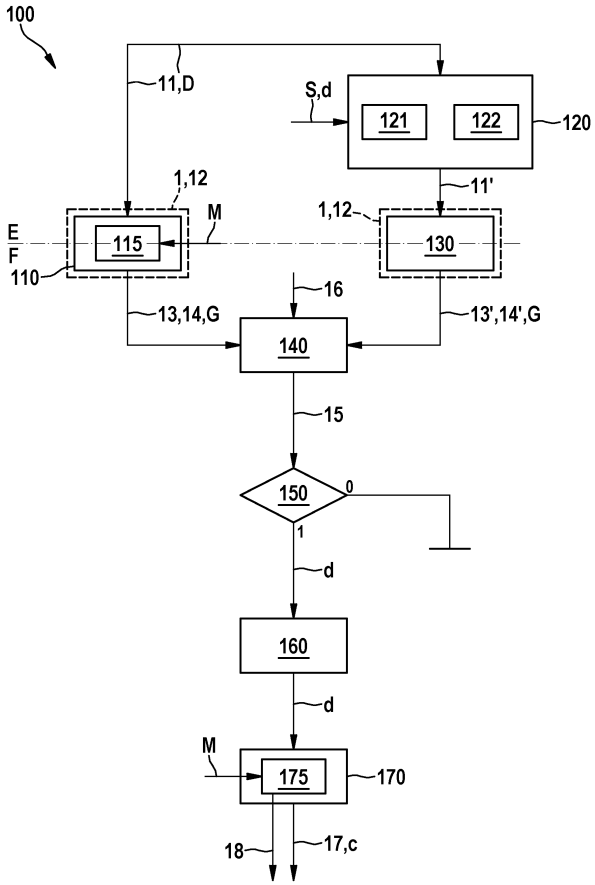
(72) Erfinder:
Fischer, Volker, 71229 Leonberg, DE; Metzen, Jan Hendrik, 71034 Böblingen, DE

Rechercheantrag gemäß § 43 PatG ist gestellt.

Die folgenden Angaben sind den vom Anmelder eingereichten Unterlagen entnommen.

(54) Bezeichnung: **Messung der Anfälligkeit von KI-Modulen gegen Täuschungsversuche**

(57) Zusammenfassung: Verfahren (100) zur Messung der Anfälligkeit eines KI-Moduls (1) gegen Täuschungsversuche mit den Schritten:
• zu einem vorgegebenen Auf-Datensatz (11) im Eingaberaum E wird die Klassifikation (13) und/oder Regression (14), auf die das KI-Modul (1) den Auf-Datensatz (11) abbildet, als ungestörtes Ergebnis (13, 14) ermittelt (110);
• dem Auf-Datensatz (11) wird mindestens eine Störung S mit einer Dimensionalität d < D aufgeprägt (120), so dass mindestens ein gestörter Datensatz (11') im Eingaberaum E entsteht;
• die Klassifikation (13') und/oder Regression (14'), auf die das KI-Modul (1) den gestörten Datensatz (11') abbildet, wird als gestörtes Ergebnis (13', 14') ermittelt (130);
• die Abweichung (15) des gestörten Ergebnisses (13', 14') von dem ungestörten Ergebnis (13, 14) wird mit einer vorgegebenen Metrik (16) ermittelt (140);
• in Antwort darauf, dass die Abweichung (15) ein vorgegebenes Kriterium erfüllt (150), wird festgestellt (160), dass das KI-Modul (1) in Bezug auf den Auf-Datensatz (11) für Täuschungsversuche mit Dimensionalität d anfällig ist.
Verfahren (200) zur Herstellung eines KI-Moduls (1), prüft Anfälligkeit des KI-Moduls (1) mit dem Verfahren (100).
Verfahren (300) zur Überwachung eines Erfassungsbereichs (2), prüft Anfälligkeit des verwendeten KI-Moduls mit dem Verfahren (100).
Computerprogramm.



Beschreibung

[0001] Die vorliegende Erfindung betrifft die Sicherung von Systemen, die Messdaten mit KI-Modulen auswerten, gegen Täuschungsversuche insbesondere mit sogenannten „adversarial examples“.

Stand der Technik

[0002] In vielen Bereichen werden KI-Module mit einer lernfähigen künstlichen Intelligenz, wie beispielsweise künstlichen neuronalen Netzwerken, eingesetzt, um Messdaten auszuwerten. Dies ist besonders dann vorteilhaft, wenn die Messdaten, beispielsweise Bilddaten, sehr hochdimensional sind und aus ihnen eine vergleichsweise niederdimensionale Klassifikation und/oder Regression zu ermitteln ist. Beispielsweise können Bilddaten mit KI-Modulen dahingehend klassifiziert werden, welche Objekte und Verkehrszeichen in einer Straßenszene enthalten sind.

[0003] Derartige KI-Module sind mit sogenannten „adversarial examples“ angreifbar. Dabei handelt es sich um gezielt böswillig eingebrachte Veränderungen von Eingangsdaten des KI-Moduls mit dem Ziel, eine Klassifikation in eine unzutreffende Klasse herbeizuführen. So kann beispielsweise ein Verkehrszeichen durch Aufkleber oder Beschriftungen, die einem menschlichen Fahrer kaum auffallen, so verändert werden, dass es von dem Netzwerk als ein anderes Verkehrszeichen erkannt wird. Dementsprechend reagiert das Fahrzeug falsch auf das Verkehrszeichen.

[0004] Aus der DE 10 2018 200 724 A1 ist ein Verfahren zur Generierung von Datensignalstörungen bekannt, mit dem typische Angriffe dieser Art künstlich nachgebildet werden können. Diese Nachbildungen können verwendet werden, um die Anfälligkeit des Klassifikationsnetzwerks zu studieren und Gegenmaßnahmen zu erproben

Offenbarung der Erfindung

[0005] Im Rahmen der Erfindung wurde ein Verfahren zur Messung der Anfälligkeit eines KI-Moduls gegen Täuschungsversuche entwickelt. Das KI-Modul umfasst eine trainierbare künstliche Intelligenz in Form einer parametrisierten internen Verarbeitungskette. Die interne Verarbeitungskette kann insbesondere beispielsweise ein künstliches neuronales Netzwerk, KNN, umfassen. Die Parameter der internen Verarbeitungskette können dann beispielsweise Gewichte sein, mit denen Eingaben, die an Neuronen angelegt werden, zu Aktivierungen dieser Neuronen verrechnet werden.

[0006] Das mit dem Verfahren zu untersuchende KI-Modul ist dazu ausgebildet, mittels seiner internen Verarbeitungskette Eingabe-Datensätze aus einem Eingaberaum **E** mit Dimensionalität **D** auf eine Klassifikation und/oder Regression in einem Ausgaberaum **F** mit Dimensionalität G < D abzubilden. Die Klassifikation gibt beispielsweise zu einem Eingabe-Datensatz und einem vorgegebenen Kanon an Klassen jeweils an, mit welchen Konfidenzen der Eingabe-Datensatz jeder dieser Klassen angehört. Die Regression gibt beispielsweise zu einem Eingabe-Datensatz und einer interessierenden reellwertigen Größe an, welche Werte dieser Größe mit welchen Konfidenzen im Lichte des Eingabe-Datensatzes stimmig sind.

[0007] In den meisten Anwendungen gilt G << D. So lebt beispielsweise ein 512x512 Pixel großes Bild in einem Eingaberaum **E** mit einer Dimensionalität D=262.144. Die Anzahl der verschiedenen Objekte, auf deren Vorhandensein das Bild untersucht werden soll, bestimmt die Dimensionalität G des Ausgaberaums **F** und ist typischerweise kleiner als 1000.

[0008] Bei dem Verfahren wird zu einem vorgegebenen Auf-Datensatz im Eingaberaum **E** die Klassifikation und/oder Regression, auf die das KI-Modul den Auf-Datensatz abbildet, als ungestörtes Ergebnis ermittelt.

[0009] Dem Auf-Datensatz wird mindestens eine Störung **S** aufgeprägt, die eine Dimensionalität **d < D** aufweist. Auf diese Weise entsteht mindestens ein gestörter Datensatz im Eingaberaum **E**. Zu diesem Zweck kann insbesondere vorteilhaft ein „Adversarial Example“ für das KI-Modul als Störung **S** ermittelt werden. Insbesondere für KI-Module, die KNNs umfassen, sind Algorithmen bekannt, mit denen sich „Adversarial Examples“ mit frei wählbarer Dimensionalität **d** erzeugen lassen.

[0010] Die Klassifikation und/oder Regression, auf die das KI-Modul den gestörten Datensatz abbildet, wird als gestörtes Ergebnis ermittelt. Die Abweichung des gestörten Ergebnisses von dem ungestörten Ergebnis wird mit einer vorgegebenen Metrik ermittelt. Diese Metrik kann insbesondere auf die konkrete Anwendung zugeschnitten sein, in der das vom KI-Modul gelieferte Ergebnis, also die Klassifikation und/oder Regression, verwendet werden soll. Die Metrik kann dann beispielsweise dadurch motiviert sein, wie störend die Abweichung sich in der Anwendung auswirkt.

[0011] Wird das KI-Modul beispielsweise verwendet, um im Rahmen eines Fahrassistenzsystems oder eines Systems für das ganz oder teilweise automatisierte Fahren Verkehrszeichen in einer Szenerie zu erkennen, so führt beispielsweise die Verwechslung eines Tempo-50-Schildes mit einem Tempo-60-Schild nur zu einer geringen Störung des Verkehrs. Wird hingegen ein Tempo-50-Schild irrtümlich für ein Stopp-Schild gehalten, kann ein unvermitteltes Abbremsen des Fahrzeugs die Folge sein

und einen Auffahrunfall provozieren. Wird umgekehrt das Stopp-Schild für ein Tempo-50-Schild gehalten, sind die potentiellen Folgen noch schwerer, weil das Fahrzeug möglicherweise seitlich und im Wesentlichen ungebremst vom Vorfahrtsberechtigten gerammt wird. Die Metrik kann in dieser Anwendung also beispielsweise die Abweichung nach der Schwere der möglichen Auswirkungen im Verkehr bewerten.

