Modellierung eines Photovoltaik-Fehlererkennungsansatzes unter Berücksichtigung von maschinellem Lernen

Philipp Akharath [1], Jaqueline Altkrüger [1], Harkiran Sahota [1], Volker Herbort [1], Henrik te Heesen [2]

Abstract: Die Photovoltaik spielt eine zentrale Rolle bei der Transformation des globalen Energiesystems zu einer emissionsfreien Energieversorgung. In Deutschland tragen rund 1,8 Millionen installierte Photovoltaikdachanlagen mit einer Nennleistung bis zu 30 kWp zur elektrischen Energieerzeugung bei. Dennoch zeigt sich insbesondere in dieser Anlagenklasse, dass durch eine fehlende Fernüberwachung der Stromproduktion und ein fehlendes Qualitätssicherungskonzept technische Störungen auftreten können und es damit zu einer Minderung des Ertrags dieser Photovoltaikanlagen kommt [HHR18; HHR19; SH21]. Bislang wurden zur Erkennung von solchen Anomalien manuelle Verfahren verwendet; *Machine-Learning*-Konzepte können eine automatisierte und adaptive Alternative darstellen. Hierfür wird die Implementierung eines *Isolation Forests* mit dem Ansatz von Leloux [Le20] hinsichtlich des methodischen Aufbaus verglichen. Zur Bewertung der verschiedenen Ergebnisse werden exemplarische Ertragsdaten von einzelnen Anlagen analysiert. Der vorliegende Datensatz besteht aus den jeweiligen Erträgen der Anlagen in fünfminütigen Intervallen sowie den nötigen Stammdaten von Photovoltaikanlagen im Südwesten Deutschlands. Die Gegenüberstellung der Verfahren zeigt, dass die Anomalieerkennung durch *Isolation Forests* Betriebsstörungen von Photovoltaikanlagen automatisch identifizieren kann.

Keywords: Photovoltaik; Datenanalyse; Maschinelles Lernen; Fehlererkennung; Anomalie; Isolation Tree; Isolation Forest

1 Einleitung

Die weltweite Anzahl von Photovoltaik (PV)-Installationen ist in den letzten Jahren durch die globale Energiewende stetig gestiegen. Mit dem Ziel, den Klimawandel zu stoppen, leisten insbesondere PV-Dachanlagen einen großen Beitrag zur Minderung von CO_2-Emissionen in Deutschland, durch die Erzeugung von elektrischer Energie [HHR19]. In Deutschland wurden bis April 2021 knapp 1,8 Millionen Dachanlagen errichtet [Bu20]. Die durch Photovoltaikanlagen erzeugte elektrische Energie hat sich in den letzten zehn Jahren mehr als verdoppelt, wie in Abbildung 1 zu sehen ist. Im Jahr 2020 betrug die Energieerzeugung 50,7 TWh [Bu21].

[1] Technische Hochschule Ulm, Fakultät Informatik, Prittwitzstraße 10, 89075 Ulm, Deutschland
volker.herbort@thu.de
[2] Hochschule Trier, Umwelt-Campus Birkenfeld, Institut für Betriebs- und Technologiemanagement, Campusallee, 55768 Hoppstädten-Weiersbach, Deutschland
h.teheesen@umwelt-campus.de

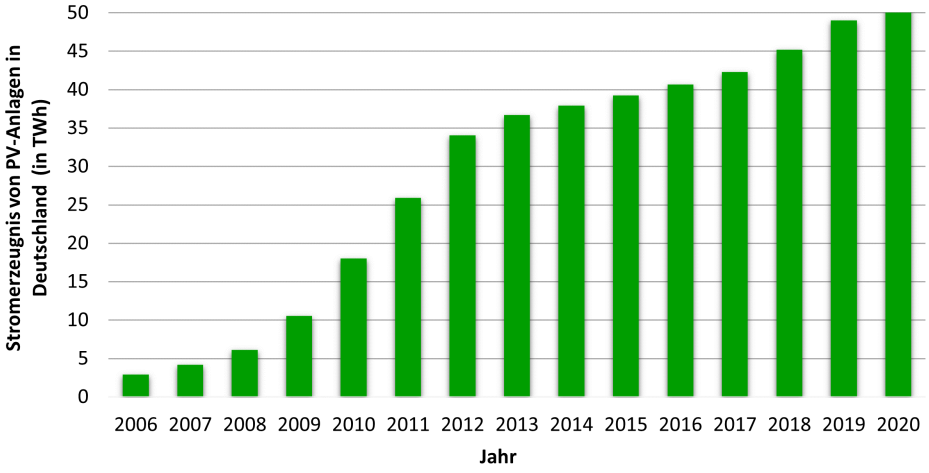

Abb. 1: Verlauf der produzierten Energie durch PV-Anlagen in Deutschland (Datenquelle [Bu21]).

Das Auftreten von technischen Störungen bei Photovoltaikanlagen kann jedoch die Energieproduktion und damit die Effizienz einer Anlage stark beeinträchtigen [Ga17; HH16; HHR18; SH21]. Dies kann durch verschiedene Arten von Fehlern verursacht werden. Einer der häufigsten Fehler ist eine technische Störung des Wechselrichters. Störungen bei den Photovoltaikmodulen treten seltener auf, können aber auch langfristig zu einer Ertragsminderung führen. Wird ein Fehler nicht rechtzeitig erkannt und behoben, beeinträchtigt dies die Produktivität und damit die Profitabilität einer PV-Anlage erheblich, weshalb ein verlässliches Fehlererkennungsverfahren unerlässlich ist [Ga17].

Solche Fehler lassen sich mithilfe von Anomalieerkennungsverfahren identifizieren, wobei abnorme Datenpunkte innerhalb eines Datensatzes ermittelt werden. Verfahren wie diese sind in unterschiedlichen Anwendungsdomänen einsetzbar, um seltene, aber signifikante Ereignisse zu erkennen [AMH16]. Ein automatisierter Ansatz zur Erkennung solcher Anomalien reduziert den händischen Aufwand zur Kalibrierung und ist adaptiv. Alternativ hierzu wurden bislang manuelle Verfahren zur Fehlererkennung eingesetzt, die mittels statistischer Methoden ein Profil regulärer Daten erstellen. Zur Ermittlung von Anomalien werden dabei Grenzwerte benötigt, welche von Domänenexperten festgelegt werden [St21].

Im Folgenden wird ein manueller Ansatz mit einer automatisierten Methode zur Anomalieerkennung im Kontext von Photovoltaikanlagen verglichen. Dabei soll die Applikabilität von *Machine-Learning*-Ansätzen zur Anomalieerkennung auf Basis exemplarischer Resultate nachgewiesen werden.

