Optimierung der Anomalieerkennung in Brauereidaten mithilfe von Mehrheitsabstimmung

1st Nguyen Anh Nguyen
Technische Hochschule Ostwestfalen-Lippe
Lemgo, Deutschland
nguyen.nguyen@stud.th-owl.de

2nd Eduard Berzeminskij
Technische Hochschule Ostwestfalen-Lippe
Lemgo, Deutschland
eduard.berzeminskij@stud.th-owl.de

***Abstract*—Anomalieerkennung bei Brauereidaten zielt darauf ab, potentielle abnormale Verhaltensweisen zu erfassen, indem die geloggten Daten über einen bestimmten Zeitraum beobachtet werden. In dieser Arbeit schlagen wir eine maschinelle Lernmethode basierend auf Mehrheistimmung vor, um eine bestimmte Anomalie automatisiert erkennen zu werden.**

***Index Terms*—Anomalieerkennung, Brauereidaten, Mehrheitsabstimmung, LSTM**

I. Einleitung (Nguyen Anh Nguyen, Eduard Berzeminskij)

Anomalieerkennung ist eine wichtige entscheidende Aufgabe in vielen Bereichen wie Cybersicherheit, Betrugserkennung [5] und wird intensiv erforscht. Anomalien sind Abweichungen vom Erwarteten wie zum Beispiel ein negativer Messwert in einer Reihe von sonst positiven Messwerten. Hier untersuchen wir verschiedene Messwerte einer Hochautomatisierte Filterung von Bier-Hefe-Suspensionen um manuelle, vom Menschen ausgeführte Proben und deren Auswirkungen in den Daten, von den automatischen Anpassungen der Maschine, zu unterscheiden mithilfe verschiedener Machine Learning Methoden. Dabei beschreiben wir hier die Anomalie als eine abgenommene Probe in der Filtrierungsanlage.

In diesem Artikel stellen wir zwei Ansätze zur Erkennung von Anomalien vor, einer der auf Mehrheitsabstimmung und ein long- and short-term memory (LSTM) Netzwerk basiert. Mehrheitsabstimmung ist eine einfache, aber effektive Technik, um durch die Verwendung mehrerer Algorithmen optimale Ergebnisse zu erzielen. Wir wenden diese Technik auf die Aufgabe der Anomalieerkennung an, indem wir die Vorhersagen mehrerer individueller Modelle aggregieren und die Mehrheitsvorhersage als endgültige Entscheidung verwenden.

II. Daten (Eduard Berzeminskij)

Wir haben die Daten von der Filtrierungsanlage in 15 Ordner erhalten, welche nach dem Datum der Messung benannt sind. In den jeweiligen Ordner, sind immer folgende Dateien dabei: *measurements.csv*, *final.csv* & gefolgt von 0-3 *filt_X.csv* Dateien. *measurements.csv* beinhaltet Zeiten, an den Proben aus der Maschine entnommen wurde. *final.csv* beinhaltet alle Sensordaten des Tages. *filt_X.csv* sind Teilmengen von *final.csv*, in denen man die Probenentnahme wieder erkennen kann und diese wurden daher intensiver betrachtet. Der Datensatz besteht (meistens) aus 23 Kanälen, worunter Zeiten, Temperaturen, Druckmessungen, Stromstärken und anderem zu finden ist. Zunächst veranschaulichen wir hier eine durch eine Probenentnahme verursachte Anomalie der Maschine mithilfe der Druckwerte. Man erkennt einen Druckabfall (siehe Abbildung 1) gefolgt von einer Steigung des Drucks in der Maschine. Dabei liegt die Schwierigkeit darin, dass die Maschine am Anfang eines Filtrierungsvorgangs sich erst ”einschwingen” muss und am Ende den Druck anpasst. Dabei gibt es einen Tag, an den gewollt die Druckeinstellung innerhalb der Maschine abfällt, aber keine Probe entnommen wurde. Diesen Fall möchten wir nicht als Anomalie erkennen. Die Daten in ihrer originalen Form sind ungelabled, um supervised Methoden zu ermöglichen wurden die manuell gelabelt. Dabei orientierte man sich lose an der Datei measurements.csv und durch wiederholten Plotten der Daten wurde jeder Druckfabfall, der durch eine Probe verursacht wurde gelabelt.

