Projekt Name

Explainable AI for Automated Production Systems (XAPS)

Typ

Abschlussbericht

Datum

27.06.2023

Förderkennzeichen

01IS19084A-F

Autoren

Marco Ehl, Malte Hellmeier, Haydar Qarawlus, Chris Scharpenberg, Prof. Dr. Steffen Staab, Rodrigo Lopez Portillo Alcocer, Prof. Dr. Jan Jürjens, Alexander Sudhoff, Mahmood Al-Doori

Zuwendungsgeber

Das diesem Bericht zugrunde liegende Vorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen 01IS19084A-F gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei den Autoren.

GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

Eckdaten

Das im Rahmen der Fördermaßnahme *Erklärbarkeit und Transparenz des Maschinellen Lernens und der Künstlichen Intelligenz* (2019) mit einer Fördersumme von 1.197.498 € geförderte Projekt ist Teil der Umsetzung der KI-Strategie der Bundesregierung und der Hightech-Strategie 2025. Projektzeitraum: 01.01.2020 - 31.12.2022

Zuwendungsempfänger

Fraunhofer ISST
Universität Koblenz-Landau
Universität Stuttgart
Old World Computing GmbH
HELLA GmbH & Co. KGaA

Inhaltsverzeichnis

Abbildungs- und Tabellenverzeichnis

I. Kurzbericht

Die Herausforderung

Ziel des Projektes XAPS ist es, mit Methoden des maschinellen Lernens komplexe Abhängigkeiten zwischen Fehlern und Bearbeitung frühzeitig zu entdecken und diese mit Methoden der Künstlichen Intelligenz (KI) so zu erklären, dass die Betreibenden das automatisierte Produktionssystem effizient und effektiv optimieren können. Dazu verknüpft XAPS digitale Beschreibungen der Fabrik und des Produkts, den digitalen Zwilling, aus dem Manufacturing Execution System mit maschinellem Lernen und innovativen Methoden der erklärbaren Künstlichen Intelligenz. Die XAPS-Plattform bietet somit die Grundlage für ein ganzheitliches Lösungsportfolio zur Steuerung und Überwachung von automatisierten Produktionssystemen.

Anknüpfungspunkte aus Wissenschaft und Technik

Das Projekt hat an vorhandene Ergebnisse aus Wissenschaft und Technik angeknüpft. Dazu gehörte die Erstellung einer Architekturdokumentation auf Basis des freien Industrie- und Forschungsstandards *arc42* (https://www.arc42.de/). In der Anforderungserhebung wurde bei der Ausgestaltung der Workshops auf vorhandene und bewährte Methoden des Requirements Engineerings zurückgegriffen. Im Bereich der Entwicklungsleistung wurde auf die vorhandene *RapidMiner* Plattform (https://rapidminer.com/) aufgebaut. Die Verfahren in der Operatorenentwicklung basieren auf Inhalten aktueller Forschungspublikationen, wie beispielhaft die der *Local Interpretable Model-Agnostic Explanation* (LIME - https://doi.org/10.1145/2939672.2939778) für die resultierende Explanation Engine. Das Buch *Entwurfsmuster: Elemente wiederverwendbarer objektorientierter Software* (Gamma, E. 2004) bietet einen Anknüpfungspunkt, da die Arbeit als Standardwerk in der Softwarearchitektur und Designmuster betrachtet wird, was uns erlaubt, auf etablierten und bewährten Prinzipien aufzubauen.

Zusammenarbeit mit anderen Forschungseinrichtungen

Auf nationaler Ebene konnte sich das Projekt-Konsortium mit den relevanten Initiativen und Projekten vernetzen, z.B. der Plattform Industrie 4.0, um weitere Sichtbarkeit zu erzeugen und somit zusätzliche Forschungsprojekte akquirieren zu können. Auf internationaler Ebene ist insbesondere eine Vernetzung mit der International Data Spaces Association (IDSA) interessant, die im Bereich intelligente Produktion hervorragend international ist.
Folgende verwandte, noch laufenden Projekte können als Anknüpfungspunkte dienen, um gesammelte Erkenntnisse weiter zu verwerten: IIP-Ecosphere, Automated Compliance Checks for Construction, Renovation or Demolition Works (ACCORD), Interdisziplinärer Hub zur Vermittlung von Kompetenzen in Entwicklung, Umgang und Anwendung von erklärbaren, vertrauenswürdigen, resilienten und sicheren KI-Verfahren (IH - evrsKI) und Engineering Trustworthy Data-intensive Systems (EnTrust).