2 Grundlagen

2.1 State of the Art

Ein manuelles Anomalieerkennungsverfahren wird von Leloux [Le20] vorgestellt. In der Arbeit wird ein neuartiger Leistungsindikator - *Performance to Peer* (*P2P*) - zum Zweck der Fehlererkennung einer Solaranlage eingeführt, welcher allein auf dem Leistungsvergleich zwischen der betrachteten Anlage, im Folgenden als Fokusanlage bezeichnet, und den umliegenden Anlagen, im Folgenden als Nachbaranlagen bezeichnet, basiert. Dabei wird davon ausgegangen, dass Anlagen, die nah beieinander liegen (und damit vergleichbaren Einstrahlungsbedingungen ausgesetzt sind), und deren technische Parameter (Nennleistung, Ausrichtung, Neigung, etc.) ähnlich sind, vergleichbare spezifische Erträge produzieren.

Das P2P-Verfahren kann in zwei Phasen unterteilt werden. Zunächst erfordert die obige Annahme eine Einschätzung der Nachbaranlagen hinsichtlich ihrer Vergleichbarkeit mit der betrachteten Anlage. Hierbei beschränkt sich Leloux auf die umliegenden Nachbarn innerhalb eines Radius von 15 km. Der Ablauf der Bewertung der in Frage kommenden Nachbarn einer ausgewählten Fokusanlage ist in Abbildung 2 ersichtlich. Dabei wird die beschriebene Sequenz für jede der Nachbaranlagen ausgeführt. Um die Energieproduktion unterschiedlicher Anlagen effektiv Vergleichen zu können, werden die Ertragswerte im Bezug auf die installierte Nennleistung normalisiert. Hierzu wird der Kapazitätsauslastungsfaktor (*Capacity Utilisation Factor CUF*)

$$CUF = \frac{E_{\mathrm{PV}}}{P \cdot T} \tag{1}$$

aus der Energieproduktion E_{PV} über einen Zeitraum T und der Nennleistung P der Anlage berechnet. Zur Bestimmung des Ähnlichkeitsgrades einer Fokusanlage und einer ihrer Nachbaranlagen wird das sogenannte Kapazitätsauslastungsverhältnis (*Capacity Utilisation Ratio CUR*) zwischen zwei CUF-Werten zu einem bestimmten Zeitpunkt ermittelt als

$$CUR = \frac{CUF_{focus}}{CUF_{peer}} \quad . \tag{2}$$

Das Kapazitätsauslastungsverhältnis wird für jedes Fokus-Nachbar-Paar berechnet. Die Energieproduktion einer guten Nachbaranlage korreliert dabei stark mit der Fokusanlage, resultierend in einer geringen Schwankung des CUR-Wertes. Um diese Variabilität in den Werten untersuchen zu können ist die Auswahl einer geeigneten Zeitspanne wichtig. Diese sollte einerseits hinreichend lang sein, um statistisch repräsentativ zu sein und alle eventuellen Betriebszustände zu berücksichtigen, andererseits aber hinreichend kurz, um wesentliche Veränderungen in den Betriebszuständen wahrzunehmen. Leloux legt sich dabei auf einen effektiven Zeitraum von einem Monat fest. Als robustes Maß der Variabilität

in den berechneten CUR-Werten über den erwählten Zeitraum eines Fokus-Nachbar-Paares wird schließlich die mittlere absolute Abweichung vom Median ($MAD(CUR)$) angewendet. Mittels

$$MAD(CUR) = med(|CUR_i - med(CUR)|) \tag{3}$$

lässt sich so die mittlere absolute Abweichung vom Median, MAD, berechnen, welche zur Ermittlung des Gewichts w durch

$$w = \frac{1}{MAD^4} \tag{4}$$

verwendet wird. Der normalisierte Wert λ

$$\lambda_i = \frac{w_i}{\sum w_i} \tag{5}$$

dient schließlich dazu, die einzelnen Nachbarn einer Anlage bei einem Vergleich zu gewichten. Diese Gewichtungen, welche exemplarisch über einen erwählten Zeitraum berechnet werden, können anschließend zu jedem Zeitpunkt einer generellen Einschätzung der Nachbaranlagen dienen und sind nicht nur auf den angegebenen Zeitabschnitt zu beschränken.

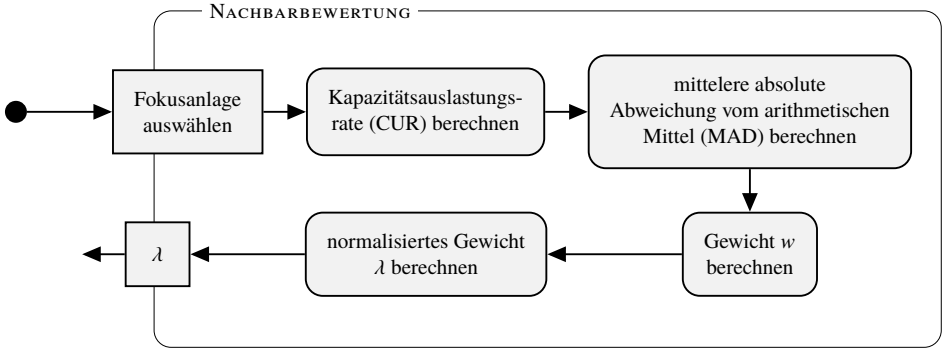

Abb. 2: Ablauf zur Bewertung der Nachbaranlagen gemäß Leloux [Le20].

Nachdem die Gewichtungen λ aller Nachbarn einer betrachteten Fokusanlage ermittelt wurden, kann in der zweiten Phase des Verfahrens nun der $P2P$-Wert zur Schätzung der Leistung eben dieser Anlage, durch den Vergleich mit den Nachbaranlagen zu einem beliebigen Zeitpunkt ermittelt werden. Über

$$P2P = \frac{CUF_{\text{focus}}}{CUF_{\text{ref}}} \tag{6}$$

lässt sich der $P2P$-Wert berechnen, indem der CUF-Wert der Anlage zur fraglichen Zeit mit den CUF-Werten der umliegenden Anlagen verglichen wird. Hierfür wird ein Referenz-CUF als gewichtetes 50. Perzentil, Q_{50}, aller CUF Werte der Nachbaranlagen berechnet über

$$CUF_{\text{ref}} = Q_{50}(CUF, \lambda) \quad . \tag{7}$$

Ein $P2P$-Wert von 1 indiziert, dass die Anlage im Vergleich zu den Nachbaranlagen einen vergleichbaren Ertrag erzielt. Aufgrund der Art, wie der $P2P$-Wert berechnet wird, zeigt ein Wert größer eins, dass der Ertrag dieser Anlage den relativen Ertrag seiner Nachbarn übertrifft. Ebenso ist es möglich, dass der $P2P$-Wert kleiner eins ist und der Ertrag damit kleiner als der Referenzertrag der Nachbaranlagen ist.