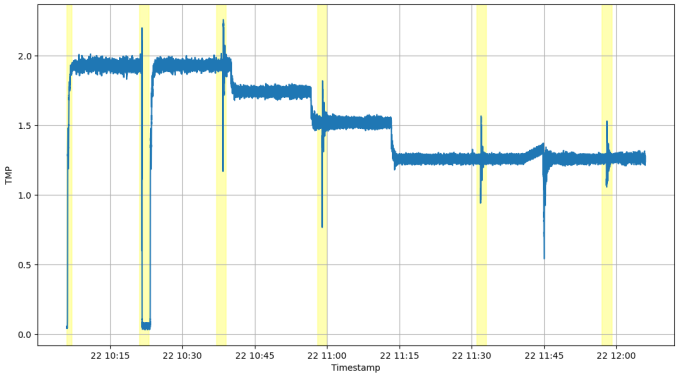

Abb. 1. Beispiel vom Druckabfall der Datei filt_1.csv am 22.06.2022
In Gelb: Zu erkennende Anomalie In Rot: Sonstiger Druckabfall

III. Methoden

A. Mehrheitabstimmung - Majority Voting (Nguyen Anh)

Die Mehrheitsabstimmung bei der Anomalieerkennung ist eine Methode, bei der mehrere Algorithmen verwendet werden, um Datenpunkte als anomal oder normal zu klassifizieren. In diesem Fall verwenden wir FB Prophet, die Drei-Sigma-Regel, den einfachen gleitenden Durchschnitt (Simple Moving Average - SMA) und exponentiellen gleitenden Durchschnitt (Exponential Moving Average

- EMA). FB Prophet ist ein Vorhersagemodell, das auf Zeitreihen basiert und das Verhalten von Daten in der Vergangenheit analysiert, um zukünftige Werte vorherzusagen. Die Drei-Sigma-Regel ist ein statistisches Konzept, bei dem Datenpunkte, die mehr als 2 oder 3 Standardabweichungen von dem Durchschnitt entfernt sind, als anomal betrachtet werden. Der SMA ist ein gleitender Durchschnitt, bei dem der aktuelle Wert durch den Durchschnitt der letzten n Werte ersetzt wird. Der EMA ist ähnlich, aber es wird mehr Gewicht auf die jüngsten Werte gelegt. Wenn mehr als 3 Algorithmen einen Datenpunkt als anomal klassifizieren, wird er als solcher betrachtet. Auf diese Weise können wir die Genauigkeit der Anomalieerkennung erhöhen, indem wir mehrere Ansätze kombinieren.

Drei-Sigma-Regel:
Die Drei-Sigma-Regel ist eine statistische Regel, die besagt, dass in einer normalverteilten Stichprobe mit hoher Wahrscheinlichkeit (siehe Abbildung 2) die wahren Werte einer Population innerhalb von drei Standardabweichungen der Stichprobenmittelwerte liegen [1]. Wenn beispielsweise die Messwerte eines Prozesses normalverteilt sind und die Stichprobenmittelwerte innerhalb von drei Standardabweichungen des Mittelwerts der Population liegen, können wir davon ausgehen, dass der Prozess innerhalb der gewünschten Toleranzen arbeitet. Wenn ein Messwert jedoch außerhalb dieses Bereichs liegt, könnte dies auf ein Problem mit dem Prozess hinweisen und weitere Untersuchungen erforderlich machen. Um die Drei-Sigma-

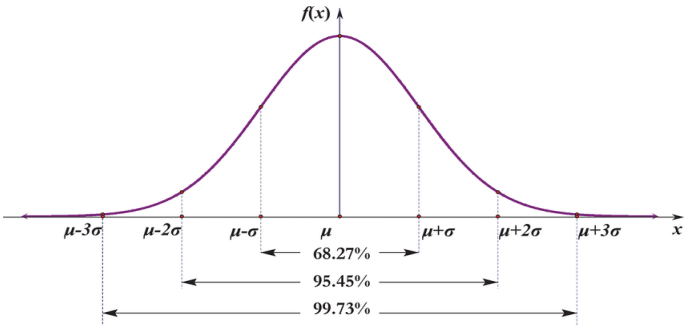

Abb. 2. Die Konfidenzintervalle entsprechen der 3-Sigma-Regel der Normalverteilung [4].