Das Ergebnis

Die XAPS-Plattform liefert dem Benutzer reaktiv oder proaktiv intuitiv nutzbare Erklärungen zu Problemen in der Produktion. Die XAPS-Plattform nutzt existierende Konfigurations-, Steuerungs-, Monitoring- und Sensordaten des Manufacturing Execution Systems, um die physikalische Fabrik als digitale Fabrik zu repräsentieren, und setzt sie mit den digitalen Zwillingen der Produkte in Beziehung.

Ablauf des Vorhabens

Die Projektlaufzeit startete im Januar 2020 und endete im Dezember 2022. Inhaltlich wurden die Aufgaben in 5 Arbeitspakete (AP) aufgeteilt, von denen die meisten in Kooperation zwischen mehreren Projektpartnern durchgeführt wurden.
Das Ziel von AP1 war es, wichtige Elemente des Projekts und des Umfeldes zu erfassen und zu beschreiben. Dieses modellbasierte Vorgehen ermöglicht eine klare, visuelle Darstellung des Systems, die zur Verbesserung der Kommunikation und des Verständnisses unter den Projektbeteiligten beiträgt. Es bietet eine Abstraktion, die es ermöglicht, komplexe Systeme zu verstehen.
In AP2 lag ein großer Fokus auf der Datenvorverarbeitung und Erstellung von Modellen des maschinellen Lernens. Diese werden zum einen direkt für die Fehlervorhersage von Teilen in der Produktionsanlage genutzt, und zum anderen als Basis für die Erklärungs-Engine. Für die Erklärungen untersuchten wir Multi-Context Argumentationssysteme, die Potenzial für interpretierbares maschinelles Lernen und Parallelen zur Mechanik neuronaler Netze zeigten. Anschließend verlagerten wir den Schwerpunkt auf modernste modellagnostische Methoden, um eine bessere Anpassungsfähigkeit an verschiedene Modelle zur Ausfallvorhersage zu erreichen.
Schon früh konnten in AP3 Vorarbeiten für die grafische Oberfläche geleistet werden, welche im späteren Verlauf als Basis für den Demonstrator in AP4.3 dienten. Zudem lag in AP3 der Schwerpunkt auf kontextabhängigen und interaktiven Erklärungen, die in der Benutzerschnittstelle die Entscheidungsfindung für die Nutzer vereinfachen. Zudem erlaubt die Untersuchung von einzelnen Vorhersagen und ihren Begründungen ein besseres Verständnis für die Dynamik des Systems.
In AP4 wurden zu Beginn des Projektes Anwendungsfälle analysiert und erhoben. Dieses umfasste mehrere Workshops auf Basis des Requirements Engineerings und des Changemanagements zur Überprüfung der Inhalte, wie Qualitätsziele und Umsetzungen in Teilkomponenten. Auf Basis dieser Anwendungsfälle und Ziele und dem Design einer Benutzerschnittstelle von AP3 wurde in AP4 ein Demonstrator entwickelt, welcher relevante Daten aus dem Betrieb mit den Komponenten von allen Arbeitspaketen vereint: Vorhersagen und Konfidenzen des Fehlervorhersagemodells, Ausgaben der statischen Modellanalyse, sowie Erklärungen und Feedback der Erklärungs-Engine.
In AP5 wurde zum Management und der Verwertung durch zweiwöchentliche digitale Meetings sichergestellt, dass alle Projektpartner stets auf dem neuesten Stand sind, und dass erforderliche Kollaborationen zwischen Partnern koordiniert werden. Weiterhin wurden die Ergebnisse für die Verwertung aufbereitet, zusammengetragen und über eine Website (https://explainable-ai.de/) veröffentlicht.

II. Eingehende Darstellung

1. Erzielte Ergebnisse

Das Hauptergebnis des Projektes ist eine Software, die von Betreibern von automatisierten Produktionssystemen eingesetzt werden kann, um Fehler in der Produktion zu entdecken und zu erklären. Für die Nutzer (durch HELLA vertreten) dieser Software ist das Dashboard die zentrale Komponente für die Informationsgewinnung.
Das Dashboard greift auf weitere Komponenten zurück: Eine statische und dynamische Modellanalyse, ein Fehlervorhersagemodell und eine Erklärungs-Engine.
Diese Komponenten zusammen bieten die Funktionalität, die durch das Dashboard den Nutzern einfach und zentral zugänglich gemacht wird. Bei der Entwicklung der Systemarchitektur wurde auf die Berücksichtigung der ISO25010 Qualitätsziele geachtet. Speziell für die Machine-Learning-Komponenten wurde auf die funktionelle Eignung, Zuverlässigkeit und Wartbarkeit geachtet.
Besonders um die Nutzbarkeit der Software zu gewährleisten, wurden während des Systementwurfs mögliche Anwendungsfälle spezifiziert. Diese Anwendungsfälle wurden bei der Systemevaluation auf ihre Erfüllung überprüft. Eine detaillierte Beschreibung der Qualitätsziele und Anwendungsfälle befindet sich in Abschnitt II.3.