[0012] Wird das KI-Modul beispielsweise verwendet, um in einem Zutrittskontrollsystem auf Grund von Messdaten zu entscheiden, ob ein gültiges Zugangsmedium vorgelegt wurde, so wirkt es sich letztendlich nicht aus, wenn ein ungültiges Zugangsmedium mit einem anderen verwechselt wird. Der Zutritt wird nach wie vor nicht gewährt. Werden zwei gültige Zugangsmedien miteinander verwechselt, kann sich dies beispielsweise dahingehend auswirken, dass der Zutritt mit einer falschen Berechtigungsstufe gewährt wird oder dass in Protokolldateien über den Zutritt die falsche Person vermerkt wird. Wird ein gültiges Zugangsmedium irrtümlich für ein ungültiges gehalten, erhält eine berechtigte Person keinen Zutritt, was den Betriebsablauf stören kann. Am schlimmsten wirkt es sich aus, wenn ein ungültiges Zugangsmedium für ein gültiges gehalten wird und einer völlig unberechtigte Person der Zugang gewährt wird.

[0013] Die Auswirkung einer Störung **S** auf das Ergebnis wird durch das Zusammenspiel aus der Anfälligkeit des KI-Moduls für die Störung **S** und der Stärke dieser Störung **S** bestimmt. Um gezielt die Anfälligkeit des KI-Moduls zu messen, kann die Metrik beispielsweise die Auswirkung pro Einheit Stärke der Störung **S** angeben oder auf eine Einheitsstärke **S** der Störung normiert sein.

[0014] In Antwort darauf, dass die Abweichung ein vorgegebenes Kriterium erfüllt, wird festgestellt, dass das KI-Modul in Bezug auf den Auf-Datensatz für Täuschungsversuche mit Dimensionalität **d** anfällig ist. Das Kriterium kann beispielsweise bei einer skalaren Metrik das Überschreiten oder Unterschreiten eines vorgegebenen Schwellwerts umfassen.

[0015] Es wurde erkannt, dass diejenigen Störungen **S**, die sich im Kontext der jeweiligen Anwendung nennenswert auf das Ergebnis der Klassifikation und/oder Regression auswirken, in einem Unterraum C $\subset$ E mit Dimensionalität c < **D** leben. Wenn nun eine Störung **S** mit einer kleineren Dimensionalität als c auf einen Eingabe-Datensatz aufgeprägt wird, dann ist die Wahrscheinlichkeit gering, dass diese Störung **S** in **C** liegt. Ist beispielsweise c=20 und wird nur eine eindimensionale Störung aufgeprägt, so ist es unwahrscheinlich, dass die der Störung entsprechende eindimensionale Gerade den irgendwo im Eingaberaum **E** liegenden 20-dimensionalen Unterraum **C** schneidet. Hat die Störung **S** hingegen eine Dimensionalität von c oder größer, steigt die Wahrscheinlichkeit, dass sie in **C** liegt, deutlich an. Dies gilt im Besonderen dann, wenn gezielt „Adversarial Examples“ als Störungen **S** gewählt werden.

[0016] Die Störung **S** muss also eine bestimmte Mindest-Dimensionalität haben, um sich merklich auf das Ergebnis auszuwirken. Es wurde erkannt, dass diese Mindest-Dimensionalität ein zuverlässiges Maß dafür ist, wie robust das KI-Modul gegen Täuschungsversuche ist. Dies sei am Beispiel von Bildern als Eingabe-Datensätzen illustriert. Wie zuvor erläutert, entspricht die Dimensionalität **D** des Eingaberaums **E** der Anzahl der Bildpixel. Die Mindest-Dimensionalität, ab der sich eine Störung **S** merklich auswirkt, entspricht in diesem Beispiel der Anzahl der Pixel, die durch die Störung **S** beeinflusst werden. Für das maximal anfällige Negativ-Beispiel eines KI-Moduls ist dann der Unterraum C $\subset$ **E** eindimensional. Das bedeutet, dass eine Veränderung des Bildes in lediglich einem Pixel bereits ausreicht, um die Klassifikation und/oder Regression signifikant zu ändern. Hat der Unterraum C $\subset$ **E** hingegen fast die Dimensionalität **D**, so müssen beinahe alle Pixel des Bildes verändert werden, um das KI-Modul zu täuschen.

[0017] Je größer wiederum die Mindest-Dimensionalität einer Störung **S** sein muss, desto schwieriger ist es, diese Störung unbemerkt auf Eingabe-Datensätze aufzuprägen. In dem eingangs genannten Beispiel des Stopp-Schildes etwa kann bei einem anfälligen KI-Modul bereits ein unscheinbarer Aufkleber auf dem Schild ausreichen, damit das Schild falsch klassifiziert wird. Um hingegen ein robusteres KI-Modul zu täuschen, müsste das Schild so stark verändert werden, dass der Manipulationsversuch optisch offensichtlich ist und sich jemand der Beseitigung dieser Veränderung annimmt. Beispielsweise kann das von jedem auch noch so robusten KI-Modul gelieferte Klassifikationsergebnis mit absoluter Sicherheit verändert werden, indem das Stopp-Schild schlichtweg abmontiert und durch ein anderes Schild ersetzt wird.

[0018] Das bedeutet, dass die Mindest-Dimensionalität einer Störung **S**, die sich nach Maßgabe der vorgegebenen Metrik auf die jeweilige Anwendung signifikant auswirkt, ein besonders zutreffendes Maß für die Anfälligkeit des KI-Moduls gegen solche Störungen ist, die sich unbemerkt in die Eingabe-Datensätze einschleusen lassen.

[0019] Daher werden in einer besonders vorteilhaften Ausgestaltung dem gleichen Auf-Datensatz mehrere Störungen **S** mit unterschiedlichen Dimensionalitäten **d** aufgeprägt. Beispielsweise kann hierbei mit einer niedrigen Dimensionalität **d** der Störungen **S**, etwa d=1, angefangen und die Dimensionalität **d** der Störungen **S** dann schrittweise gesteigert werden, etwa jeweils um **1**. Die kleinste Dimensionalität c=min (d), für die nach Maßgabe der besagten Metrik eine Anfälligkeit des KI-Moduls gegen Täuschungsversu-

che festgestellt wird, wird als Maß für die Anfälligkeit des KI-Moduls in Bezug auf den Auf-Datensatz gewertet. Hierfür kann beispielsweise die Größe s=D-c als „intrinsische Dimension“ der sich stark auswirkenden Störungen gewählt werden. Je kleiner c ist, d.h., je größer die Anfälligkeit ist, desto größer wird s.

[0020] Es wurde weiterhin erkannt, dass die durch c bzw. s gemessene Anfälligkeit im Eingaberaum **E**, in dem die Eingabe-Datensätze leben, nicht homogen verteilt ist. Insbesondere kann es solche Eingabe-Datensätze geben, für die die Anfälligkeit des KI-Moduls gegen Täuschungsversuche besonders groß ist (sogenannte „corner cases“). Die Situation ist in gewisser Weise vergleichbar mit einer Wanderroute im Gebirge, die größtenteils auch für Ungeübte begehbar ist, jedoch auch einige exponierte Stellen aufweist, an denen ein Fehltritt unmittelbar zum Absturz führt. Das Verfahren ermöglicht es, solche „corner cases“ zu identifizieren und zu quantifizieren. Dadurch können wiederum Gegenmaßnahmen ergriffen werden, um die Anfälligkeit zu vermindern. Beispielsweise kann das KI-Modul in Bereichen um die „corner cases“ herum mit zusätzlichen Lern-Daten trainiert werden.

[0021] In einer weiteren vorteilhaften Ausgestaltung wird die Anfälligkeit des KI-Moduls, etwa gemessen in Form von c bzw. s, in Bezug auf mehrere Datensätze aus einer vorgegebenen Menge M ermittelt. Über die auf die so ermittelten Anfälligkeiten wird eine zusammenfassende Statistik ermittelt. Diese zusammenfassende Statistik kann beispielsweise einen Mittelwert, und/oder eine Varianz, und/oder eine Häufigkeitsverteilung, und/oder einen schlechtesten Wert der ermittelten Anfälligkeiten beinhalten.

[0022] So kann es beispielsweise für viele Anwendungen ausreichend sein, dass die Anfälligkeit im Mittel in einem vorgegebenen Rahmen bleibt, während in anderen Anwendungen, in denen bei einer Fehlfunktion Sach- oder Personenschäden drohen, eine „worst-case“ Abschätzung der Anfälligkeit benötigt wird.

[0023] Die neu gewonnene Quantifizierbarkeit der Anfälligkeiten von KI-Modulen gegen Täuschungsversuche kann insbesondere als Feedback genutzt werden, um KI-Module gegen Täuschungsversuche zu härten.

[0024] Daher bezieht sich die Erfindung auch auf ein Verfahren zur Herstellung eines KI-Moduls, welches eine trainierbare künstliche Intelligenz in Form einer parametrisierten internen Verarbeitungskette umfasst und dazu ausgebildet ist, mittels dieser internen Verarbeitungskette Eingabe-Datensätze aus einem Eingaberaum **E** mit Dimensionalität **D** auf eine Klassifikation und/oder Regression in einem Ausgaberaum **F** mit Dimensionalität G < D abzubilden.

[0025] Bei diesem Verfahren wird die Architektur der internen Verarbeitungskette durch sogenannte Hyperparameter festgelegt. Diese Hyperparameter beziehen sich auf Freiheitsgrade, die bei der Gestaltung der Architektur bestehen. Beispielsweise umfasst die Architektur eines KNN sowohl die Abfolge von Schichten bestimmter Typen (etwa „10x alternierend eine Faltungsschicht und eine Max-Pooling-Schicht, dann eine vollvernetzte Schicht) als auch die konkreten Abmessungen dieser Schichten (etwa „512×512 Neuronen in der Eingangsschicht“).

[0026] Es wird KI-Modul mit der internen Verarbeitungskette gebildet, die eine Architektur gemäß den festgelegten Hyperparametern aufweist. Diese Architektur weist dann noch freie Parameter für das Training auf, im Falle eines KNN etwa die Gewichte, mit denen die einem jeden Neuron zugeführten Eingaben zu einer Aktivierung dieses Neurons verrechnet werden. Beim Training des KI-Moduls werden diese freien Parameter anhand einer Menge **L** von Lern-Datensätzen und zugehörigen Lern-Ergebnissen so optimiert, dass das KI-Modul die Lern-Datensätze nach Maßgabe einer Fehlerfunktion mit einer vorgegebenen Genauigkeit auf die Lern-Ergebnisse abbildet. Beispielsweise können die Lern-Datensätze Bilder von Straßenszenen umfassen, und das zu jedem Bild jeweils gehörende Lern-Ergebnis gibt an, welche Objekte das KI-Modul in dem Bild erkennen sollte.