Regel anzuwenden, muss zunächst μ das arithmetische Mittel und σ die Standardabweichung der Daten berechnet werden.

$$\sigma = \sqrt{\frac{1}{n}\left[(a_1-\mu)^2 + (a_2-\mu)^2 + ... + (a_n-\mu)^2\right]} \quad (1)$$
$$= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(a_i-\mu)^2}$$

Danach berechnen wir die Schwellenwerte, um die Maximal- und Minimalwerte festgelegt zu werden.

$$\begin{aligned} \mathbf{Min_{3\sigma}} &= \mu - 3\sigma \\ \mathbf{Max_{3\sigma}} &= \mu + 3\sigma \end{aligned} \quad (2)$$

Alle Datenpunkte, die den Schwellenwert überschreiten, werden als Ausreißer oder Anomalie betrachtet.

Simple Moving Average - SMA:
Der Simple Moving Average (SMA) ist eine Technik zur Glättung von Zeitreihen, die häufig in der Finanzindustrie und anderen Bereichen verwendet wird, in denen es darum geht, Trends und Mustererkennung in Daten zu identifizieren. Der SMA ist einfach anzuwenden und eignet sich besonders für die Analyse von Daten, die keine saisonalen Muster aufweisen. Er ist jedoch weniger geeignet für die Analyse von Daten mit saisonalen Mustern, da in diesem Fall die Glättung kurzfristige saisonale Schwankungen unterdrücken kann. Der SMA wird berechnet, indem man alle Datenpunkte eines bestimmten Zeitraums (z.B. 3 Minuten, 3 Tage) addiert und durch die Anzahl der Datenpunkte n teilt [2]. Der resultierende Wert gibt den Durchschnittswert für den Zeitraum an.

$$\mathbf{SMA} = \frac{P_T + P_{T-1} + ... + P_{T-(n-1)}}{n} \quad (3)$$

P_T steht für den Datenpunktwert zum Zeitpunkt T. Die Länge der SMA, also die Anzahl der verwendeten Datenpunkte, kann angepasst werden, um unterschiedliche Glättungseffekte zu erzielen. Ein längerer SMA wird die Daten stärker glätten und kurzfristige Schwankungen unterdrücken, während ein kürzerer SMA die Daten weniger glätten und mehr kurzfristige Schwankungen zeigen wird. Anschließend werden die SMA-Bänder berechnet, um Anomalien in Zeitreihendaten zu identifizieren. Sie werden erstellt, indem man zwei Linien über einem SMA-Diagramm zeichnet, die die sich n Standardabweichungen von dem SMA entfernen. Die Formel für die SMA-Bänder lautet wie folgt:

$$\begin{aligned} LowerBand &= SMA - (n \cdot \sigma) \\ UpperBand &= SMA + (n \cdot \sigma) \end{aligned} \quad (4)$$

Wenn ein Datenpunkt außerhalb des SMA-Bands liegt, wird er als anomal betrachtet. Dies kann darauf hinweisen, dass sich der Wert des Datenpunkts signifikant von dem Durchschnittswert des Zeitraums unterscheidet und daher ungewöhnlich ist.

Exponential Moving Average - EMA:
Der Exponential Moving Average (EMA) ist auch ein statistischer Indikator, der verwendet wird, um den Durchschnittswert einer Zeitreihe über einen bestimmten Zeitraum zu berechnen [2]. Der EMA hat im Vergleich zum SMA einige Vorteile. Zum einen reagiert der EMA schneller auf Veränderungen in den Daten, da er stärker von den jüngsten Datenpunkten abhängt. Zum anderen kann der EMA zur Glättung von saisonalen Mustern in den Daten verwendet werden, da er jedem Datenpunkt ein individuelles Gewicht zuweist. Der EMA wird berechnet, indem man x_t den aktuellen Datenpunkt mit α dem bestimmten Gewicht multipliziert und dann mit y_{t-1} dem vorherigen Gleitenden Durchschnitt verrechnet. Dieses Gewicht, auch

als Glättungsfaktor bezeichnet, wird so festgelegt, dass die jüngsten Datenpunkte ein höheres Gewicht haben als die älteren Datenpunkte. Auf diese Weise werden die jüngsten Datenpunkte stärker in die Berechnung des gleitenden Durchschnitts einbezogen und die Glättung wird sich schneller an Veränderungen in den Daten anpassen.