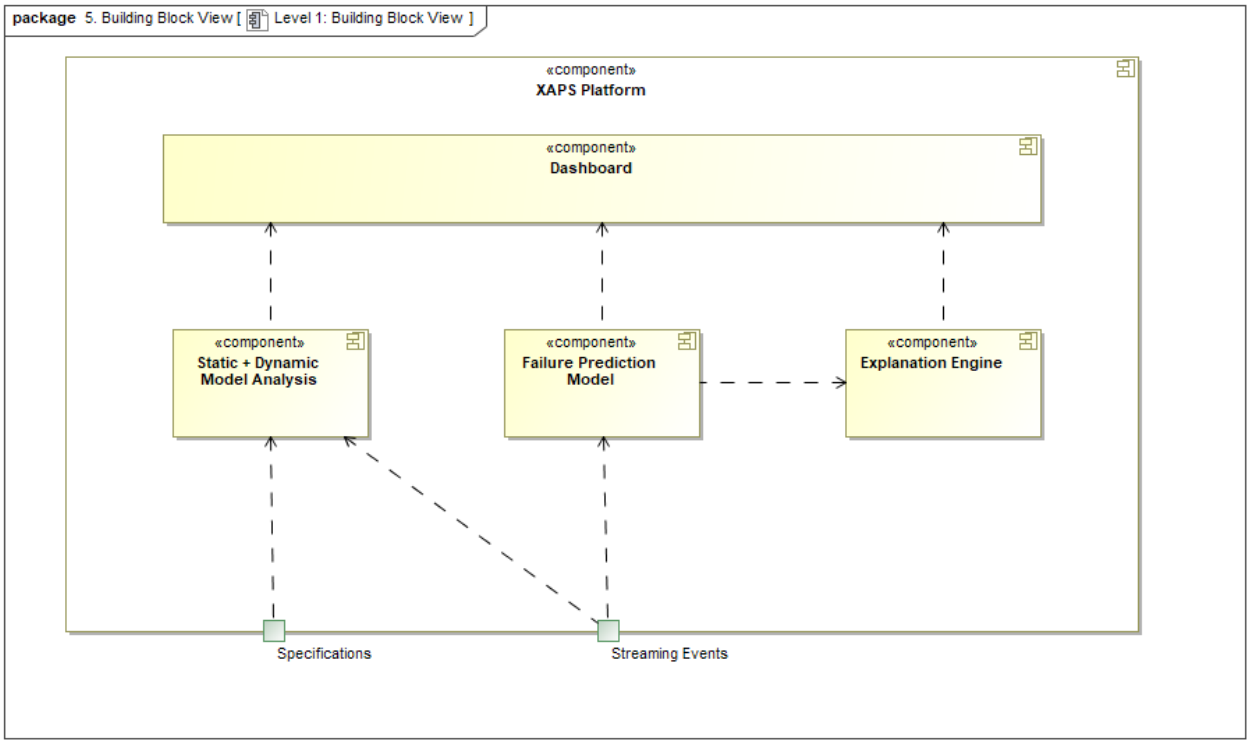

Abbildung 1: Komponentendiagramm der erstellten Software mit ihren Bausteinen

Ergebnis-Bestandteil	Kurzbeschreibung
Statische und dynamische Modellanalyse	Die Universität Koblenz-Landau führte die statische und dynamische Modellanalyse durch. Die Modellanalyse verwendet die Spezifikationen und Streaming-Ereignisse als Eingabedaten und führt basierend auf wiederverwendbaren Mustern Analysen durch. Das Ergebnis der Analysen wird für den Endbenutzer als Bericht angezeigt.
Fehlervorhersagemodell	Old World Computing GmbH (OWC) implementierte das Fehlervorhersagemodell, das die von HELLA bereitgestellten Streaming-Ereignisse verwendet, die Daten für die spätere Nutzung vorverarbeitet, und Modelle des maschinellen Lernens erstellt um bereits nach wenigen Fertigungsschritten die Ausschusswahrscheinlichkeit in späteren Fertigungsschritten zu prognostizieren. Darüber hinaus werden die Ausgaben dieser Komponente von der Erklärungs-Engine verwendet.
Erklärungs-Engine	Um Erklärungen für verschiedene problematische Teile in der Produktionslinie generieren zu können, hat die Universität Stuttgart vier hochmoderne Methoden implementiert, die den Endnutzern relevantes Feedback liefern und es ihnen ermöglichen, die Ergebnisse des Fehlervorhersagemodells zu verstehen und entsprechend zu handeln. Das Fraunhofer ISST nutzt sie, um sie in Operatoren umzuwandeln, die in die RapidMiner-Implementierung von OWC integriert werden können. Die Ergebnisse werden auch an das Dashboard übergeben.