[0027] Nach Abschluss des Trainings wird mit einer Menge M von Validierungs-Datensätzen gemäß dem zuvor beschriebenen Verfahren eine zusammenfassende Statistik der Anfälligkeiten des trainierten KI-Moduls ermittelt. Dabei sollte die Menge M der Validierungs-Datensätze vorteilhaft disjunkt zur Menge **L** der Lern-Datensätze sein, denn die Lern-Datensätze sind insoweit ausgezeichnet, als das KI-Modul gerade auf die korrekte Verarbeitung dieser Datensätze besonders optimiert ist. Relevant für die Anwendung ist jedoch gerade korrekte Funktion des KI-Moduls in unbekannten Situationen.

[0028] Es werden nun die Hyperparameter, d.h. die Architektur des KI-Moduls, dahingehend optimiert, dass nach erneutem Bilden und Trainieren des KI-Moduls die hierfür ermittelte zusammenfassende Statistik der Anfälligkeiten eine insgesamt geringere Anfälligkeit gegen Täuschungsversuche anzeigt.

[0029] Es wurde erkannt, dass die Anfälligkeit von KI-Modulen gegen Täuschungsversuche nicht nur durch ihr jeweiliges Training bestimmt wird, sondern auch durch die zu Grunde liegende Architektur. Für das Training gibt es Feedback in Form der Fehlerfunktion, die beurteilt, wie gut die vorgegebenen Lern-Ergebnisse reproduziert werden. Durch die zusammenfassende Statistik der Abhängigkeiten steht nun auch Feedback in Bezug auf die Architektur des KI-Moduls zur Verfügung.

[0030] In einer besonders vorteilhaften Ausgestaltung werden die Hyperparameter optimiert, indem Architekturen der internen Verarbeitungskette mit einem evolutionären Algorithmus erzeugt werden. Die jeweils nach dem Bilden und Trainieren des KI-Moduls mit einer Architektur ermittelte zusammenfassende Statistik der Anfälligkeiten geht in ein Gütemaß für die Bewertung dieser Architektur ein.

[0031] Im Rahmen des evolutionären Algorithmus können Architekturen der internen Verarbeitungskette beispielsweise mit den naturanalogen Verfahren der Kreuzung und Mutation so abgewandelt werden, dass immer diejenigen Architekturen mit dem besseren Gütemaß „überleben" (Survival of the Fittest). Ein evolutionärer Algorithmus ist für die Optimierung der Architekturen besonders geeignet, da er nicht voraussetzt, dass die zu optimierenden Hyperparameter kontinuierlich sind. So ist beispielsweise der Typ einer Schicht in einem KNN (etwa Faltungsschicht, Pooling-Schicht oder vollvernetzte Schicht) eine diskrete Größe, die für ein bei Optimierungsaufgaben häufig verwendetes Gradientenabstiegsverfahren schlecht fassbar ist.

[0032] In einer weiteren besonders vorteilhaften Ausgestaltung der zuvor beschriebenen Verfahren enthält mindestens ein Auf-Datensatz, und/oder mindestens ein Lern-Datensatz, mindestens einen Messwert einer physikalischen Messgröße. Auf diese Weise kann ein System, das diese Messgrößen auswertet, dagegen geschützt werden, dass durch eine gezielte geringfügige Manipulation der Messgröße die Auswertung unbemerkt in eine falsche Richtung gelenkt wird. Der Messwert kann insbesondere mit einem Sensor ermittelt werden, der eine physikalische Einwirkung registriert, deren Art und/oder Stärke durch die physikalische Messgröße charakterisiert ist.

[0033] Die Anfälligkeit eines KI-Moduls für Täuschungsversuche zu ermitteln ist in der Regel kein Selbstzweck, sondern auf eine konkrete Anwendung bezogen. Es wurde erkannt, dass die besagte Anfälligkeit im Rahmen einer solchen Anwendung unmittelbar als Maß für die Verlässlichkeit von Ergebnissen genutzt werden kann, die mit dem KI-Modul ermittelt werden. Je nach Anwendung kann eine belastbare Aussage über die Verlässlichkeit genauso wichtig sein wie das Ergebnis selbst.

[0034] Daher bezieht sich die Erfindung auch auf ein Verfahren zur Überwachung eines Erfassungsbereichs. Bei diesem Verfahren wird durch physikalische Beobachtung des Erfassungsbereichs mit mindestens einem Sensor mindestens ein Mess-Datensatz mit Messdaten erfasst.

[0035] Der Mess-Datensatz wird einem KI-Modul zugeführt, welches eine trainierbare künstliche Intelligenz in Form einer parametrisierten internen Verarbeitungskette umfasst und dazu ausgebildet ist, mittels dieser internen Verarbeitungskette Eingabe-Datensätze aus einem Eingaberaum **E** mit Dimensionalität **D** auf eine Klassifikation und/oder Regression in einem Ausgaberaum **F** mit Dimensionalität $G < D$ abzubilden. Der Eingaberaum **E** ist hier der Raum, in dem sich die möglichen Messdaten bewegen können. Der Ausgaberaum **F** und seine Dimensionalität G sind durch die mit der Überwachung verfolgte Fragestellung, etwa Art und Anzahl der in den Messdaten zu klassifizierenden Objekte.

[0036] Die vom KI-Modul ausgegebene Klassifikation und/oder Regression wird als Ergebnis der Überwachung gewertet und/oder ausgegeben. Zusätzlich wird die Anfälligkeit des KI-Moduls gegen Täuschungsversuche mit dem beschriebenen Verfahren gemessen und als Maß für die Verlässlichkeit des Ergebnisses gewertet und/oder ausgegeben.

[0037] Die auf automatisiertem Wege ermittelte Anfälligkeit des KI-Moduls für Täuschungsversuche eröffnet wiederum die Möglichkeit, automatisiert Gegenmaßnahmen gegen derartige Täuschungsversuche zu ergreifen, sofern die ermittelte Anfälligkeit ein vorgegebenes Kriterium erfüllt. Das Kriterium kann beispielsweise bei einer skalaren Anfälligkeit beinhalten, dass ein vorgegebener Schwellwert überschritten oder unterschritten wird.

[0038] Als eine beispielhafte mögliche Gegenmaßnahme kann der Mess-Datensatz einem weiteren KI-Modul zugeführt werden, wobei dieses weitere KI-Modul eine andere Architektur aufweist als das zuvor genutzte KI-Modul, und/oder wobei das weitere KI-Modul anders trainiert worden ist als das zuvor genutzte KI-Modul. Je universeller ein und derselbe Täuschungsversuch (etwa ein „Adversarial Example") auf unterschiedlich trainierte, oder sogar auf unterschiedlich strukturierte, KI-Module passt, desto schwieriger wird es, die Manipulation so den Messdaten zu überlagern, dass sie nicht anderweitig bemerkt wird.

[0039] Als eine weitere beispielhafte mögliche Gegenmaßnahme können mit einem weiteren physikalischen Sensor zusätzliche Messdaten erfasst werden. Diese zusätzlichen Messdaten können herangezogen werden, um das Ergebnis zu plausibilisieren. Beispielsweise kann zu diesem Zweck ein weiterer physikalischer Sensor aktiviert werden. Der Täuschungsversuch (etwa das „Adversarial Example") lebt im Eingaberaum **E** des KI-Moduls, aber die zusätzlichen Messdaten leben außerhalb dieses Raums **E** und werden somit von dem Täuschungsversuch nicht abgedeckt.

[0040] Als eine weitere beispielhafte mögliche Gegenmaßnahme kann das Ergebnis verworfen wer-

den. Je nach Anwendung kann dies das kleinere Übel sein im Vergleich dazu, dass ein möglicherweise manipuliertes Ergebnis weiter verarbeitet wird.

[0041] Nach dem zuvor Beschriebenen umfasst in einer besonders vorteilhaften Ausgestaltung der Erfassungsbereich mindestens einen Teil des Umfelds eines Fahrzeugs. Das vom KI-Modul gelieferte Ergebnis wird einem in dem Fahrzeug verbauten Fahrassistenzsystem oder System für das zumindest teilweise automatisierte Fahren zugeführt. Das Fahrassistenzsystem oder System für das zumindest teilweise automatisierte Fahren ist dazu ausgebildet, abhängig von dem Ergebnis ein Lenksystem, ein Antriebssystem, und/oder ein Bremssystem, des Fahrzeugs anzusteuern. Gerade diese Anwendung ist zum einen besonders sicherheitskritisch, und zum anderen sind Täuschungsversuche einfach zu realisieren, da etwa Verkehrsschilder öffentlich zugänglich sind.

[0042] Bei der Anwendung im Fahrzeug sind verschiedene spezifische Gegenmaßnahmen einzeln oder in Kombination sinnvoll. So kann beispielsweise eine für einen Fahrer des Fahrzeugs wahrnehmbare physikalische Warneinrichtung aktiviert werden. Der Fahrer des Fahrzeugs kann dazu aufgefordert werden, das vom KI-Modul gelieferte Ergebnis zu bestätigen oder richtigzustellen, und/oder die Kontrolle über das Fahrzeug zu übernehmen. In der höchsten Eskalationsstufe kann das Fahrzeug auf einer für den Ausfall der zumindest teilweise automatisierten Fahrfunktion vorgesehenen Notfalltrajektorie zum Stehen gebracht werden.

[0043] Wie zuvor bereits erwähnt, sind Zutrittskontrollsysteme eine weitere wichtige Anwendung. Daher umfasst in einer weiteren besonders vorteilhaften Ausgestaltung der Erfassungsbereich mindestens einen Teil eines Bereiches, in dem ein Zutrittskontrollsystem zur Steuerung des Zutritts zu einem Raum, einem Gelände und/oder einem informationstechnischen System die Vorlage eines Zugangsmediums erwartet. Das vom KI-Modul gelieferte Ergebnis wird dem Zutrittskontrollsystem zugeführt. Das Zutrittskontrollsystem ist dazu ausgebildet, auf der Grundlage dieses Ergebnisses zu entscheiden, ob im Erfassungsbereich ein gültiges Zugangsmedium vorhanden ist. Auf Grund dieser Entscheidung kann insbesondere eine Sperr- und/oder Alarmvorrichtung angesteuert werden, um den Zutritt zu gewähren oder zu verwehren. Versuche, durch eine gezielte Manipulation der Messdaten ein gültiges Zugangsmedium vorzutäuschen, können mit den beschriebenen Gegenmaßnahmen erschwert werden.