$$\mathbf{EMA(t)} = \begin{cases} y_t = x_t & t = 1 \\ y_t = \alpha(x_t) + (1-\alpha)y_{t-1} & \text{t} > 1 \end{cases} \quad (5)$$

Überlicherweise wird α mit folgenden Formel berechnet:

$$\alpha = \frac{2}{(n+1)} \quad (6)$$

Danach werden die EMA-Bänder berechnet, um Anomalien zu erkennen. Sie werden auch wie bei SMA-Bändern erstellt, indem man zwei Linien über einem EMA-Diagramm zeichnet, die sich n Standardabweichungen von dem EMA entfernen. Diese Linien bilden dann einen Bereich, der als EMA-Band bezeichnet wird.

$$\begin{aligned} LowerBand &= EMA - (n \cdot \sigma) \\ UpperBand &= EMA + (n \cdot \sigma) \end{aligned} \quad (7)$$

Wenn ein Datenpunkt außerhalb des EMA-Bands liegt, wird er als anomal betrachtet.

Fb Prophet:
Fb Prophet [3] ist eine Open-Source-Bibliothek, die von Facebook entwickelt wurde und zur Vorhersage von Zeitreihendaten verwendet wird. Es ist speziell für die Vorhersage von zukünftigen Werten von Zeitreihen geeignet, die saisonale Muster aufweisen, wie z.B. Verkaufszahlen oder Wetterdaten. Es berücksichtigt auch Fehler und Trends in den Daten und kann automatisch Anomalien erkennen. Das Modell von FB Prophet besteht aus drei Hauptkomponenten $g(t)$ Trend, $s(t)$ Saisonalität und $h(t)$ Feiertage. Die Formel lautet wie folgt:

$$\mathbf{y(t)} = g(t) + s(t) + h(t) + \epsilon(t) \quad (8)$$

$y(t)$ steht für die Trendfunktion, die nicht periodische Änderungen im Wert der Zeitreihendaten modelliert und $\epsilon(t)$ für den Fehler, der als letzte Komponente in der Formel hinzugefügt wird, um die Unsicherheit im Modell abzubilden, die durch unvorhergesehene Ereignisse oder ungenauen Daten entstehen kann. Für die Anomalieerkennung berechnet FB Prophet das Verhältnis von prognostiziertem Wert zu Beobachtungswert und markiert Abweichungen von einem Schwellenwert als Anomalie.

B. LSTM (EB)

Hier benutze ich ein long- and short-term memory (LSTM) Netzwerk, welche zu den recurrent neural networks (RNNs) gehört. Diese Art von Modellen schienen mir plausibel, da innerhalb eines RNN der Output eines Zeitpunkts abhängig ist von einigen vorherigen Zeitpunkten. Jedoch leiden RNNs am vanishing und exploding gradient problem, wobei ein Netzwerk entweder nicht mehr dazu lernt oder mit einer erhörten Lernrate nie zu einen minimalen Lernfehler konvergiert. LSTM Netzwerke umgehen das Problem in dem das Model irrelevante Information systematisch vergisst. Somit eignet diese sich für das Modellieren aufeinander folgenden Daten wie unsere Zeitreihe. [6]

Da hier an dem Vorgehen des LSTM-Netzwerks nichts angepasst wurde, verweise ich auf weitere Literatur für das Verständnis. [7] Im Gegensatz zu anderen LSTM-Forecasting Ansätze wird hier versucht, durch das Auslassen gelabelter Anomalien einen Rekonstruierungfehler zu maximalisieren. Der Gedanke dabei ist, das unser Modell nicht ausreichend genau den Druckabfall vorhersagen kann, da dieser nie gelernt wurde. Die resultierende Differenz zwischen dem Vorhergesagten und den realen Daten sollten unsere Anomalien markieren.