Dashboard	Das von OWC implementierte Dashboard ruft die Ergebnisse von den anderen Komponenten ab und zeigt dem Endbenutzer die Ausgaben der Modellanalyse, des Fehlervorhersagemodells, und der Erklärungs-Engine in Kombination mit den jeweils relevanten Daten der Produktionsanlage in einem lesbaren Format an.

Tabelle 1: Kurzbeschreibungen der Ergebnis-Bestandteile

Statische und dynamische Modellanalyse

Modellierungssprache und Analyseverfahren für Prozesse in automatisierten Produktionssystemen

In Arbeitspaket 1 (AP1) wurde durch die Universität Koblenz-Landau eine Modellierungssprache für Prozesse in automatisierten Produktionssystem (aPS) mit Fokus auf Ausschussrisiken, ein Analyseverfahren für solche Modelle hinsichtlich dieser Risiken und eine Methode zur Wiederverwendung von Modellteilen innerhalb und zwischen Organisationen vorgestellt. Die Lösung ermöglicht es, aPS zeit- und kosteneffizient zu modellieren und zu analysieren. Die folgende Abbildung zeigt eine einfach zu handhabende Methode zur Modellierung und Analyse von Produktionssystemen mit ihren Eingangs-, Ausgangsmessungen und Randbedingungen.

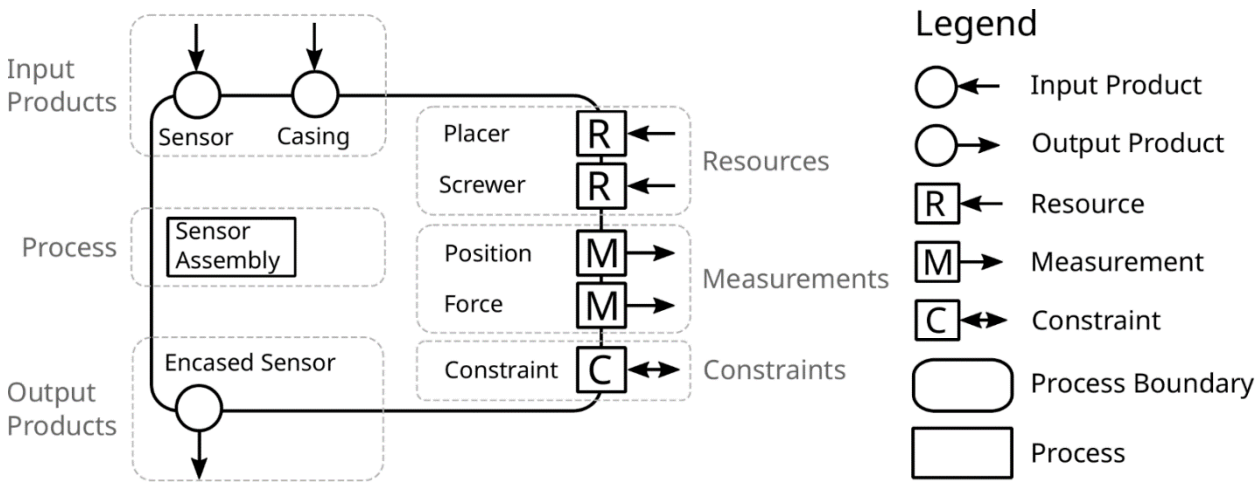

Abbildung 2: Modellierung und Analyse von Produktionssystemen

In der folgenden Tabelle sind die Probleme aufgeführt, die mit den Mustern in dieser Arbeit erkannt werden können. Wann immer das Systemmodell geladen oder von Grund auf neu erstellt wird, können sechs Problemanalysemuster verwendet werden, um die dargestellten Probleme zu erkennen.