Da es sich hierbei um eine manuelle Methode zur Erkennung von Anomalien handelt, besteht der letzte Schritt darin, einen geeigneten Schwellwert für den berechneten $P2P$-Wert festzulegen, um die Leistung der betrachteten Anlage zu kategorisieren. Fällt der Wert einer Anlage zu einem bestimmten Zeitpunkt unter den Schwellwert, so kann davon ausgegangen werden, dass ein Fehlverhalten vorliegt. Hierfür präsentiert Leloux ein Verfahren zur dynamischen Festlegung eines solchen Grenzwertes, welches im Weiteren jedoch keine Anwendung findet.

Dieser *State of the Art*-Ansatz wird im Folgenden als *Ground Truth* verwendet, um die Ergebnisse des entwickelten *Machine-Learning*-Ansatzes zur Anomalieerkennung zu verifizieren [Le20].

2.2 Isolation Forest

Der in dieser Ausarbeitung verfolgte automatisierte Ansatz zur Fehleridentifikation basiert auf der vorgeschlagenen Methodik des *Isolation Forest* nach Liu [LTZ08]. Der bislang verfolgte Prozess zur Anomalieerkennung bestand darin, intakte Datenpunkte zu erkennen und nicht übereinstimmende Fälle als Anomalien zu kennzeichnen. *Isolation Forests* hingegen isolieren Anomalien vorrangig durch iterative Partitionierung, anstatt zunächst normale Datenpunkte über Gruppierungen zu identifizieren [Ch20]. Dieser Ansatz unterscheidet sich von anderen Verfahren, wie dem *Local Outlier Factor*, durch die Tatsache, dass keine Berechnungen auf Basis der Distanz oder Dichte von Datenpunkten verwendet werden. Daher weist das Verfahren des *Isolation Forests* einen reduzierten Rechenaufwand und Speicherbedarf, sowie eine lineare Zeitkomplexität auf [JN19]. Somit lässt sich dieser

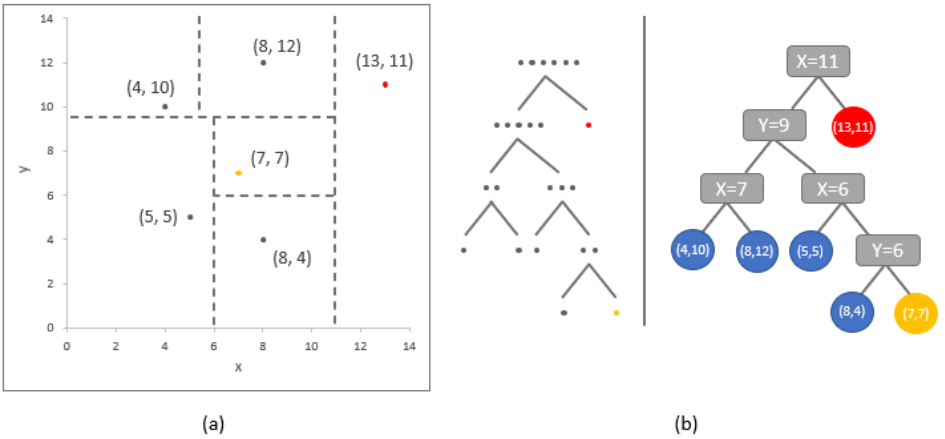

Abb. 3: Schematische Darstellung der Funktionsweise von *Isolation Forest* in der zweidimensionalen Ebene. Sechs Elementen werden jeweils ein Wertepaar (X, Y) zugewiesen (a). Über den binären Suchbaum (b) erfolgt eine Aufteilung der Instanzen und eine Isolation von Elementen (Datenquelle [Ch20])

Ansatz auch bei extrem großen Datenmengen effizient anwenden, liefert aber auch bei weniger Informationen zuverlässige Ergebnisse [LTZ08]. Dazu erfolgt zunächst die Erstellung zufälliger binärer Suchbäume, welche die Datenpunkte rekursiv partitionieren, bis alle Datenpunkte isoliert sind [Ch20].

Die Isolierung der Datenpunkte geschieht mittels eines zweistufigen Prozesses. Zunächst ist ein Training des Modells notwendig, bei dem die Baumstruktur gebildet wird. Bei der Erstellung eines Baumes wird ein zufälliges Merkmal der Daten selektiert und anschließend ein zufälliger Teilungswert zwischen dem minimalen und maximalen Wert dieses Attributes ausgewählt. Daraufhin werden die betrachteten Daten in zwei Gruppen entsprechend des Wertes dieses Attributs aufgeteilt. Dabei besteht die Annahme, dass Anomalien seltener sind als normale Daten und diese sich daher schneller vom restlichen Datensatz isolieren lassen. Dies zeigt sich durch eine kürzere Pfadlänge vom Wurzelknoten zum isolierten Datenpunkt. Dieser Vorgang wird in Abbildung 3 anhand eines zweidimensionalen Beispiels dargestellt. Die Verteilung der Datenpunkte (a) zeigt, dass die Instanz (13,11) abseits der restlichen Punkte liegt und direkt isoliert werden kann. Der so entstandene binäre Suchbaum wird in (b) dargestellt. Zur Erstellung eines *Isolation Forests* werden mehrere solcher zufällig verteilten binären Suchbäume generiert und deren Ergebnisse zusammengefasst interpretiert.

Anschließend erfolgt die Evaluierung der Daten, wobei alle Testinstanzen an jeden *Isolation Tree* innerhalb des *Isolation Forests* gereicht werden. Dabei werden *Anomaly Scores* mittels der durchschnittlichen Pfadlänge der einzelnen Datenpunkte berechnet und zurückgegeben.

Die *Anomaly Scores* kennzeichnen das Maß der Anomalie, wobei Werte nahe 1 mit großer Sicherheit als Anomalie interpretiert werden. Werte, die kleiner als 0,5 sind, werden als regulär angesehen. Falls alle Ergebnisse eines Datensatzes *Anomaly Scores* um 0,5 aufzeigen, wurden keine Anomalien durch den *Isolation Forest* identifiziert [LTZ08].

Ein zentraler Vorteil bei der Verwendung von *Isolation Forests* ist, dass keine Anomalien im Trainingsdatenset vorhanden sein müssen, um diese später auch zu identifizieren. Des Weiteren liefert dieses Verfahren schon bei vergleichsweise kleinen Datensets zuverlässige Ergebnisse. Da beide Aspekte für die vorliegenden Daten zutreffen, wird diese Methode im weiteren Verlauf angewandt [Ch20].