Für das Trainieren des LSTM-Modells wurden die Daten vorverarbeitet, indem die gelabelten Anomalien lückenlos ausgelassen wurden und die Daten mit dem Standard Skalierer skaliert wurden. Als Input bekommt das Modell ein drei Dimensionales Matrix "*TrainX*" und zwar ein Zeitfenster *n*, die Datenkanäle und die Anzahl der Datenpunkte minus *n*. Die ein dimensionale Matrix "*TrainY*" beinhaltet nur die vorherzusagende Datensätze. "*TrainX*" wird erstellt, indem man *n* Zeitpunkte aus den Daten kopiert und iterierend *(N × Datenkanäle)* Matrizen zu einer Matrix hinzufügt. "*TrainY*" hat im Gegensatz nur die "*Pressure retentate IN5*" Werte, die nach dem *n* großen Zeitfenster erscheinen und vorgesagt werden sollen. Aus allen gelabelten Daten hat man nun die "*TrainX*" Matrix und die "*TrainY*" Matrix zusammengeführt, zu einer allen gelabelten Daten übergreifende "*TrainX*" und "*TrainY*". Die wurden gemischt, damit die zeitliche Reihenfolge eine kleinere Rolle spielt.

IV. Ergebnisse

A. Mehrheitabstimmung - Majority Voting (Nguyen Anh)

Zuerst müssen wir die Ergebnisse von jeder Methode anschauen, um ein optimales Resultat zu bekommen. Danach sammeln wir die Ergebnisse und zählen die Anzahl der Anomalien, die von jeder Methode erkannt wurden. Wir analysieren den Kanal *"Pressure retentate IN5"*, der den Druckwert darstellt.

Fb Prophet:
Wir implementieren Prophet mit den Standardwerten, aber die saisonale Komponenten werden deaktiviert, weil die aufgezeichneten Daten nicht täglich sind. Da Prophet optimiert wurde, wenden wir es automatisch an. Prophet zeigt uns einen Trend und seine Schwellenwerte. Ein Datenpunkt wird als anomal betrachtet, wenn er außerhalb der Schwellenwerte liegt. Abhängig von der Komplexität der aufgezeichneten Daten kann das Modell optimale (siehe Abbildung 3) oder suboptimale (siehe Abbildung 4) Ergebnisse liefern. Da wir jedoch die Mehrheitabstimmung anwenden und dieses Modell mit Standardwerten verwendet wird, akzeptieren wir die

Ergebnisse, die das Modell liefert. Die Anomalien nach der

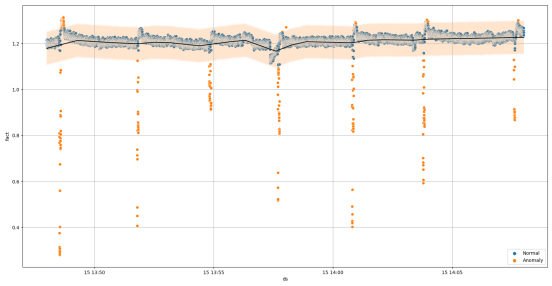

Abb. 3. Anomalien wurden von Datei filt_2.csv am 15.06.2022 durch der Methode FB Prophet erkannt

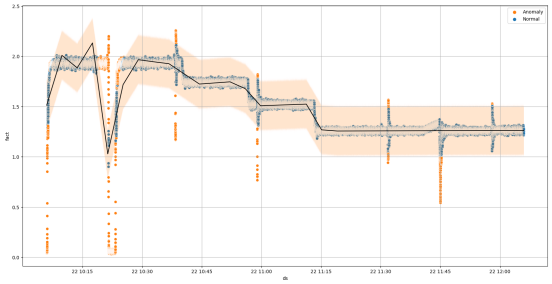

Abb. 4. Anomalien wurden von Datei filt_1.csv am 22.06.2022 durch der Methode FB Prophet erkannt

Impelementierung werden gespeichert und wir wenden die weitere Methoden an.

Drei-Sigma-Regel:
In diesem Fall werden Datenpunkte, die mehr als Standardabweichungen von 2 des Durchschnitts entfernt sind, als anomal betrachtet. Hier wenden wir die Mehrheitabstimmung an, deswegen ist es sinnvol, wenn jedes Modell von Mehrheitabstimmung empfindlicher für anomale Datenpunkte ist. In Tabelle I können wir eindeutig den Unterschied der Anzahl von Anomalien erkennen. Die Anzahl der Anomalien von 2-Sigma ist von 15% bis 93% höher als die von 3-Sigma. Es deutet darauf hin, das Modell ist empfindlicher für Anomalien, wenn wir bedenken, dass 5% des Datensatzes Anomalien sind. Die Ergebnisse von diesem Modell werden gespeichert und später mit den anderen Ergebnissen ausgewertet.