[0044] Darüber hinaus sind speziell für die Zutrittskontrolle weitere spezifische Gegenmaßnahmen einzeln oder in Kombination sinnvoll. So kann beispielsweise von der den Zutritt begehrenden Person eine zusätzliche Authentifikation gefordert werden, etwa eine PIN oder ein Passwort. Der Zutritt kann auch beispielsweise unabhängig vom Vorliegen eines gültigen Zugangsmediums für eine vorbestimmte Zeit gesperrt werden, um wiederholte Täuschungsversuche auszubremsen. Es kann auch beispielsweise ein Alarm an eine für die Sicherheit des Raums, des Geländes, bzw. des informationstechnischen Systems, verantwortliche Stelle ausgegeben werden.

[0045] Dabei werden in einer besonders vorteilhaften Ausgestaltung biometrische Messdaten einer den Zutritt begehrenden Person als Messdaten gewählt. Ein derartiges Zugangsmedium ist üblicherweise schwer oder gar nicht zu kopieren, so dass es für einen Angreifer eine ernstzunehmende Alternative darstellt, stattdessen das KI-Modul zu täuschen.

[0046] Nach dem zuvor Beschriebenen können die genannten Verfahren Gebrauch von zusätzlicher Hardware machen, etwa von zusätzlichen Sensoren. Dies ist jedoch nicht zwingend erforderlich. Die Verfahren können auch ganz oder teilweise in einer Software implementiert sein, die den unmittelbaren Kundennutzen bewirkt, dass die Anfälligkeit des KI-Moduls für Täuschungsversuche erkannt, vermindert bzw. in ihren Folgen für die jeweilige Anwendung abgeschwächt werden kann. Daher bezieht sich die Erfindung auch auf ein Computerprogramm, enthaltend maschinenlesbare Anweisungen, die, wenn sie auf einem Computer, und/oder auf einem Steuergerät, und/oder auf einem Embedded-System, ausgeführt werden, den Computer, und/oder das Steuergerät, und/oder das Embedded-System, dazu veranlassen, eines der beschriebenen Verfahren auszuführen. Ebenso bezieht sich die Erfindung auch auf einen maschinenlesbaren Datenträger oder ein Downloadprodukt mit dem Computerprogramm.

[0047] Weitere, die Erfindung verbessernde Maßnahmen werden nachstehend gemeinsam mit der Beschreibung der bevorzugten Ausführungsbeispiele der Erfindung anhand von Figuren näher dargestellt.

Ausführungsbeispiele

[0048] Es zeigt:

Fig. 1 Ausführungsbeispiel des Verfahrens **100** zur Messung der Anfälligkeit;

Fig. 2 Vergleich der Dimensionalitäten $\mathbf{c_{1a}}$, $\mathbf{c_{1b}}$, $\mathbf{c_{1c}}$ von Störungen **S**, ab denen bei KI-Modulen **1a**, **1b**, **1c** signifikante Abweichungen **15** verursacht werden;

Fig. 3 Ausführungsbeispiel des Verfahrens **200** zur Herstellung eines KI-Moduls **1**;

Fig. 4 Ausführungsbeispiel des Verfahrens **300** zur Überwachung eines Erfassungsbereichs **2**;

Fig. 5 Schematische Darstellung einer beispielhaften Anwendung des Verfahrens **300** in einem Fahrzeug **50**;

Fig. 6 Schematische Darstellung einer beispielhaften Anwendung des Verfahrens **300** in einem Zutrittskontrollsystem **60**.

[0049] **Fig. 1** zeigt ein Ausführungsbeispiel des Verfahrens **100**. Ein vorgegebener Auf-Datensatz **11**, der im Eingaberaum **E** mit Dimensionalität **E** lebt, wird in Schritt **110** durch das KI-Modul **1** und dessen interne Verarbeitungskette **12** auf eine Klassifikation **13** und/oder Regression **14** in einem Ausgaberaum **F** mit Dimension G abgebildet.

[0050] Zusätzlich wird dem gleichen Auf-Datensatz **11** in Schritt **120** eine Störung **S** mit Dimensionalität d < D aufgeprägt. Dies kann gemäß Block **121** ein „Adversarial Example“ für das verwendete KI-Modul **1** sein. Gemäß Block **122** können weiterhin verschiedene Störungen **S** mit unterschiedlichen Dimensionalitäten **d** verwendet werden. Der gestörte Datensatz **11'** wird in Schritt **130** auf eine gestörte Klassifikation **13'**, bzw. auf eine gestörte Regression **14'**, abgebildet.

[0051] In Schritt **140** wird das ungestörte Ergebnis **13**, **14** anhand einer Metrik **16** mit dem gestörten Ergebnis **13'**, **14'** verglichen. In Schritt **150** wird überprüft, ob die ermittelte Abweichung **15** ein vorgegebenes Kriterium erfüllt, d.h., ob sie beispielsweise einen vorgegebenen Schwellwert überschreitet. Ist dies der Fall (Wahrheitswert **1**), so wird in Schritt **160** festgestellt, dass das KI-Modul **1** in Bezug auf den Auf-Datensatz **11** für Täuschungsversuche mit Dimensionalität **d** anfällig ist. In Schritt **170** wird die kleinste Dimensionalität c=min(d), für die dies der Fall ist, als Maß **17** für die Anfälligkeit des KI-Moduls **1** in Bezug auf den Auf-Datensatz **11** gewertet.

[0052] Optional kann gemäß Block **115** die Anfälligkeit des KI-Moduls **1** in Bezug auf mehrere Datensätze **11** aus einer vorgegebenen Menge M ermittelt werden. Über die zugehörigen Maße **17**, c für die Anfälligkeit des KI-Moduls **1**, die in Schritt **170** jeweils ermittelt werden, kann gemäß Block **175** eine zusammenfassende Statistik **18** geführt werden.

[0053] In **Fig. 2** sind für drei beispielhafte KI-Module **1a-1c** die Abweichungen **15** aufgetragen, die sich ergeben, wenn auf ein und denselben Auf-Datensatz **11** Störungen **S** verschiedener Dimensionalitäten aufgeprägt werden und der gestörte Datensatz **11'** dann jeweils mit dem KI-Modul **1a-1c** verarbeitet wird. Mit einem in diesem Beispiel als Schwellwert implementierten Kriterium **150** wird jeweils überprüft, ob die Abweichung **15** in Bezug auf die vorliegende Anwendung signifikant ist. Die kleinste Dimensionalität **d**, bei der dies für die drei untersuchten KI-Module **1a-1c** jeweils der Fall ist, ist in **Fig. 2** mit dem Bezugszeichen $\mathbf{c_{1a}}$, $\mathbf{c_{1b}}$ bzw. $\mathbf{c_{1c}}$ bezeichnet. Je höher diese Dimensionalität $c_{1a\text{-}1c}$ ist, desto robuster ist das zugehörige KI-Modul **1a-1c** gegen Täuschungsversuche. In dem in **Fig. 2** gezeigten Beispiel schneidet das KI-Modul **1b** am schlechtesten und das KI-Modul **1c** am besten ab. Die Anfälligkeit des KI-Moduls **1a** liegt zwischen den Anfälligkeiten der KI-Module **1b** und **1c**, bezogen auf den für diese Untersuchung verwendeten Auf-Datensatz **11**.

[0054] **Fig. 3** zeigt ein Ausführungsbeispiel des Verfahrens **200** zur Herstellung eines KI-Moduls **1**. In Schritt **210** wird die Architektur der internen Verarbeitungskette **12** des KI-Moduls **1** durch Hyperparameter **12a** festgelegt. Wie zuvor beschrieben, muss es sich bei diesen Hyperparametern **12a** nicht nur um Zahlenwerte handeln. Vielmehr können die Hyperparameter **12a** beispielsweise auch Auswahlmöglichkeiten umfassen, ob etwa eine Schicht eines KNN eine Faltungsschicht, eine vollvernetzte Schicht oder eine Pooling-Schicht ist.

[0055] In Schritt **220** wird das KI-Modul **1** mit der internen Verarbeitungskette **12**, die die durch die Hyperparameter **12a** festgelegte Architektur hat, gebildet. Diese Architektur ist ein Ansatz mit noch freien Parametern **12b**, etwa den Gewichten in dem KNN. Diese Parameter **12b** werden in Schritt **230** mit Lern-Datensätzen **11a** aus einer Menge **L** und zugehörigen Lern-Ergebnissen **13a**, **14a** trainiert. Das heißt, die Parameter **12b** werden so lange variiert, bis die Lern-Datensätze **11a** nach Maßgabe einer Fehlerfunktion mit einer vorgegebenen Genauigkeit auf die Lern-Ergebnisse **13a**, **14a** abgebildet werden.

[0056] Das fertig trainierte KI-Modul **1** wird in Schritt **240** gemäß dem zuvor beschriebenen Verfahren **100** auf seine Anfälligkeit gegen Täuschungsversuche geprüft. Hierzu werden Validierungs-Datensätze aus einer Menge M als Auf-Datensätze verwendet. Über alle Validierungs-Datensätze aus der Menge M wird eine zusammenfassende Statistik **18** der jeweiligen Anfälligkeiten des KI-Moduls **1** erstellt.

[0057] Gemäß Schritt **250** werden nun die Hyperparameter **12a** dahingehend optimiert, dass nach erneutem Bilden **220** und anschließenden Trainieren **230** des KI-Moduls **1** die zusammenfassende Statistik **18** der Anfälligkeiten eine insgesamt geringere Anfälligkeit des KI-Moduls **1** gegen Täuschungsversuche anzeigt. Es werden also nacheinander Kandidaten-KI-Module mit dem Verfahren **100** auf ihre Anfälligkeit untersucht.

[0058] In dem in **Fig. 3** gezeigten Beispiel werden die Hyperparameter **12a** optimiert, indem neue Architekturen der internen Verarbeitungskette **12** des zu bildenden KI-Moduls **1** jeweils gemäß Block **251** mit einem evolutionären Algorithmus erzeugt werden. Die zusammenfassende Statistik **18** geht dann ge-

mäß Block **252** jeweils in ein Gütemaß (Fitness Function) für die jeweilige Architektur ein.