3 Methoden

Im Folgenden wird zunächst die praktische Umsetzung des manuellen Verfahrens zur Anomalieerkennung beschrieben, welche konzeptionell bereits in Abschnitt 2.1 vorgestellt wurde. Daraufhin folgt die Beschreibung der Implementierung des *Isolation Forests* basierend auf der Beschreibung in Abschnitt 2.2. Aufgrund der hohen Effektivität und Effizienz bei der Anomalieerkennung ist dies das Verfahren der Wahl in der vorliegenden Arbeit. Alternative Methoden wurden im Rahmen dieser Ausarbeitung nicht weiter beleuchtet. Die Umsetzung des *Isolation Forests* erfolgte in zwei Phasen. Aufgrund der nicht gekennzeichneten Daten wurden zuerst Zeitpunkte bestimmt, in denen Anomalien auftraten. Daraufhin fand eine Identifikation der fehlerhaften Anlagen bezogen auf die entsprechenden Zeiten statt. Diese Unterscheidung spiegelt sich in der Datenbeschaffung nicht wider, bei der Datenvorbereitung wurden aus diesem Grund zwei Verfahren unterschieden. Abschließend erfolgt die Interpretation der Ergebnisse durch den *Isolation Forest*.

3.1 Datenbeschaffung

Damit eine möglichst zuverlässige Aussage über Anomalien getroffen werden kann, sind vergleichbare Daten zu möglichst vielen PV-Anlagen in unmittelbarer Umgebung nötig. Zu diesem Zweck wurden frei verfügbare, webbasierte Ertrags- und Stammdaten des Fernüberwachungsanbieters Solare Datensysteme GmbH [So21] verwendet, um Informationen wie Postleitzahl, Nennleistung, Ausrichtung und Neigung zu beziehen [SH21]. Die Standortangaben verweisen dabei auf die Koordinaten der entsprechenden Postleitzahl und nicht dem tatsächlichen Standort der Anlage. Die benötigten Erträge in 5-Minutenintervallen wurden zusätzlich mit ein Scraping-Skript gesammelt.

Um die PV-Anlagen als Peers vergleichbar zu halten, müssen die technischen Eigenschaften (Ausrichtung, Neigung etc.) und der Standort der Peers hinreichend ähnlich sein, um zuverlässige Ergebnisse zu gewährleisten [Le20]. Aus diesem Grund wurden nur Daten von Aufdachanlagen berücksichtigt, die nach Süden ausgerichtet sind, eine Nennleistung

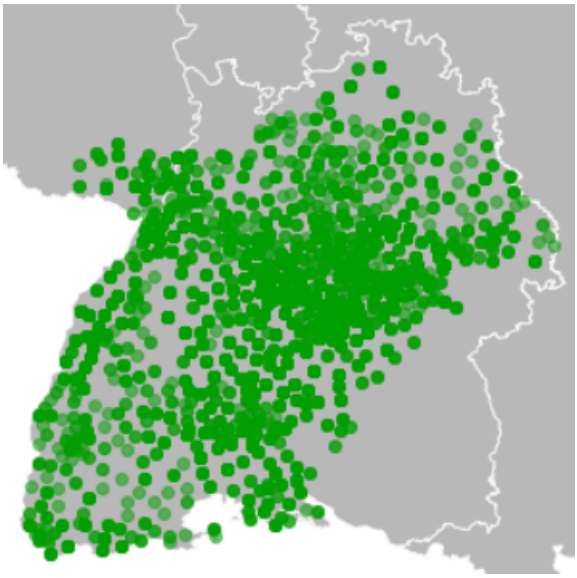

Abb. 4: Standorte der betrachteten Photovoltaikanlagen mit der Postleitzahl 7xxxx in den Bundeslädnern Baden-Württemberg und Rheinland-Pfalz.

bis 30 kW_p aufweisen und in der Postleitzahlregion beginnend mit der 7xxxx liegen (siehe Abbildung 4). Insgesamt besteht die Datenbasis aus 4 479 Photovoltaikanlagen, woraus die Daten für den exemplarischen Vergleich bezogen wurden, welche in Abschnitt 4 weiter beschrieben werden.

3.2 Datenaufbereitung

Um die Daten dem *Isolation Forest* zur Verfügung zu stellen und eine korrekte Anomalieerkennung zu gewährleisten, ist es notwendig, verschiedene Datenaufbereitungsschritte zu implementieren. Um einen Vergleich zwischen den Ergebnissen der manuellen Anomalieerkennung nach Leloux [Le20] zu ermöglichen, wurden die folgenden Schritte im Wesentlichen aus diesem Prozess nachgebildet. Demzufolge wurden die Nachbarn der betrachteten Anlage bestimmt, indem sämtliche PV-Anlagen identifiziert werden, die innerhalb eines 15 km-Radius liegen. Daraufhin wurden für jeden dieser Nachbarn die Daten im angegebenen Zeitintervall zusammengetragen. Dafür wurden sämtliche Datenpunkte für die untersuchten Anlagen auf ihre summierten Erträge von 5-Minutenintervallen auf eine Stunde aggregiert. Falls Anlagen nicht zwölf Datenpunkte innerhalb des Stundenintervalls aufwiesen, wurden diesen Anlagen in der Datenanalyse nicht weiter berücksichtigt. Anschließend werden die Stromerträge in Bezug auf die installierte Nennleistung normiert. Dieses Vorgehen ist in Abbildung 5 dargestellt.

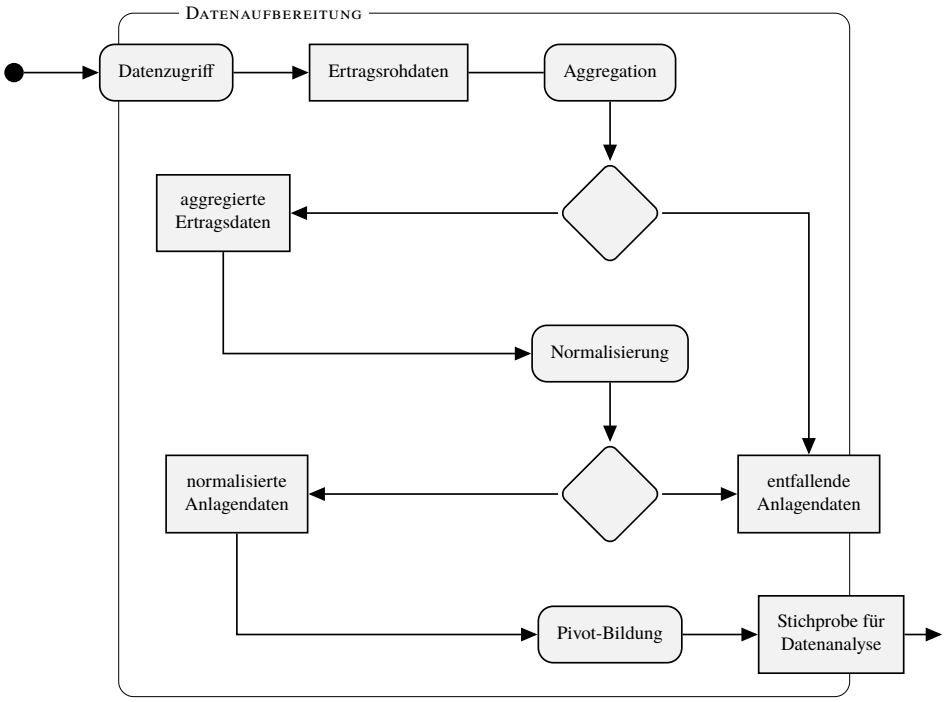

Abb. 5: Datenaufbereitung für die Datenanalyse. Aus den Rohertragsdaten werden Stundenertragswerte aggregiert. PV-Anlagen mit Datenlücken werden aus der Datenauswertung ausgeschlossen. Die Erträge werden normalisiert und pivotiert.