Table I. Anzahl der Anomalien nach der Implementierung 2-Sigma und 3-Sigma

Datei	2-Sigma	3-Sigma	% Untersch. in der Anzahl
2022_05_20_filt1	2203	1139	0,93
2022_05_20_filt2	615	418	0,48
2022_06_15_filt2	195	170	0,15
2022_06_15_filt3	627	401	0,56

Simple Moving Average - SMA:
Für den SMA müssen wir einen bestimmten Zeitraum definieren und dazu vergleichen wir die Anzahl der Anomalien über 30 Sekunden, 1 Minute, 2 Minuten, 3 Minuten bzw. 300, 600, 1200 und 1800 Datenpunkte. Wir haben ein Experiment mit der Datei filt_1.csv am 22.06.2022 und beobachten die Anzahl der Anomalien. Aus Tabelle II ist ersichtlich, dass wir umso mehr Anomalien erkennen können, je länger wir den Zeitraum definieren. Außerdem stecken wir die SMA-Bänder

Table II. Die Anzahl der Anomalien von Datei filt_1.csv am 22.06.2022 durch SMA in angegebenen Zeiträumen

2022_06_22_filt_1	
Datenpunkte	**Anzahl der Anomalien**
300	457
600	726
1200	888
1800	1177

für die Daten mit der Standardabweichung von 3 ab. Die Standardabweichung von 3 anstelle von 2 kann tatsächlich dazu beitragen, dass es weniger falsch positive Ergebnisse gibt, da die Anforderungen an Abweichungen von den normalen Daten höher sind (siehe die Tabelle III). Nach der

Table III. Die falsch positive Ergebnisse von 4 Daten, wenn die SMA-Bänder mit der Standabweichung von 2 und 3 eingestellt sind

Datei	Falsch positive Ergebnisse	
	2-Sigma	3-Sigma
2022_05_20_filt_1	17	9
2022_05_20_filt_2	16	7
2022_06_15_filt_2	4	0
2022_06_15_filt_3	5	2

Bestimmung des Zeitraums und der SMA-Bänder erhalten wir das Ergebnis wie in der Abbildung 5 dargestellt. Ein Nachteil des SMA ist jedoch, dass es nicht in der Lage ist, Anomalien zu erkennen, die in den ersten Zeitperioden auftreten. Dies liegt daran, dass der SMA erst nach einigen Zeitperioden berechnet werden kann und daher keine Informationen über die Anfangsdaten hat, um Anomalien zu erkennen.

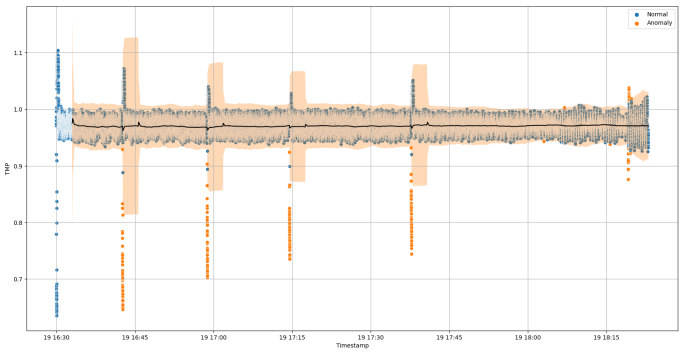

Abb. 5. Anomalien von Datei filt_3.csv am 19.07.2022 durch der Methode SMA

Exponential Moving Average - EMA:
Ähnlich wie bei SMA wählen wir einen Zeittraum von 3 Minuten bzw. 1800 Datenpunkte und die EMA-Bänder sind auch die Standardabweichung von 3. Der EMA hat auch das Problem bei der Erkennung von Anomalien in den ersten Zeitperioden. Dies liegt daran, dass der EMA ähnlich wie

der SMA, eine gewichtete Mittelwertsmethode ist, die darauf abzielt, den durchschnittlichen Wert einer Zeitreihe von Daten über einen bestimmten Zeitraum hinweg zu berechnen. Der EMA gibt jedoch mehr Gewicht auf die jüngsten Daten und verringert das Gewicht auf ältere Daten, was ihm ermöglicht schneller auf Veränderungen in den Daten zu reagieren (siehe Abbildung 6). Die anomale Datenpunkten werden auch gespeichert und mit den anderen Ergebnissen ausgewertet.