Name	Definition
Produkt ist einer Risikoquelle ausgesetzt (Product Exposed to Risk Source)	Ein Produkt ist einer Risikoquelle ausgesetzt und kann geschädigt werden
Ressource ist einer Risikoquelle ausgesetzt (Resource Exposed to Risk Source)	Eine Ressource ist einer Risikoquelle ausgesetzt und kann geschädigt werden.
Unbeschränkte geänderte Eigenschaft (Unconstrained Changed Property)	Eine Eigenschaft wurde geändert und hat keine Einschränkung.

Ungemessene geänderte Eigenschaft (Unmeasured Changed Property)	Eine Eigenschaft wurde geändert und hat keine Messung
Ungenutzte Messung (Unused Measurement)	Eine Messung wird nicht verwendet. Setzen Sie eine Einschränkung für den Messwert
Abweichung zwischen Eingabe und Ausgabe (Input Output Mismatch)	Zwei Prozesse sind miteinander verbunden. Der erste Prozess stellt ein Produkt her, und der zweite Prozess verwendet das Produkt nicht als Input. Dies kann auf eine Fehlkonfiguration zurückzuführen sein.

Tabelle 2: Definition der Problemanalysemuster

Die folgende Abbildung zeigt die Liste der Problemanalysemuster im Prototyp des Tools. Um die Methode leicht anpassbar zu machen und die Anforderungen und Qualitätsziele zu erfüllen, ist sie in einer Webanwendung implementiert, die ohne Installation genutzt werden kann.

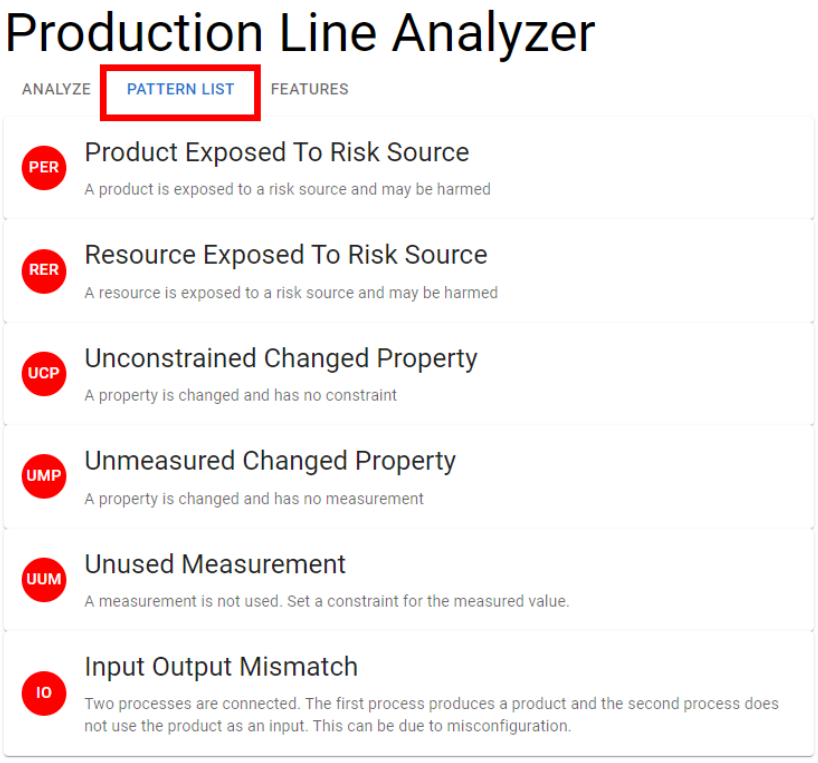

Abbildung 3: Problemanalysemuster in der Anwendung

Formalisierung des beschriebenen Modells

Um das beschriebene Modell zu formalisieren, wird das RDF (Resource Description Framework) verwendet. Das RDF ist ein Rahmenwerk zur Beschreibung von verknüpften Daten in Form von Graphen auf der Grundlage von Subjekt-Prädikat-Objekt-Tripeln. Die Verwendung eines weit verbreiteten Frameworks wie das RDF hat den Vorteil, dass es von Tools unterstützt wird, dokumentiert ist, und vorhandene Wissensquellen importiert werden können. Um das Modell im RDF zu formalisieren, werden die Konzepte von Input- und Output-Produkten, Ressourcen, Messungen und Beschränkungen in Form von Subjekt-Prädikat-Objekt-Tripeln ausgedrückt. Das Modell kann aus einer RDF-Datei geladen oder mit einem Editor eingegeben und verändert werden.

Eingabemodell der Analyse und Beispielszenario

In diesem Abschnitt wird das Eingabemodell der Analyse anhand eine realitätsnahen Beispielszenarios erläutert. Die folgende Abbildung zeigt die visuelle Darstellung des Eingabemodells der Analyse.