Die Ertragsdaten auf stündlicher Basis werden nach der Datenbereinigung pivotiert, um die aggregierten Zeitstempel in Zeilen und die zugehörigen Anlagen in Spalten zu erhalten. Die Zellen werden anschließend mit den spezifischen Erträgen gefüllt. Dieser Datensatz diente als Basis zur Bestimmung des Zeitpunktes, zu welchem eine Anomalie auftritt.

Für den zweiten Schritt wird bestimmt, welche der betrachteten PV-Anlagen zu dem eben bestimmten Zeitpunkt Ertragsanomalien erzeugen. Hierfür werden die Nachbarn der betrachteten Anlage bestimmt und deren Daten mithilfe der Nennleistung normalisiert. Eine Aggregation nach der Zeit entfällt. Die Pivotierung der Daten zeigt die Anlagen als Zeilen und 5-Minutenintervalle für die betreffende Stunde als Spalten. Dieses Vorgehen ist in Abbildung 6 veranschaulicht.

Die räumliche Nähe zweier Peers ist ein entscheidender Faktor für ihre Ähnlichkeit. Da bereits kleinere regionale Wetteränderungen die spezifischen Erträge einer PV-Anlage beeinflussen können, muss der Standort eine übergeordnete Rolle spielen. Um sich auf lokal nahe gelegene Peers zu konzentrieren, schlägt Leloux vor, nur benachbarte Anlagen

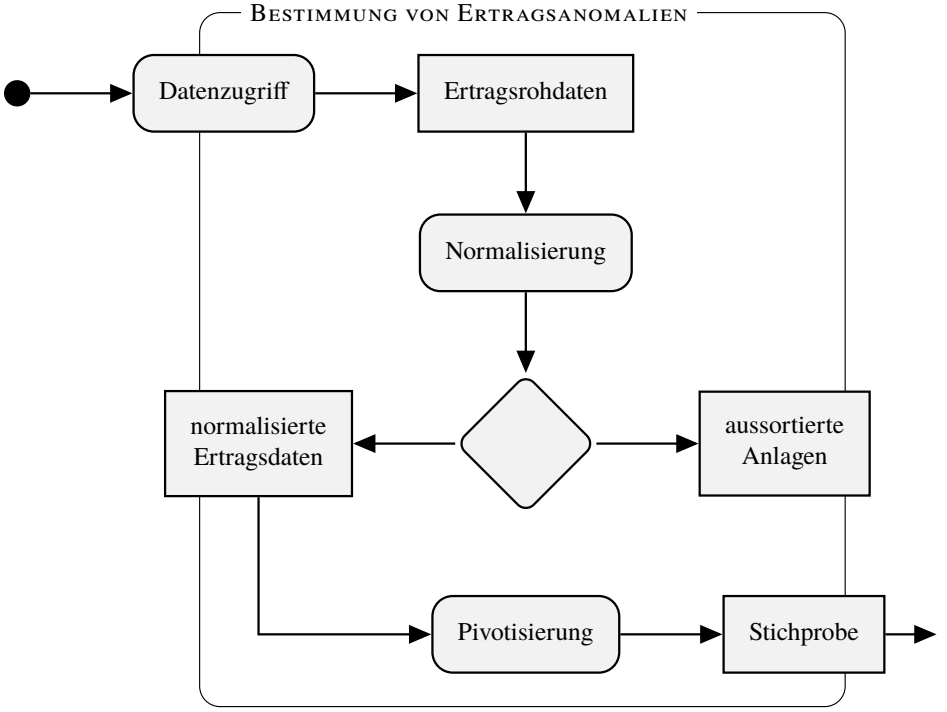

Abb. 6: Identifikation von Ertragsanomalien durch Bestimmung von Nachbaranlagen.

innerhalb eines Bereichs von 15 km zu berücksichtigen [Le20]. Um dies für den weiteren maschinellen Lernansatz umsetzen zu können, liefern die Datenvorbereitungsfunktionen die notwendigen Datenmatrizen für einen bestimmten gegebenen Fokus-Peer und filtert alle Anlagen, die nicht in diesem räumlichen Bereich liegen. Dies schränkt auch die Interpretation der Ergebnisse auf den oder die Fokus-Peer(s) ein, da für Anlagen, die am Rande dieses 15 km Radius liegen, Nachbarinformationen fehlen und die Ergebnisse nicht als vergleichbar angesehen werden können.

3.3 State-of-the-Art-Implementierung

Um einen *Ground Truth* für die folgende Analyse zu schaffen, wurde der in Abschnitt 2.1 beschriebene Ansatz von Leloux [Le20] implementiert. Hierbei soll es möglich sein, die Stromertragsproduktion einer Fokusanlage unter Einbeziehung ihrer Nachbarn zu bewerten. Dies bedarf zunächst einer Einschätzung aller Nachbarn in Bezug auf ihre Vergleichbarkeit zur Fokusanlage. Für die Berechnungen zur Gewichtung der Nachbaranlagen werden exemplarisch die Erzeugnisse der PV-Anlagen im Sommermonat August 2019 in Betracht

gezogen. Für jede Fokusanlage werden dabei ausschließlich die Nachbarn innerhalb eines 15 km-Radius beachtet. In der hier realisierten Variante wird zusätzlich die Restriktion eingeführt, dass lediglich Anlagen berücksichtigt werden, deren Standort in der Postleitzahlregion beginnend mit einer 7 liegt, um die Anzahl der PV-Anlagen zu reduzieren und die Machbarkeit des Vorgehens zu demonstrieren.

Für die somit infrage kommenden Nachbarn einer Fokusanlage werden für jeden der 31 Tagen 12 CUF zwischen 8 und 20 Uhr berechnet. Dazu werden die 5-Minutenwerte auf Stundenwerte aggregiert. Anlagen, welche nicht durchgehend Werte im festgelegten Zeitintervall liefern, werden verworfen. Anschließend können die stündlichen CUR-Werte für den betrachteten Monat berechnet werden, welche zur Ermittlung des einzelnen MAD Wertes pro Nachbar verwendet werden. Daraus werden schließlich die Gewichtung w und das normalisierte Gewicht λ berechnet.