[0059] **Fig. 4** zeigt ein Ausführungsbeispiel des Verfahrens **300** zur Überwachung eines Erfassungsbereichs **2**. In Schritt **310** wird mit einem Sensor **3** mindestens ein Mess-Datensatz **11** erfasst, der Messdaten **3a** enthält. In Schritt **320** wird dieser Mess-Datensatz dem KI-Modul **1** zugeführt. Das KI-Modul **1** erzeugt mit seiner internen Verarbeitungskette **12** aus dem Mess-Datensatz **11** eine Klassifikation **13** und/oder eine Regression **14**. In Schritt **330** wird diese Klassifikation **13** und/oder Regression **14** als Ergebnis **13**, **14** der Überwachung gewertet und/oder ausgegeben.

[0060] In Schritt **340** wird mit dem Verfahren **100** die Anfälligkeit **17**, **18** des KI-Moduls **1** gegen Täuschungsversuche gemessen. Diese Anfälligkeit **17**, **18** wird in Schritt **350** als Maß für die Verlässlichkeit des Ergebnisses **13**, **14** gewertet und/oder ausgegeben.

[0061] In **Fig. 4** sind zusätzlich noch verschiedene Gegenmaßnahmen dargestellt, die einzeln oder Kombination ergriffen werden können in Antwort auf eine Feststellung in Schritt **360**, dass die ermittelte Anfälligkeit **17**, **18** ein vorgegebenes Kriterium erfüllt (Wahrheitswert **1**).

[0062] So kann beispielsweise bei Anwendung des Verfahrens **300** in einem Fahrzeug **50**, in dem die Ergebnisse **13**, **14** einem in **Fig. 4** nicht eingezeichneten Fahrassistenzsystem **52a** oder System **52b** für das zumindest teilweise automatisierte Fahren zugeführt werden,

- gemäß Block **361** eine für den Fahrer des Fahrzeugs wahrnehmbare physikalische Warneinrichtung aktiviert werden,
- gemäß Block **362** der Fahrer des Fahrzeugs **50** dazu aufgefordert werden, das Ergebnis **13**, **14** zu bestätigen oder richtigzustellen,
- gemäß Block **363** der Fahrer des Fahrzeugs **50** dazu aufgefordert werden, die Kontrolle über das Fahrzeug **50** zu übernehmen, und/oder
- gemäß Block **364** das Fahrzeug **50** auf einer für den Ausfall der zumindest teilweise automatisierten Fahrfunktion vorgesehenen Notfalltrajektorie zum Stehen gebracht werden.

[0063] Weiterhin kann beispielsweise bei Anwendung des Verfahrens **300** in einem Zutrittskontrollsystem **60** zur Steuerung des Zutritts zu einem Raum **61**, einem Gelände **62** und/oder einem informationstechnischen System **63**, das auf der Grundlage der Ergebnisse **13**, **14** das Vorhandensein eines gültigen Zugangsmediums im Erfassungsbereich **2** prüft,

- gemäß Block **365** der Zutritt unabhängig vom Vorliegen eines gültigen Zugangsmediums **65** für eine vorbestimmte Zeit gesperrt werden,
- gemäß Block **366** eine zusätzliche Authentifikation von der den Zutritt begehrenden Person gefordert werden, und/oder
- gemäß Block **367** ein Alarm an eine für die Sicherheit des Raums **61**, des Geländes **62**, bzw. des informationstechnischen Systems **63**, verantwortliche Stelle ausgegeben werden.

[0064] Allgemein kann beispielsweise

- gemäß Block **370** der Mess-Datensatz **11** einem weiteren KI-Modul **1'** zugeführt werden, das anders trainiert ist und/oder eine andere Architektur aufweist, und/oder
- gemäß Block **380** ein weiterer Sensor **3'** für die Gewinnung zusätzlicher Messdaten **3a'** herangezogen werden, um das Ergebnis **13**, **14** gemäß Block **385** zu plausibilisieren, und/oder
- gemäß Block **390** das Ergebnis **13**, **14** verworfen werden.

[0065] Ist die Anfälligkeit **17**, **18** hingegen nicht auffällig hoch (Wahrheitswert **0** bei der Prüfung in Schritt **360**), so können die Ergebnisse **13**, **14** beispielsweise gemäß Block **395** im Fahrassistenzsystem **52a** oder im System **52b** für das zumindest teilweise automatisierte Fahren genutzt werden. Die Ergebnisse **13**, **14** können aber auch beispielsweise gemäß Block **396** im Zutrittskontrollsystem **60** genutzt werden.

[0066] Die Anwendung des Verfahrens **300** im Fahrzeug **50** ist in **Fig. 5** noch einmal kurz skizziert. Der Erfassungsbereich **2** ist hier Teil des Umfelds **51** des Fahrzeugs **50**. Das Fahrassistenzsystem **52a**, und/oder das System **52b** zumindest teilweise automatisierten Fahren, greifen abhängig von den mit dem Verfahren **300** aus den Mess-Datensätzen **11** abgeleiteten Ergebnissen **13**, **14** in die Fahrdynamik des Fahrzeugs ein, indem ein Lenksystem **53**, ein Antriebssystem **54**, und/oder ein Bremssystem **55**, des Fahrzeugs **50** angesteuert werden.

[0067] Die Anwendung des Verfahrens **300** im Zutrittskontrollsystem **60** ist in **Fig. 6** noch einmal kurz skizziert. Das Zutrittskontrollsystem **60** prüft, ob im Erfassungsbereich **2** ein gültiges Zugangsmedium **65** für den Zutritt zum Raum **61**, Gelände **62** oder IT-System **63** vorgelegt wird. In dem in **Fig. 6** gezeigten Beispiel ist das Zugangsmedium **65** eine Hand mit bestimmten biometrischen Merkmalen. Abhängig von den mit dem Verfahren **300** aus den Mess-Datensätzen **11** abgeleiteten Ergebnissen **13**, **14** steuert das Zutrittskontrollsystem **60** eine Sperr- und/oder Alarmvorrichtung **64** an, um den Zutritt zu gewähren oder zu verwehren.

ZITATE ENTHALTEN IN DER BESCHREIBUNG

Diese Liste der vom Anmelder aufgeführten Dokumente wurde automatisiert erzeugt und ist ausschließlich zur besseren Information des Lesers aufgenommen. Die Liste ist nicht Bestandteil der deutschen Patent- bzw. Gebrauchsmusteranmeldung. Das DPMA übernimmt keinerlei Haftung für etwaige Fehler oder Auslassungen.

Zitierte Patentliteratur

- DE 102018200724 A1 [0004]

Patentansprüche

1. Verfahren (100) zur Messung der Anfälligkeit eines KI-Moduls (1) gegen Täuschungsversuche, wobei das KI-Modul (1) eine trainierbare künstliche Intelligenz in Form einer parametrisierten internen Verarbeitungskette (12) umfasst und dazu ausgebildet ist, mittels dieser internen Verarbeitungskette (12) Eingabe-Datensätze (11) aus einem Eingaberaum E mit Dimensionalität D auf eine Klassifikation (13) und/oder Regression (14) in einem Ausgaberaum F mit Dimensionalität G < D abzubilden, mit den Schritten:
• zu einem vorgegebenen Auf-Datensatz (11) im Eingaberaum E wird die Klassifikation (13) und/oder Regression (14), auf die das KI-Modul (1) den Auf-Datensatz (11) abbildet, als ungestörtes Ergebnis (13, 14) ermittelt (110);
• dem Auf-Datensatz (11) wird mindestens eine Störung S mit einer Dimensionalität d < D aufgeprägt (120), so dass mindestens ein gestörter Datensatz (11') im Eingaberaum E entsteht;
• die Klassifikation (13') und/oder Regression (14'), auf die das KI-Modul (1) den gestörten Datensatz (11') abbildet, wird als gestörtes Ergebnis (13', 14') ermittelt (130);
• die Abweichung (15) des gestörten Ergebnisses (13', 14') von dem ungestörten Ergebnis (13, 14) wird mit einer vorgegebenen Metrik (16) ermittelt (140);
• in Antwort darauf, dass die Abweichung (15) ein vorgegebenes Kriterium erfüllt (150), wird festgestellt (160), dass das KI-Modul (1) in Bezug auf den Auf-Datensatz (11) für Täuschungsversuche mit Dimensionalität d anfällig ist.

2. Verfahren (100) nach Anspruch 1, wobei ein Adversarial Example für das KI-Modul (1) mit Dimensionalität d als Störung S ermittelt wird (121).

3. Verfahren (100) nach einem der Ansprüche 1 bis 2, wobei dem gleichen Auf-Datensatz (11) mehrere Störungen S mit unterschiedlichen Dimensionalitäten d aufgeprägt werden (122) und wobei die kleinste Dimensionalität c=min(d), für die eine Anfälligkeit des KI-Moduls (1) gegen Täuschungsversuche festgestellt wird (160), als Maß (17) für die Anfälligkeit des KI-Moduls (1) in Bezug auf den Auf-Datensatz (11) gewertet wird (170).

4. Verfahren (100) nach einem der Ansprüche 1 bis 3, wobei die Anfälligkeit des KI-Moduls (1) in Bezug auf mehrere Datensätze (11) aus einer vorgegebenen Menge M ermittelt wird (115) und wobei über die auf die so ermittelten Anfälligkeiten eine zusammenfassende Statistik (18) ermittelt wird (175).

5. Verfahren (100) nach Anspruch 4, wobei die zusammenfassende Statistik (18) einen Mittelwert, und/oder eine Varianz, und/oder eine Häufigkeitsverteilung, und/oder einen schlechtesten Wert der ermittelten Anfälligkeiten (17) beinhaltet.