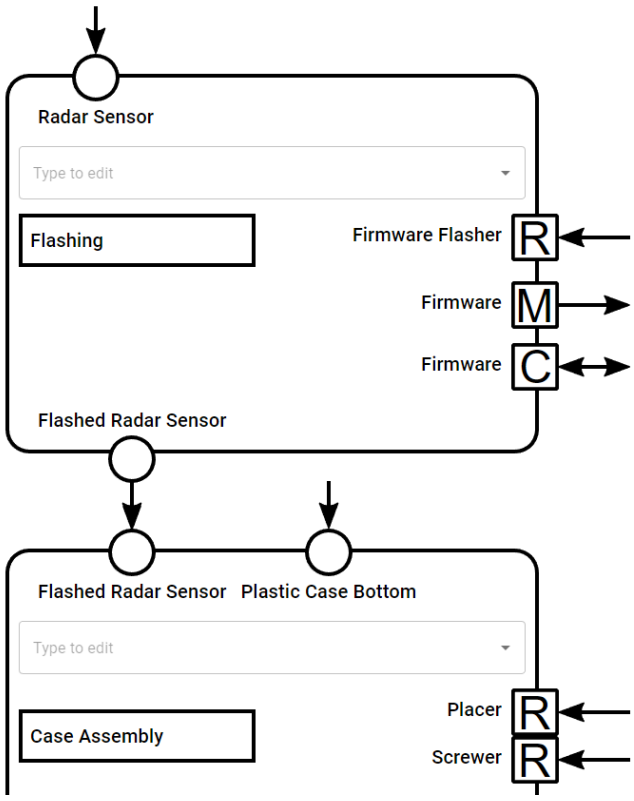

Abbildung 4: Visuelle Darstellung des Eingabemodells

Das Beispielszenario besteht aus zwei Prozessen, die mit einer Produktionslinie verkettet sind. Der erste Prozess ist das "Flashing" und hat das Radarsensorprodukt als Input. Er verwendet die Ressource "Firmware Flasher", um die Firmware des Sensors zu

aktualisieren. Die Firmware-Version wird gemessen und in einem Versions-Constraint überprüft. Das geflashte Sensorausgangsprodukt wird als Eingangsprodukt an die nächste Station weitergegeben.

Der zweite Prozess nennt sich Paketmontage und nimmt den geflashten Radarsensor und einen Plastikverpackungsboden als Eingangsprodukt. Der Prozess verwendet die Placer-Ressource, um die beiden Eingangsprodukte übereinander zu legen. Dann verwendet der Prozess die Schraubendreher-Ressource, um die Teile zusammenzuschrauben. Die Position des Placers und die Kraft des Schraubers werden gemessen. Der zusammengebaute Sensor ist der Output und wird an den nächsten Prozess weitergegeben.

Im Folgenden wird das Ergebnis der Analyse im Eingabemodell visuell dargestellt. Die visuelle Ausgabe der Analyse ermöglicht es, Probleme im Systemkontext schnell zu erkennen. In dem Beispielszenario ist der geflashte Radarsensor verwundbar für Krafteinwirkung, deswegen wird die Schraubendreher-Ressource, welche Kraft ausübt und alle Messungen und Randbedingungen, welche die Kraft messen oder beschränken, rot markiert.