Anschließend können in der zweiten Phase die $P2P$-Werte einer Fokusanlage berechnet und die Funktionsfähigkeit an einem bestimmten Tag eingeschätzt werden. Hierzu werden zunächst die stündlichen CUF-Werte dieser Anlage berechnet und für die Nachbaranlagen über die dazugehörigen λ-Werte die stündlichen Referenz CUF-Werte ermittelt. Somit können zwölf $P2P$-Werte ermittelt werden, um eine stündliche Einschätzung der Fokusanlage zu ermöglichen. Zur statischen Einschätzung der Anlagen muss ein Grenzwert definiert werden, welcher zur Kategorisierung der Anlagen dient. Dieser wird auf 0,85 festgelegt, um sämtliche Anlagen mit einem geringeren $P2P$-Wert als nicht funktionsfähig zu kategorisieren. Der Wert ergibt sich durch die Angabe von Leloux [Le20], dass eine funktionsfähige Anlage auch einen $P2P$-Wert von 90 Prozent annehmen kann. Ein zusätzlicher Interpretationsspielraum soll kleinere Abweichungen in den ersten Analysen zu Gunsten des *Isolation Forests* ausgleichen, weshalb dieser Grenzwert etwas verringert wird.

3.4 Dateninterpretation

An dieser Stelle werden die vorbereiteten Daten dem *Machine-Learnign*-Algorithmus übergeben. Dabei ist die Handhabung dieses Prozesses abhängig von den verwendeten Werkzeugen. Für den Einsatz des *Isolation-Forest*-Verfahrens wird die Python3-Bibliothek *Scikit-Learn* verwendet. Da die Anzahl der eingesetzten Bäume erheblichen Einfluss auf das Ergebnis hat, wurde der Wert 100 verwendet, welcher nach Liu [LTZ08] als optimal angesehen wird. Alle weiteren Parameter des Modells wurden entsprechend der Standardwerte übernommen.

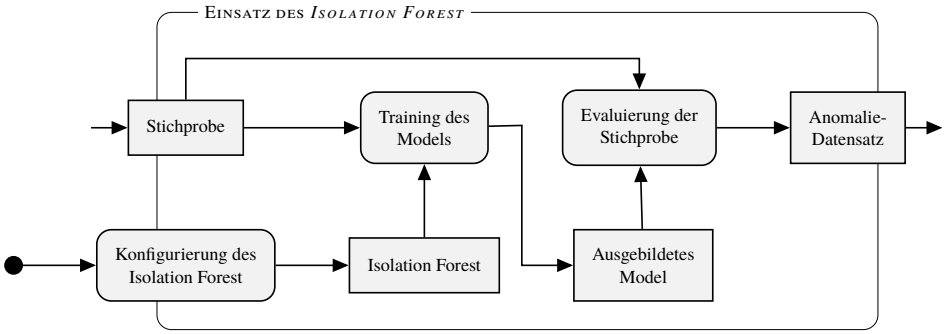

Abb. 7: Ermittlung von Anomalien durch den Einsatz des *Isolation Forest*.

In Abbildung 7 wird der *Isolation-Forest*-Prozess dargestellt. Dabei wird der Datensatz jeweils zum Training und zur Evaluation verwendet und anschließend den ermittelten Anomalien zugeordnet.

Für das Ermitteln einer Anomalie bezogen auf eine bestimmte Anlage und in Relation zu den dazugehörigen Nachbarn war ein iteratives Vorgehen mit den jeweiligen Datenvorbereitungen erforderlich. Aufgrund der fehlenden Kennzeichnung der vorliegenden Daten hinsichtlich ihrer Funktionstüchtigkeit wurden zunächst Tage ermittelt, welche eine Anomalie aufwiesen. Hierfür wurde die Datenvorbereitung nach Zeitpunkten verwendet. Der daraus resultierende anomalische Datensatz wurde herangezogen, um eine weitere Eingrenzung zu treffen. Dabei wurde der zu untersuchende Tag extrahiert, um daraufhin die Datenvorbereitung nach Anlagen durchzuführen. Nach Identifikation der anomalischen Anlagen mithilfe des *Isolation Forests* konnte der Datensatz der Anomalien zu den jeweiligen Anlagen zugeordnet werden. Das Format dieses Resultats ermöglicht einen Vergleich mit den jeweiligen *P2P*-Werten bezogen auf eine Anlage.

4 Ergebnisse und Diskussion

Ersichtlich ist, dass sowohl das Verfahren von Leloux [Le20] als auch die hier dargestellte Methode mittels eines *Isolation Forests* vielversprechende Ergebnisse liefern kann, um die Funktionsfähigkeit von Photovoltaikanlagen zu bewerten. Die Ergebnisse der exemplarischen Datenanalyse sind in Tabelle 1 dargestellt. Hierbei werden einzelne Anlagen an einem bestimmten Tag und zu einer bestimmten Stunde analysiert und die Ergebnisse der beiden Verfahren gegenübergestellt. Das Verfahren von Leloux [Le20] liefert den in der Studie entwickelten *P2P*-Wert. Dieser kennzeichnet die Leistung einer PV-Anlage im Vergleich zu den umliegenden Nachbarn. Werte, die dabei unter dem in Abschnitt 3.3 festgelegten

Tab. 1: Gegenüberstellung exemplarischer Resultate aus der Datenanalyse. Der *P2P*-Wert wurde nach Leloux [Le20] bestimmt. Die Klassifizierung nach dem *Isolation Forest* beträgt 1 für eine fehlerfreie Photovoltaikanlage und −1 für eine Anlage mit einer technischen Störung.

Anlagen-ID	Datum	Zeitspanne	*P2P*	*Isolation Forest*	
				Klassifizierung	*Anomaly Score*
12246	10.08.2019	14:00-15:00	1,033	1	0,372
15863	01.09.2020	11:00-12:00	1,311	-1	0,580
24874	10.08.2019	14:00-15:00	0,126	-1	0,697
25731	11.06.2020	09:00-10:00	0,860	1	0,366
25731	20.06.2020	09:00-10:00	0,857	1	0,386

Grenzwert von 0,85 liegen, kennzeichnen eine mangelhafte Anlage. Auch eine starke Überperformance bei Werten über 1,15 kann ein Kennzeichen für ein Fehlverhalten sein. Der *Isolation Forest* liefert eine automatisierte Identifizierung der Anomalien; eine -1 für abnorme und eine 1 für reguläre Werte. Diese Einschätzung geschieht anhand der *Anomaly Scores*, welche bei einem Wert nahe 1 eine Diskrepanz in den Daten feststellen.