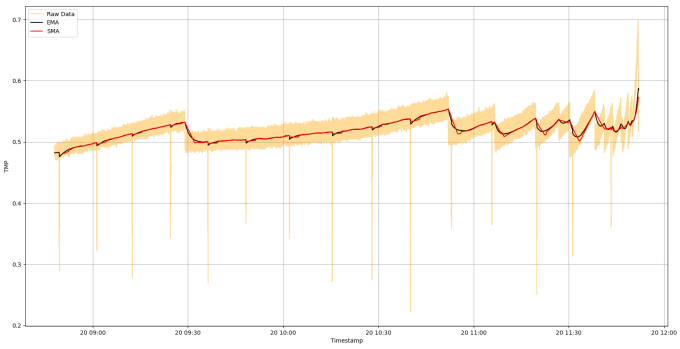

Abb. 6. EMA und SMA für die Datei filt_1.csv am 20.05.2022. Die EMA-Linie reagiert schneller mit der Veränderung in der Datei als die SMA-Linie

Mehrheitabstimmung:
Nach der Anwendung von vier Modellen sammeln wir die Ergebnisse und zählen die Anzahl der Anomalien, die von jeder Methode erkannt wurden. Danach werden der Schwellenwert für die Mehrheitsabstimmung auf 3/4 festgelegt. Das bedeutet, dass mindestens drei der vier Algorithmen eine Anomalie erkennen müssen, damit die Zeitreihe als anomal klassifiziert wird. In Tabelle IV können wir den Unterschied zwischen 2 und 3 Abstimmungen sehen. Mit 2 Abstimmungen werden alle anomalen Datenpunkte erkannt (Recall-Score ist 1,0), aber das Modell liefert auch mehr falsch positiven Ergebnissen. Mit 3 Abstimmungen ist das Modell stabiler, wenn der F1-Score immer mehr als 0,8. Hier gibt es nur ein Problem, dass die Anomalien in den ersten Zeitperioden nicht erkannt werden, weil SMA und EMA diese Anomalie in diesen Zeitperioden nicht erkannt werden können. Dennoch kann das Modell Anomalien gut erkennen und die falsch positive Ergebnisse einschränken.

Table IV. Genauigkeit des Modells mit 2 und 3 Abstimmungen bei 5 Daten. Der F1-Score mit 3 Abstimmungen ist besser als mit 2 Abstimmungen und das Modell hat weniger falsch positive Ergbenisse.

Datei	ab 3 Abstimmungen			ab 2 Abstimmungen		
	F1-Score	Recall	Precision	F1-Score	Recall	Precision
2022_05_20_filt_1	0,882	0,937	0,833	0,653	1,0	0,484
2022_05_20_filt_2	0,937	0,937	0,937	0,842	1,0	0,727
2022_06_15_filt_2	0,923	0,857	1,0	1,0	1,0	1,0
2022_06_15_filt_3	0,9	0,9	0,9	0,909	1,0	0,83
2022_06_29_filt_2	0,941	0,888	1,0	0,6	1,0	0,428

Ingesamt kann das Modell Anomalien gut erkennen, der F1-Score liegt bei am häufigsten 0,7 oder höher. Einige Daten, die zu komplex oder nichtnormalverteilte sind, haben den eher niedrigen F1-Score (siehe Tabelle V).

Table V. Der F1-Score des Modells "Mehrheitabstimmung" bei 20 Daten

Datei	F1-Score
2022_05_20_filt_1	0,882
2022_05_20_filt_2	0,937
2022_06_01_filt_1	0,558
2022_06_15_filt_1	1,0
2022_06_15_filt_2	0,923
2022_06_15_filt_3	0,9
2022_06_22_filt_1	0,769
2022_06_24_filt_1	0,842
2022_06_29_filt_1	0,714
2022_06_29_filt_2	0,941
2022_07_13_filt_1	0,666
2022_07_13_filt_2	0,5
2022_07_13_filt_3	0,25
2022_07_19_filt_1	0,714
2022_07_19_filt_2	0,727
2022_07_19_filt_3	0,8
2022_07_27_filt_1	0,8
2022_07_27_filt_2	0,8
2022_08_03_filt_1	0,857
2022_08_03_filt_2	0,833

B. LSTM (EB)

Zunächst haben wir nur eine Reihe von hervorgesagten Druckwerten, dessen Differenz zu den realen Daten wir berechnen müssen. Vergleichen wir die Differenz zu den gelabelten Anomalien, erkennen wir im besten Fall größere Differenzen um unsere gelabelten Anomalien herum.