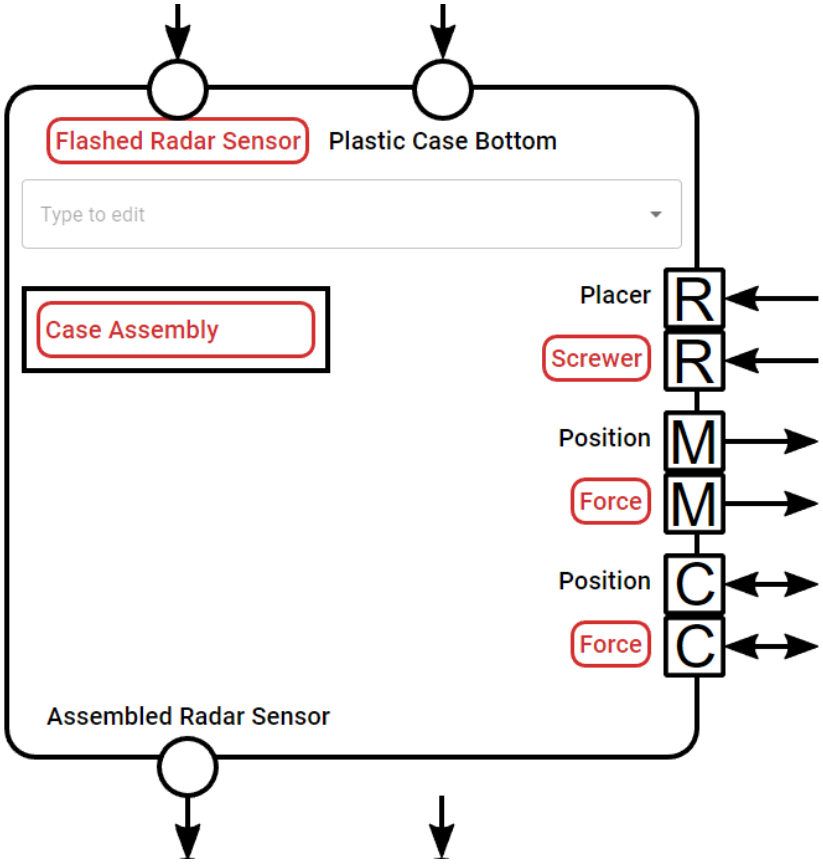

Abbildung 5: Visuelle Darstellung des Resultats der Analyse

Darüber hinaus stellt die folgende Abbildung die textliche Beschreibung eines Problems dar, das mit der vorgeschlagenen Problemanalyse gefunden wurde.

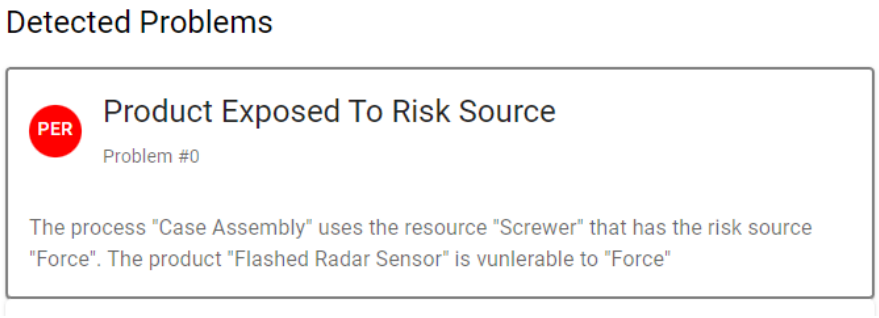

Abbildung 6: Textuelle Darstellung eines erkannten Problems

Dynamische Modellierung

Das dynamische Softwaremodell zielt darauf ab, den zeitlich aktuellen Zustand eines bestimmten Prozesses zu zeigen, während er ausgeführt wird. Es ermöglicht die Erfassung von Datenschnappschüssen, die jeder Phase des Prozesses entsprechen. Die erfassten Daten enthalten sowohl operative als auch parametrische Daten. Einerseits stellen die operativen Daten den Zustand der Sensoren und Maschinen dar, d.h. wenn eine bestimmte Fertigungskomponente in den Fertigungsstrom geladen oder aus ihm entnommen wird. Andererseits stellen die parametrischen Daten den physikalischen Zustand der Maschinen dar, wie z. B. Temperatur und Druck. Das dynamische Modell besteht aus zwei Hauptansichten oder Schichten: der digitalen Modelldarstellung und dem digitalen Schatten. Die digitale Modelldarstellung enthält ein Eins-zu-eins-Abbild der physischen Realität, d. h. der Stationen, der Fertigungskomponenten (Teile), der vorgesehenen Parameter und der akzeptierten Parameterwertebereiche. Der digitale Schatten kann Objekte enthalten oder aufnehmen, die dem digitalen Modell entsprechen, Instanzwerte verschiedener Zeitstempel des Prozesses haben und ein Mittel zur Abfrage und Erfassung von Daten zu bestimmten Zeitpunkten des Prozesslebenszyklus bieten.