Durch die Übereinstimmung der Ergebnisse in Tabelle 1 des *Isolation Forests* mit den eingangs als *Ground Truth* angenommenen *P2P*-Werten kann ersteres als plausible Alternative zu manuellen Verfahren angenommen werden. Die Zusammenhänge der verschiedenen Werte können bei näherer Betrachtung einzelner Zeilen erkannt werden. Die Anlage mit der ID 25731 hat beispielsweise am 11.06.2020 einen *P2P*-Wert von 0,860; somit lässt sie sich mit dem dazugehörigen statischen Grenzwert von 0,85 als funktionsfähig einstufen. Die respektiven Resultate des *Isolation Forest* stufen diese Anlage ebenso als funktionstüchtig ein, ersichtlich in der Klassifizierung von 1 und dem dazugehörigen *Anomaly Score* von 0,366. Dabei kennzeichnet der *Anomaly Score*, dass das Modell des *Isolation Forests* diese Stichprobe nur mit einer Wahrscheinlichkeit von 36 Prozent als Anomalie klassifizieren konnte. Anhand dieser exemplarischen Resultate können erste Annahmen getroffen werden, dass ein solcher *Machine-Learning*-Ansatz als funktionsfähig betrachtet werden kann.

Im Laufe der Implementierung haben sich besondere Charakteristiken des *Isolation Forests* gezeigt, welche für die Bestimmung von Anomalien bei den vorliegenden Daten zu Photovoltaikanlagen besonders von Nutzen sind. Wesentlich ist eine umfassende Datenaufbereitung der Rohdaten, um PV-Anlagen mit offensichtliche Fehlkonfigurationen, die zu Artefakten bei der Datenanalyse führen können, auszuschließen. Für den Vergleich von Anlagen zu deren Nachbarn ist des weiteren die Tatsache relevant, dass *Isolation Forests* gut auf kleineren Datensets anwendbar sind. Dies ist vor allem für Regionen mit einer geringeren Anlagendichte relevant.

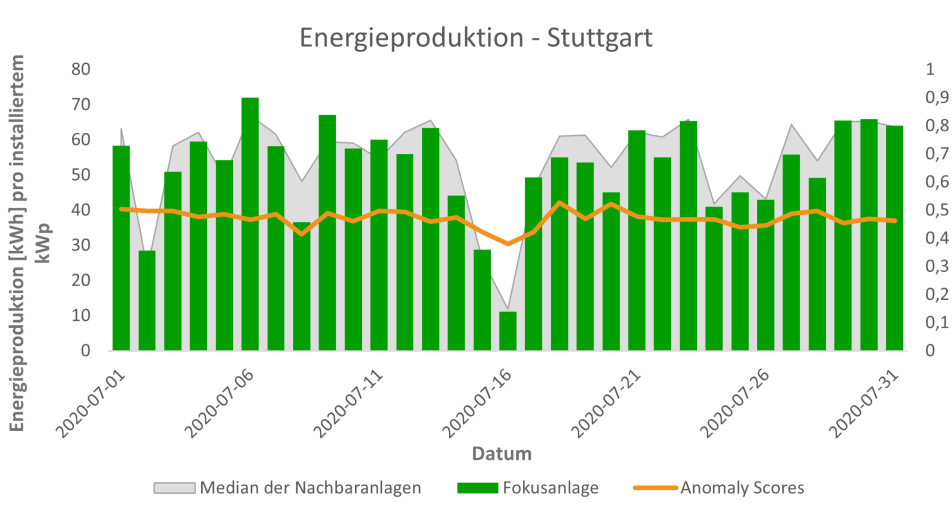

Abb. 8: Spezifischer Ertrag (grüne Balken) und *Anomaly Score* (orange Linie) der Photovoltaikanlage mit der ID 12246 (Standort Stuttgart) im Juli 2020.

Da die beiden Anlagen in Ostfildern (Anlagen-ID 24874) und Stuttgart (Anlagen-ID 12246) mit einer Distanz von circa 9 km als direkte Nachbarn angesehen werden, folgt eine detaillierte Interpretation der Ergebnisse dieser beiden Installationen.

Die Anlage in Stuttgart wird vom *Isolation Forest* als regulär eingestuft. Zur Validierung dieser Einschätzung werden die entsprechenden Daten manuell analysiert. In Abbildung 8 sind die aggregierten Messwerte der einzelnen Tage vom Juli 2020 als grüne Balken dargestellt. Dabei bildet die x-Achse die Zeit in Tagen ab. Die linke y-Achse zeigt die Energieproduktion in kWh, die rechte y-Achse veranschaulicht den *Anomaly Score*. Die graue Fläche im Hintergrund zeigt die summierten Erträge der Nachbaranlagen im Verhältnis zu deren Nennleistung. Der *Anomaly Score* der einzelnen Tage wird durch den Verlauf der orangen Linie gekennzeichnet.

Auffällig ist, dass sowohl der Ertrag der Fokusanlage als auch die der Nachbaranlagen deutlich schwanken. Dies lässt sich durch die lokalen Wetter- und Einstrahlungsbedingungen erklären [De21]. Besonders die Tage am 02.07.2020, 15.07.2020 und 16.07.2020 sind hiervon betroffen. Ersichtlich ist dies auch in Abbildung 8, da sowohl die Fokusanlage als auch die Nachbaranlagen an diesen Tagen besonders geringe Erträge aufweisen. Durch den *Isolation Forest* werden in dem gesamten Monat keinerlei Anomalien verzeichnet, was durch durchgängige Werte um 0,5 ersichtlich ist [LTZ08].

Die Anlage in Ostfildern hingegen weist für alle Tage im Juli 2020 Anomalien auf (Abbildung 9). Die Abbildung ist ebenso zu interpretieren wie Abbildung 8. Hier ist vor

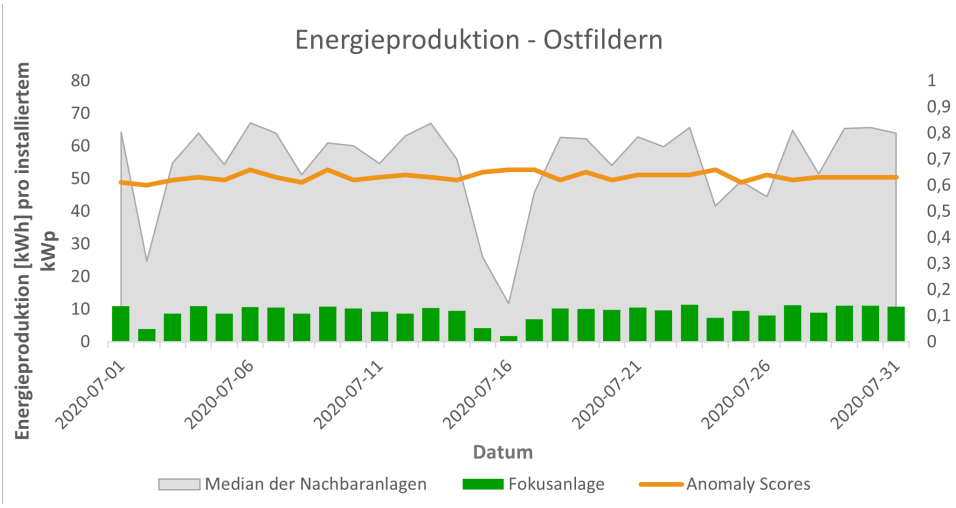

Abb. 9: Spezifischer Ertrag (grüne Balken) und *Anomaly Score* (orange Linie) der Photovoltaikanlage mit der ID 24874 (Standort Ostfildern) im Juli 2020.

allem auffällig, dass die Erträge der Fokusanlage im Verhältnis zu deren Nennleistung deutlich geringer sind als die normalisierten Erträge der Nachbarn. Grund hierfür kann entweder eine falsche Angabe der Nennleistung oder ein technischer Ausfall eines oder mehrerer Wechselrichter sein.