Hierbei nutzen wir die oben genannten Methoden, um die Anomalien innerhalb der Differenz zu erkennen. Und gleichen diese Anomalien der Differenz mit unseren gelabelten Anomalien ab, um eine Bewertung der Ergebnisse zu machen. In einem LSTM Modell, welches ein fünftel Sekunde als Zeitfenster für das Trainieren als auch die Vorhersage hat, sind die Ergebnisse wie folgt:

Table VI. Genauigkeit des Modells bei 5 Daten

Datei	ab 3 Abstimmungen		
	F1-Score	Recall	Precision
2022_05_20_filt_1	0,517	0,937	0,357
2022_06_01_filt_1	0,298	0,928	0,178
2022_06_15_filt_1	0,777	1,0	0,636
2022_06_22_filt_1	0,333	0,857	0,206
2022_06_24_filt_1	0,285	0,888	0,17

V. Fazit

A. Mehrheitabstimmung - Majority Voting (Nguyen Anh)

Mehrheitsabstimmung ist eine Methode der Anomalieerkennung, die verwendet wird, um die Ergebnisse mehrerer Anomalieerkennungsalgorithmen zusammenzuführen und zu einem einzigen Ergebnis zu konsolidieren. Es kann verwendet werden, um die Genauigkeit der Anomalieerkennung zu verbessern, indem es die Vorteile mehrerer Methoden kombiniert. Ein wichtiger Faktor bei der Anwendung der Mehrheitsabstimmung ist die Schwellenwert-Einstellung, die von der Anzahl der verwendeten Algorithmen abhängt. Je mehr Algorithmen verwendet werden, desto höher ist der Schwellenwert, um sicherzustellen, dass die Mehrheit der

Algorithmen eine Anomalie erkennt, bevor die Zeitreihe als anomal markiert wird. In diesem Artikel mit Brauereidaten haben wir die Anomalien vom Druckwert recherchiert und ein Modell basierend auf vier Algorithmen Fb Prophet, Drei-Sigma-Regel, SMA und EMA erstellt. Als Ergebnis erhalten wir 16 von 20 Daten mit F1-Score größer als 0,7.

B. LSTM (EB)

Allgemein bieten sich LSTMs für die Erkennung von Anomalien an, jedoch in dieser Ausführung nicht. Was man in den schlechten Ergebnissen wieder erkennen kann. Dies liegt an einer sehr hohen Anzahl an falschen positiven.

VI. Diskussion

A. LSTM (EB)

Die Auswahl der Fenstergrößen für das Modell war unzureichend überlegt, Angesicht der Tatsache, das diese in diesem Modell eine große rolle spielt. [8] Sowohl hätte man auch die Daten glätten können, um das Rauschen in den Daten zu reduzieren. Oder die Frequenz der Daten reduzieren können, um Rechenleistung sparend eine größere Zeitspanne zu betrachten.

References

[1] Pukelsheim, Friedrich. "The three sigma rule." The American Statistician 48.2 (1994): 88-91.

[2] Hansun, Seng. "A new approach of moving average method in time series analysis." 2013 conference on new media studies (CoNMedia). IEEE, 2013.

[3] Taylor, Sean J., and Benjamin Letham. "Forecasting at scale." The American Statistician 72.1 (2018): 37-45.

[4] Wang, Bin, Wenzhong Shi, and Zelang Miao. "Confidence analysis of standard deviational ellipse and its extension into higher dimensional Euclidean space." PloS one 10.3 (2015): e0118537.

[5] Yao, Danfeng, et al. "Anomaly detection as a service: challenges, advances, and opportunities." Synthesis Lectures on Information Security, Privacy, and Trust 9.3 (2017): 1-173.

[6] Géron, Aurélien. Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow. " O'Reilly Media, Inc.", 2022.

[7] Siami-Namini, Sima, Neda Tavakoli, and Akbar Siami Namin. "The performance of LSTM and BiLSTM in forecasting time series." 2019 IEEE International Conference on Big Data (Big Data). IEEE, 2019.

[8] Homayouni, Hajar, et al. "An autocorrelation-based lstm-autoencoder for anomaly detection on time-series data." 2020 IEEE International Conference on Big Data (Big Data). IEEE, 2020.