6. Verfahren (200) zur Herstellung eines KI-Moduls (1), welches eine trainierbare künstliche Intelligenz in Form einer parametrisierten internen Verarbeitungskette (12) umfasst und dazu ausgebildet ist, mittels dieser internen Verarbeitungskette (12) Eingabe-Datensätze (11) aus einem Eingaberaum E mit Dimensionalität D auf eine Klassifikation (13) und/oder Regression (14) in einem Ausgaberaum F mit Dimensionalität G < D abzubilden, mit den Schritten:
• die Architektur der internen Verarbeitungskette (12) wird durch Hyperparameter (12a) festgelegt (210),
• es wird ein KI-Modul (1) mit dieser internen Verarbeitungskette (12) gebildet (220),
• das KI-Modul (1) wird trainiert (230), indem die Parameter (12b) der internen Verarbeitungskette anhand einer Menge L von Lern-Datensätzen (11a) und zugehörigen Lern-Ergebnissen (13a, 14a) so optimiert werden, dass das KI-Modul (1) die Lern-Datensätze (11a) nach Maßgabe einer Fehlerfunktion mit einer vorgegebenen Genauigkeit auf die Lern-Ergebnisse (13a, 14a) abbildet;
• mit einer Menge M von Validierungs-Datensätzen (11) wird mit dem Verfahren (100) nach einem der Ansprüche 4 bis 5 eine zusammenfassende Statistik (18) der Anfälligkeiten des trainierten KI-Moduls (1) ermittelt (240);
• die Hyperparameter (12a) werden dahingehend optimiert (250), dass nach erneutem Bilden (220) und Trainieren (230) des KI-Moduls (1) die hierfür ermittelte zusammenfassende Statistik (18) der Anfälligkeiten eine insgesamt geringere Anfälligkeit gegen Täuschungsversuche anzeigt.

7. Verfahren (200) nach Anspruch 6, wobei die Hyperparameter (12a) optimiert werden (250), indem Architekturen der internen Verarbeitungskette (12) mit einem evolutionären Algorithmus erzeugt werden (251), wobei die jeweils nach dem Bilden und Trainieren des KI-Moduls (1) mit einer Architektur ermittelte zusammenfassende Statistik (18) der Anfälligkeiten in ein Gütemaß für die Bewertung dieser Architektur eingeht (252).

8. Verfahren (100, 200) nach einem der Ansprüche 1 bis 7, wobei mindestens ein Auf-Datensatz (11), und/oder mindestens ein Lern-Datensatz (11a), mindestens einen Messwert einer physikalischen Messgröße enthält.

9. Verfahren (300) zur Überwachung eines Erfassungsbereichs (2) mit den Schritten:
• durch physikalische Beobachtung des Erfassungsbereichs (2) mit mindestens einem Sensor (3) wird mindestens ein Mess-Datensatz (11) mit Messdaten (3a) erfasst (310);
• der Mess-Datensatz (11) wird einem KI-Modul (1) zugeführt (320), welches eine trainierbare künstliche Intelligenz in Form einer parametrisierten internen Verarbeitungskette (12) umfasst und dazu ausgebildet ist, mittels dieser internen Verarbeitungskette (12)

Eingabe-Datensätze aus einem Eingaberaum E mit Dimensionalität D auf eine Klassifikation (13) und/oder Regression (14) in einem Ausgaberaum F mit Dimensionalität G < D abzubilden;
• die vom KI-Modul (1) ausgegebene Klassifikation (13) und/oder Regression (14) wird als Ergebnis (13, 14) der Überwachung gewertet und/oder ausgegeben (330);
• die Anfälligkeit (17, 18) des KI-Moduls (1) gegen Täuschungsversuche wird mit einem Verfahren (100) nach einem der Ansprüche 1 bis 5 gemessen (340);
• die ermittelte Anfälligkeit (17, 18) wird als Maß für die Verlässlichkeit des Ergebnisses (13, 14) gewertet und/oder ausgegeben (350).

10. Verfahren (300) nach Anspruch 9, wobei in Antwort darauf, dass die ermittelte Anfälligkeit (17, 18) ein vorgegebenes Kriterium erfüllt (360),
• der Mess-Datensatz (11) einem weiteren KI-Modul (1') zugeführt wird (370), wobei dieses weitere KI-Modul (1') eine andere Architektur aufweist als das zuvor genutzte KI-Modul (1), und/oder wobei das weitere KI-Modul (1') anders trainiert worden ist als das zuvor genutzte KI-Modul (1); und/oder
• mit einem weiteren physikalischen Sensor (3') zusätzliche Messdaten (3a') erfasst werden (380) und diese zusätzlichen Messdaten (3a') herangezogen werden (385), um das Ergebnis (13, 14) zu plausibilisieren; und/oder
• das Ergebnis (13, 14) verworfen wird (390).

11. Verfahren (300) nach einem der Ansprüche 9 bis 10, wobei der Erfassungsbereich (2) mindestens einen Teil des Umfelds (51) eines Fahrzeugs (50) umfasst und wobei das vom KI-Modul (1) gelieferte Ergebnis (13, 14) einem in dem Fahrzeug (50) verbauten Fahrassistenzsystem (52a) oder System (52b) für das zumindest teilweise automatisierte Fahren zugeführt wird (395), wobei das Fahrassistenzsystem (52a) oder System (52b) für das zumindest teilweise automatisierte Fahren dazu ausgebildet ist, abhängig von dem Ergebnis (13, 14) ein Lenksystem (53), ein Antriebssystem (54), und/oder ein Bremssystem (55), des Fahrzeugs (50) anzusteuern.

12. Verfahren (300) nach Anspruch 11, wobei in Antwort darauf, dass die ermittelte Anfälligkeit (17, 18) ein vorgegebenes Kriterium erfüllt (360),
• eine für einen Fahrer des Fahrzeugs (50) wahrnehmbare physikalische Warneinrichtung aktiviert wird (361),
• der Fahrer des Fahrzeugs (50) dazu aufgefordert wird, das Ergebnis (13, 14) zu bestätigen oder richtigzustellen (362),
• der Fahrer des Fahrzeugs (50) dazu aufgefordert wird, die Kontrolle über das Fahrzeug (50) zu übernehmen (363), und/oder
• das Fahrzeug (50) auf einer für den Ausfall der zumindest teilweise automatisierten Fahrfunktion vorgesehenen Notfalltrajektorie zum Stehen gebracht wird (364).

13. Verfahren (300) nach einem der Ansprüche 9 bis 10, wobei der Erfassungsbereich (2) mindestens einen Teil eines Bereiches umfasst, in dem ein Zutrittskontrollsystem (60) zur Steuerung des Zutritts zu einem Raum (61), einem Gelände (62) und/oder einem informationstechnischen System (63) die Vorlage eines Zugangsmediums (65) erwartet, wobei das vom KI-Modul (1) gelieferte Ergebnis (13, 14) dem Zutrittskontrollsystem (60) zugeführt wird (396) und wobei das Zutrittskontrollsystem (60) dazu ausgebildet ist, auf der Grundlage dieses Ergebnisses (13, 14) zu entscheiden, ob im Erfassungsbereich (2) ein gültiges Zugangsmedium (65) vorhanden ist.

14. Verfahren (300) nach Anspruch 13, wobei in Antwort darauf, dass die ermittelte Anfälligkeit (17, 18) ein vorgegebenes Kriterium erfüllt (360), der Zutritt unabhängig vom Vorliegen eines gültigen Zugangsmediums (65) für eine vorbestimmte Zeit gesperrt wird (365), und/oder eine zusätzliche Authentifikation von der den Zutritt begehrenden Person gefordert wird (366), und/oder ein Alarm an eine für die Sicherheit des Raums (61), des Geländes (62), bzw. des informationstechnischen Systems (63), verantwortliche Stelle ausgegeben wird (367).

15. Verfahren (300) nach einem der Ansprüche 13 bis 14, wobei biometrische Messdaten einer den Zutritt begehrenden Person als Messdaten (3a, 3a') gewählt werden.

16. Computerprogramm, enthaltend maschinenlesbare Anweisungen, die, wenn sie auf einem Computer, und/oder auf einem Steuergerät, und/oder auf einem Embedded-System, ausgeführt werden, den Computer, und/oder das Steuergerät, und/oder das Embedded-System, dazu veranlassen, ein Verfahren (100, 200, 300) nach einem der Ansprüche 1 bis 15 auszuführen.