5 Zusammenfassung

Eine automatisierte Identifikation von technischen Störungen bei Photovoltaikanlagen, insbesondere von Dachanlagen mit einer Nennleistung bis 30 kW_p, dient einer schnellen Fehlererkennung, um eine Beseitigung von technischen Störungen einzuleiten. Hierzu wird die Tauglichkeit automatisierter Anomalieidentifikationsverfahren zur effektiven Einschätzung der Ertragsqualität einzelner Photovoltaikanlagen geprüft.

Zusammenfassend lässt sich festhalten, dass ein *Isolation Forest* als ein solches automatisiertes Verfahren mittels maschinellem Lernen anwendbar ist und eine gute Alternative zu den herkömmlichen, manuellen Verfahren (wie zum Beispiel das *P2P*-Verfahren nach Leloux [Le20]) bietet. Die Anwendung des *Isolation Forests* kommt im Gegensatz zum *P2P*-Ansatz ohne vorherige Gewichtung von Nachbaranlagen und Parametrisierung von Grenzen zur Fehlererkennung aus. Anhand eines exemplarischen Datensatzes aus den 4 479 PV-Anlagen im Südwesten Deutschlands konnte gezeigt werden, dass technische Störungen und damit verbundene Stromproduktionsausfälle durch den *Isolation Forest* erkannt werden können.

Eine ausführliche Analyse des Ansatzes anhand einer quantitativen Evaluierung ist Umfang der zukünftigen Forschungsarbeit. Auch die mögliche Parametrisierung des *Isolation Forests* zur verbesserten Anomalieerkennung ist Gegenstand weiterer Ausarbeitungen. Die Klassifizierung der identifizierten Anomalien nach verschiedenen Fehlerarten bietet ebenfalls Grundlage zur Anwendung weiterer *Machine-Learning*-Ansätze.

Zukünftig könnte die Umsetzung von Anomalieerkenunng mithilfe von automatisierten Verfahren innerhalb privater oder gewerblich genutzter Photovoltaikanlagen eine effektive Möglichkeit für Anwender bieten, Ausfälle zu erkennen und diesen gegenwirken zu können. Eine zentrale Implementierung des präsentierten Verfahrens ermöglicht den Zugriff auf die notwendigen Daten benachbarter Anlagen und das Bereitstellen eines Services zum Abfragen der Funktionsfähigkeit einzelner Anlagen. Durch die Realisierung einer Fehlerfrüherkennung könnten so die Gesamtstörungszeiten von Photovoltaikanlagen reduziert werden. Infolgedessen ließe sich der Stromertrag von Photovoltaikanlagen verbessern und somit ein nachhaltiger Beitrag zur Energiewende erreichen.

Literatur

[AMH16] Ahmed, M.; Mahmood, A. N.; Hu, J.: A survey of network anomaly detection techniques. Journal of Network and Computer Applications 60/, S. 19–31, 2016.

[Bu20] Bundesnetzagentur: Marktstammdatenregister, 2020.

[Bu21] Burger, B.: Burger, Bruno: Nettostromerzeugnis in Deutschland im Jahr 2020, 2021.

[Ch20] Chen, H.; Ma, H.; Chu, X.; Xue, D.: Anomaly detection and critical attributes identification for products with multiple operating conditions based on isolation forest. Advanced Engineering Informatics 46/, S. 101139, 2020.

[De21] Deutscher Wetterdienst: Deutscher Wetterdienst: Climate Data Center, 2021.

[Ga17] Garoudja, E.; Harrou, F.; Sun, Y.; Kara, K.; Chouder, A.; Silvestre, S.: Statistical fault detection in photovoltaic systems. Solar Energy 150/, S. 485–499, 2017.

[HH16] te Heesen, H.; Herbort, V.: Development of an algorithm to analyze the yield of photovoltaic systems. Renewable Energy 87/, 2016.

[HHR18] te Heesen, H.; Herbort, V.; Rumpler, M.: Untersuchung des Ertrags von Photovoltaikdachanlagen bis 30 kWp in Deutschland im Zeitraum 2014 bis 2017. In: Workshops der INFORMATIK 2018-Architekturen, Prozesse, Sicherheit und Nachhaltigkeit. Köllen Druck+ Verlag GmbH, 2018.

[HHR19] te Heesen, H.; Herbort, V.; Rumpler, M.: Performance of roof-top PV systems in Germany from 2012 to 2018. Solar Energy 194/June 2021, S. 128–135, Dez. 2019.

[JN19] John, H.; Naaz, S.: Credit card fraud detection using local outlier factor and isolation forest. Int. J. Comput. Sci. Eng. 7/4, S. 1060–1064, 2019.

[Le20] Leloux, J.; Narvarte, L.; Desportes, A.; Trebosc, D.: Performance to Peers (P2P): A benchmark approach to fault detections applied to photovoltaic system fleets. Solar Energy 202/, S. 522–539, 2020.

[LTZ08] Liu, F. T.; Ting, K. M.; Zhou, Z.-H.: Isolation forest. In: 2008 eighth ieee international conference on data mining. IEEE, S. 413–422, 2008.

[SH21] Schardt, J.; te Heesen, H.: Performance of roof-top PV systems in Germany from 2012 to 2019. Solar Energy 217/June 2021, S. 235–244, 2021.

[So21] Solar Datensysteme GmbH: Solar Datensysteme GmbH: Energy Management System Solar-Log, 2021.

[St21] Steenwinckel, B.; De Paepe, D.; Hautte, S. V.; Heyvaert, P.; Bentefrit, M.; Moens, P.; Dimou, A.; Van Den Bossche, B.; De Turck, F.; Van Hoecke, S. et al.: FLAGS: A methodology for adaptive anomaly detection and root cause analysis on sensor data streams by fusing expert knowledge with machine learning. Future Generation Computer Systems 116/, S. 30–48, 2021.