Es folgen 6 Seiten Zeichnungen

Anhängende Zeichnungen

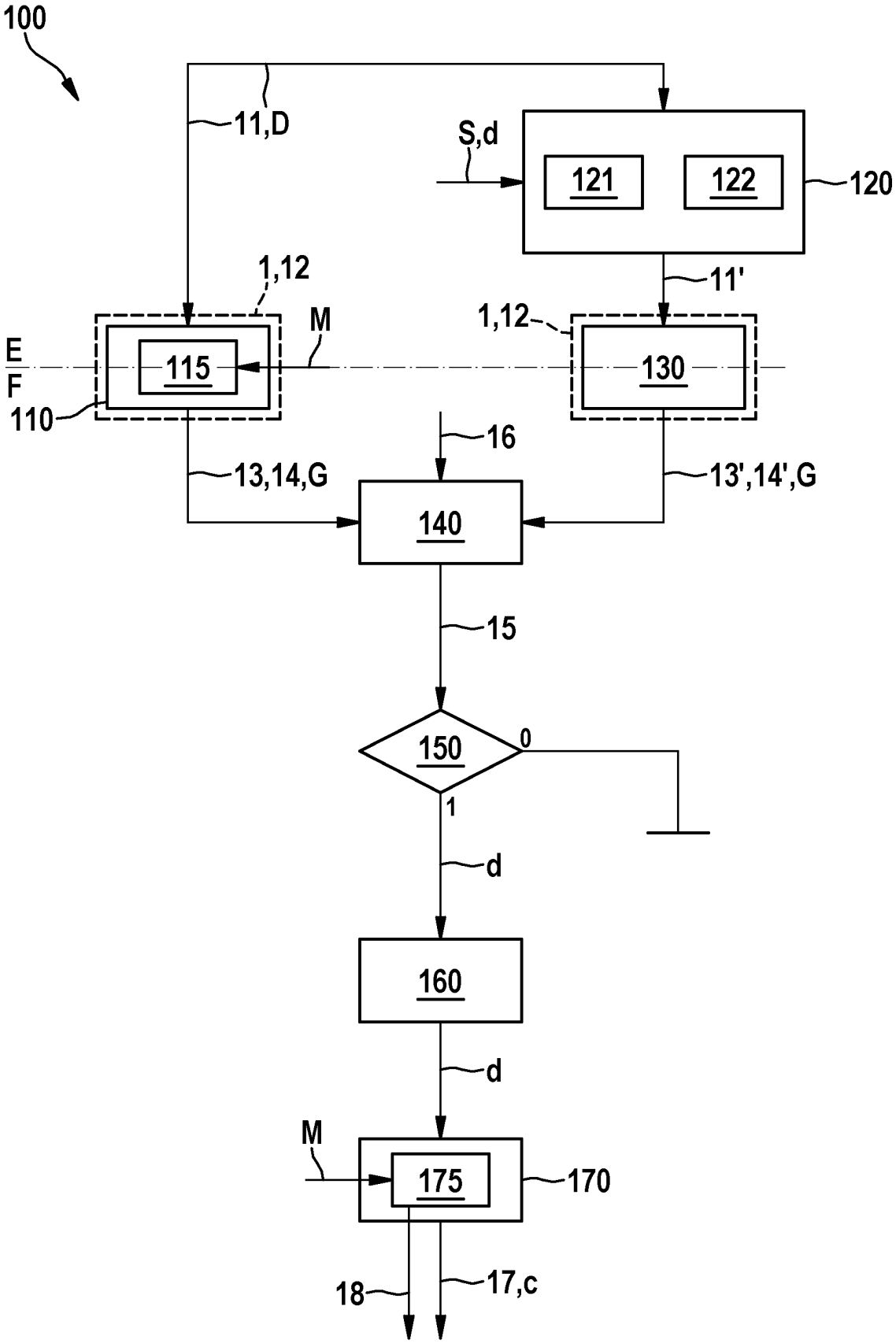


Fig. 1

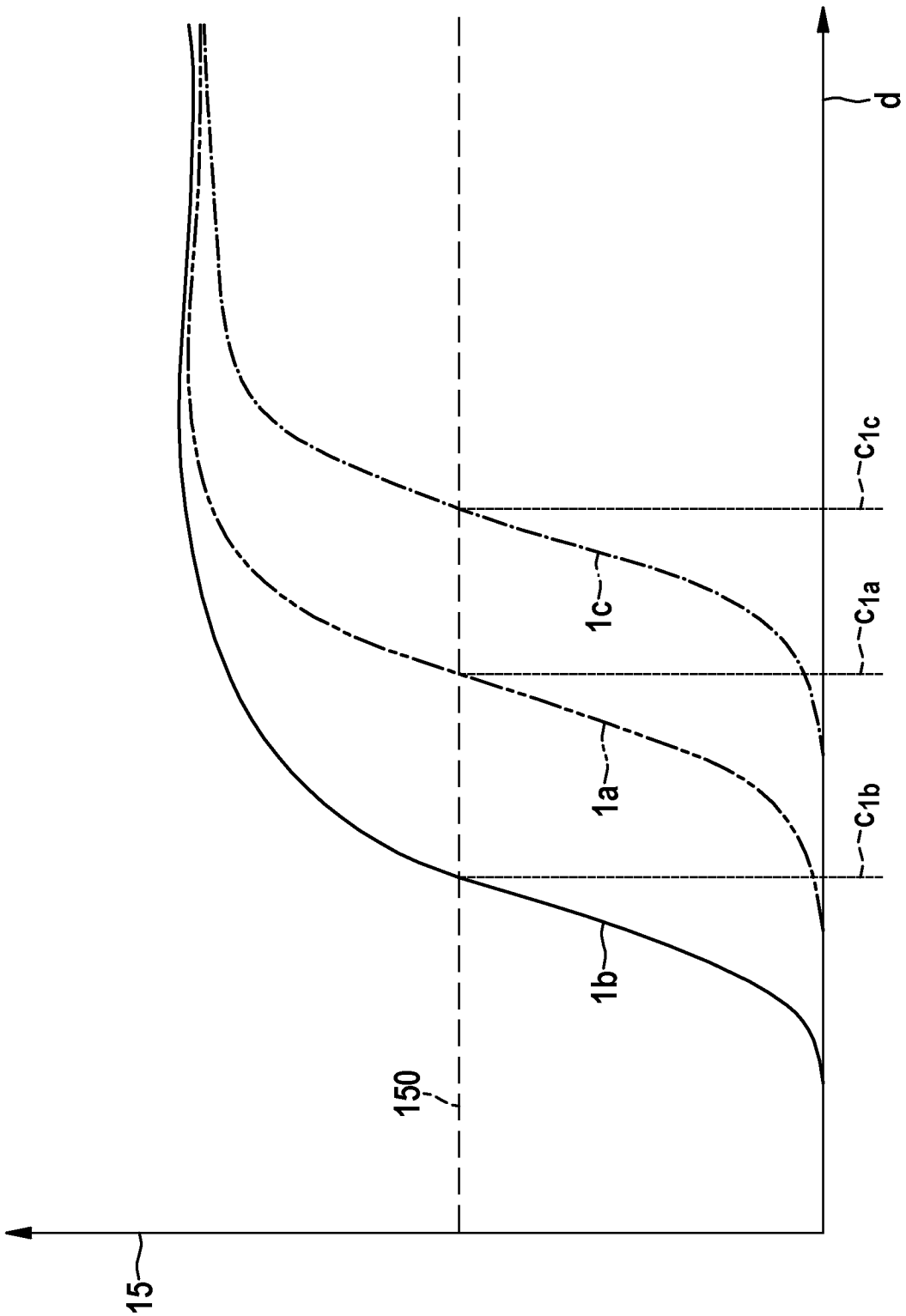


Fig. 2

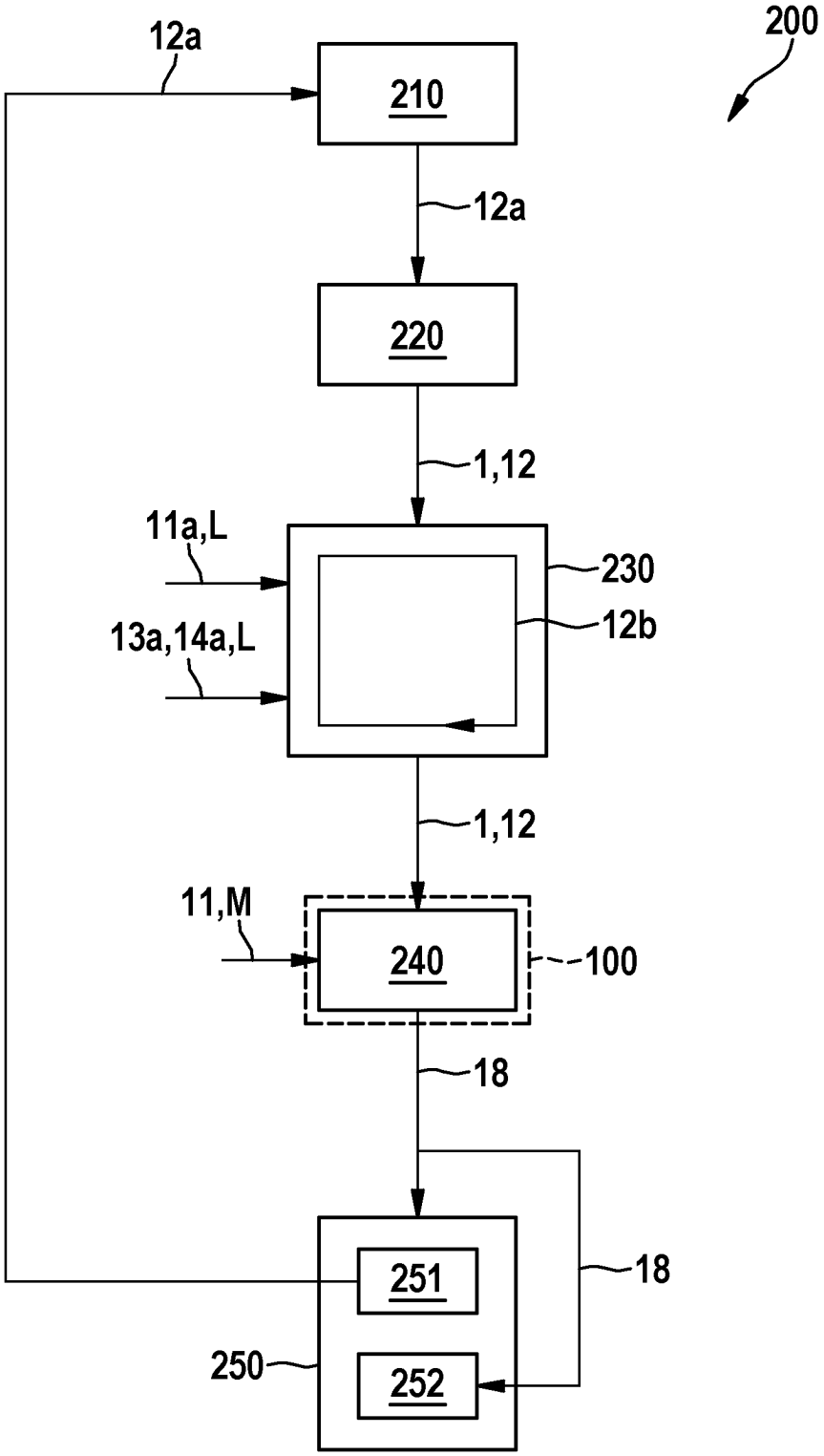

200
12a
210
12a
220
1,12
11a,L
230
12b
13a,14a,L
1,12
11,M
240
100
18
18
251
250
252

Fig. 3

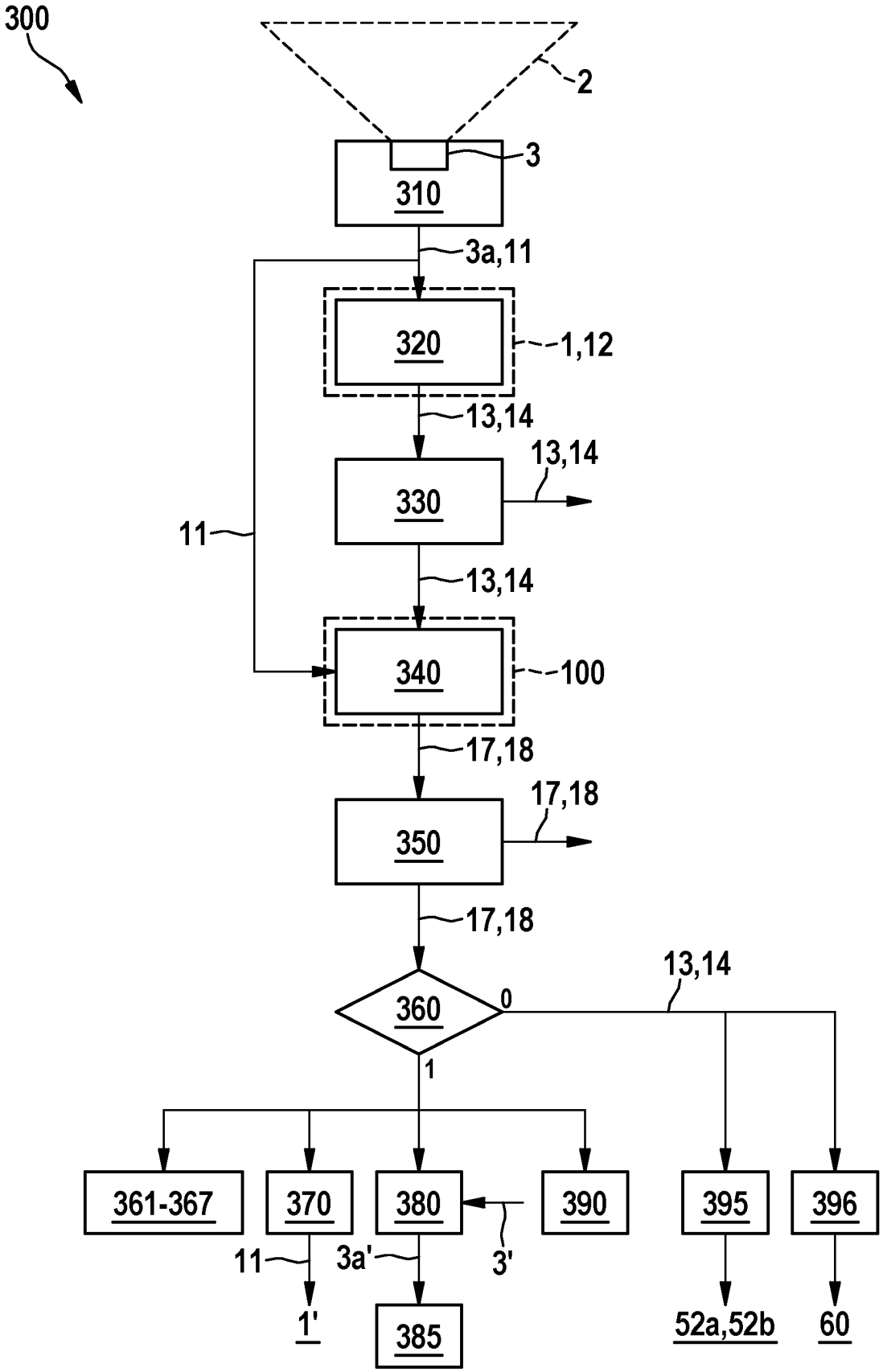

300
2
3
310
3a,11
320
1,12
13,14
13,14
330
11
13,14
340
100
17,18
17,18
350
17,18
13,14
360
0
1
361-367
370
380
390
395
396
11
3a'
3'
1'
385
52a,52b
60

Fig. 4

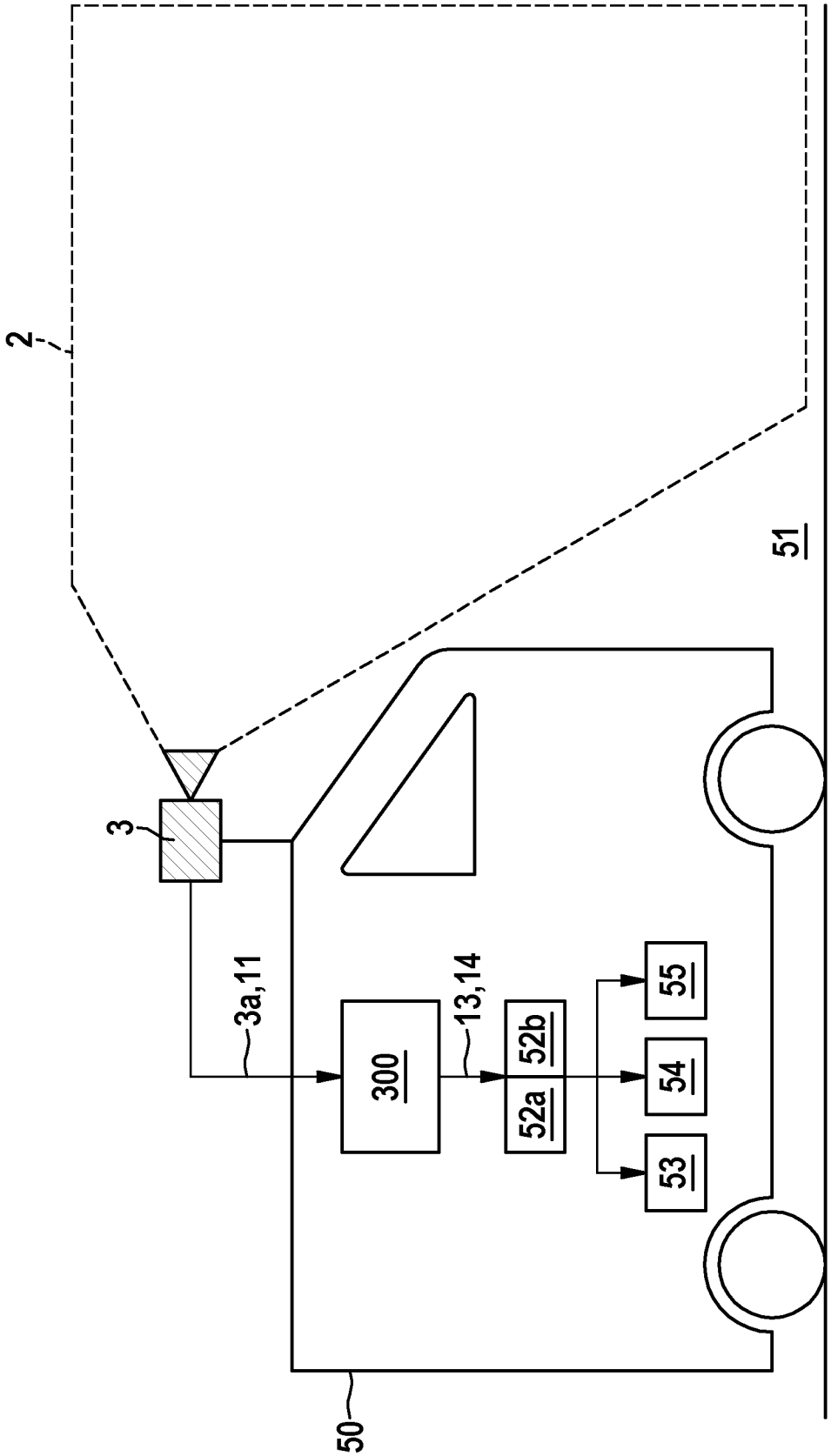


Fig. 5

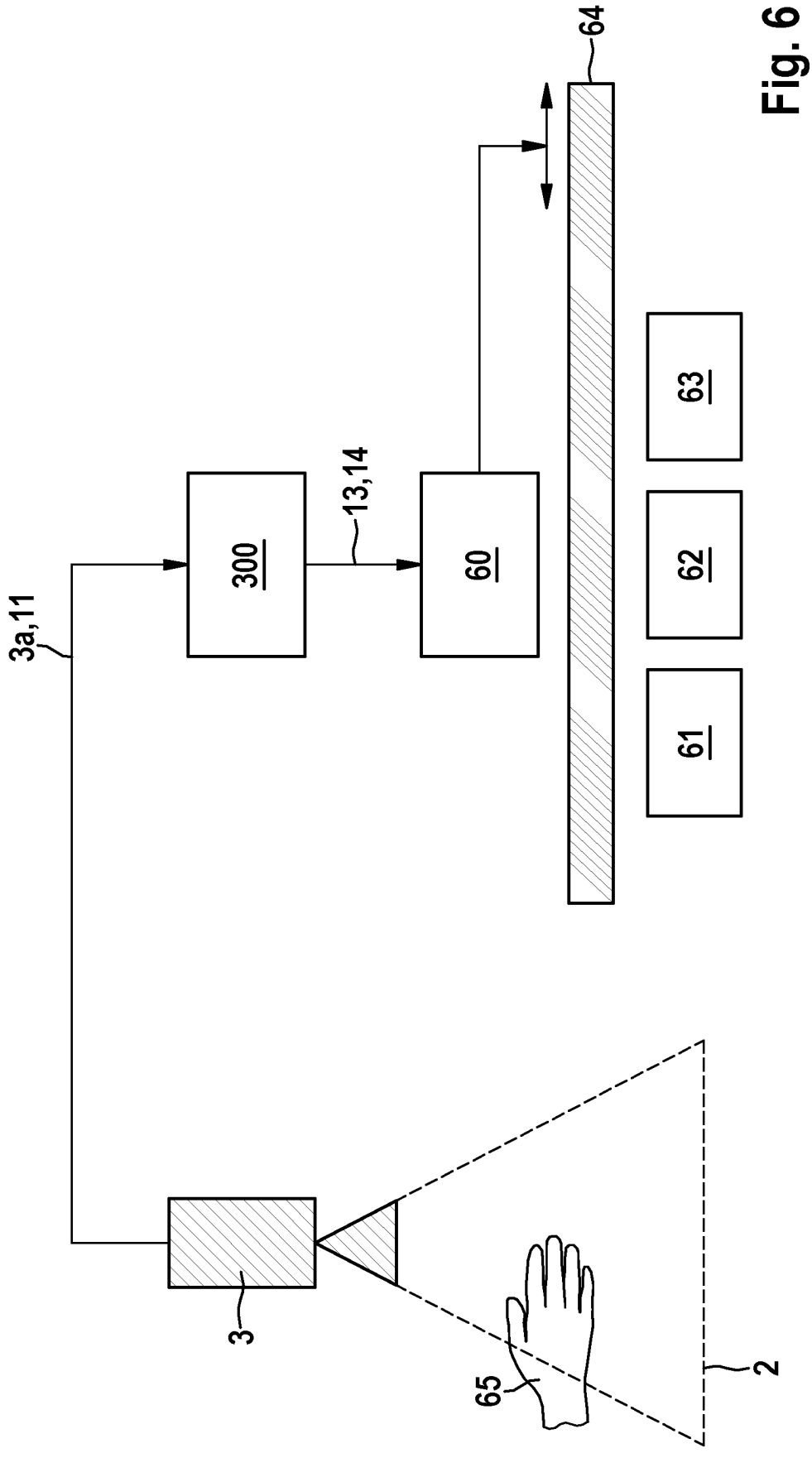


Fig. 6