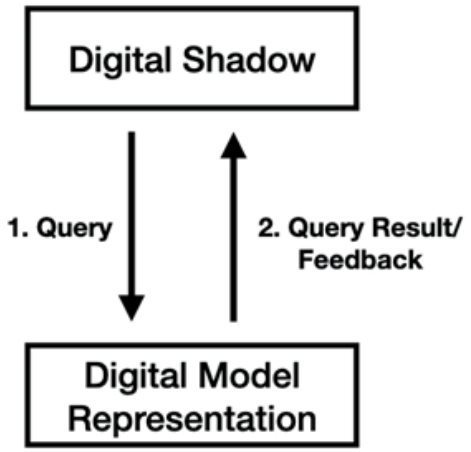

Abbildung 7: Design der dynamischen Modellierung und Analyse

Fehlervorhersagemodell

In AP2 (Extraktion von Erklärungen zur Fehleranalyse und Erkennen von Verbesserungspotentialen in der digitalen Fabrik) wurde ein Fehlervorhersagemodell anhand einer größeren Menge historischer Produktionsdaten in mehreren Iterationen gemäß dem branchenübergreifenden Standardprozess für Data Mining CRISP-DM erstellt. Dieses Fehlervorhersagemodell verwendet Messungen als Eingabe und liefert die direkten Vorhersagen, ob diese Messungen mit erhöhter Wahrscheinlichkeit zu Fehlern führen, sowie die dazugehörigen Konfidenzen der Vorhersagen.

Zur Erstellung des Fehlervorhersagemodells wurden Workshops mit HELLA durchgeführt, um das Business-Verständnis also grundlegendes Verständnis der automatisierten Produktionsanlage sicherzustellen. Anschließend wurden anhand von zunächst nur kleinen Samplen, und später größeren Datenmengen über längere Zeiträume iterativ das Datenverständnis und die Datenvorverarbeitung angepasst, um die Produktionsdaten zusammenzuführen und in ein Format zu bringen, welches für das Trainieren von Modellen des maschinellen Lernens geeignet ist. Dafür wurden die Daten der einzelnen Stationen so kombiniert, dass Messungen der früheren Stationen und Einstellparameter aller Stationen als Eingabegrößen, und der Ausschuss in späteren Stationen als zu vorhersagende Labels gegenübergestellt werden.

Zudem wurden diejenigen Messungen identifiziert, die für die Vorhersage von Fehlern im weiteren Verlauf der Produktionslinie am besten geeignet sind, und aus diesen Messungen wurden Features generiert, die möglichst sauber und aussagekräftig sind. Auf Basis dieser Features haben wir verschiedene Machine-Learning-Modelle trainiert, um den Ausschuss in späteren Stationen anhand der Messungen der früheren Stationen vorherzusagen,

angefangen bei simplen linearen Modellen und Entscheidungsbäumen, bis hin zu Support Vector Machines und Gradient Boosted Decision Trees. Um sicherzustellen, dass unsere Modelle gut verallgemeinern und präzise Vorhersagen für zukünftige Daten treffen können, wurde die Vorhersagegüte anhand von nicht im Training genutzten Daten validiert.

Da die Erklärungs-Engine aus modellagnostischen Methoden besteht, gibt es zwar keine spezifischen Limitationen für die Art des Modells, allerdings hängt die Erklärungsgüte der Erklärungs-Engine stark von der Vorhersagegüte der Modelle ab. Durch die hohe Modularität könnte das zugrundeliegende Modell zwar ausgetauscht werden, aber für die finale Version des Demonstrators wurden optimierte Gradient Boosted Decision Trees eingesetzt, da diese die höchste Vorhersagegüte auf Validierungsdaten aufgezeigt haben.

Erklärungs-Engine

Die Aufgabe der von der Universität Stuttgart entwickelten Erklärungs-Engine besteht darin, wertvolles Feedback für die Endbenutzer der XAPS-Plattform zu generieren, indem sie Tools bereitstellt, um die Vorhersagen des Fehlervorhersagemodells besser zu verstehen, um somit eine Fehleranalyse und eine Identifikation potenzieller Verbesserungen in der automatisierten Produktionsanlage zu ermöglichen. Dieses Feedback in Form von sogenannten eXplainable Artificial Intelligence (XAI)-Methoden ist entscheidend für den Aufbau vertrauenswürdiger und benutzerorientierter maschineller Lernmodelle.

Die endgültige Version der XAPS-Plattform umfasst mehrere hochmoderne Methoden, die numerische und visuelle Informationen liefern, um die komplexen inneren Abläufe der Produktionslinie von unserem Partner HELLA zu beleuchten. Beispielsweise kann die Erklärungs-Engine bei einem bestimmten Ausfall in der Produktionslinie auf die Station oder Maschine hinweisen, die höchstwahrscheinlich für diesen bestimmten Ausfall verantwortlich ist.

Durch die Analyse und Nutzung der Erkenntnisse aus den realen Ausfällen in den vom Partner HELLA bereitgestellten Daten sowie der Vorhersagen des Fehlervorhersagemodells ist die Erklärungs-Engine auch in der Lage, dem Benutzer eine Vorstellung davon zu vermitteln, welche Maschinen wahrscheinlich generell Ausfälle verursachen. Dies könnte darauf hinweisen, dass eine Maschine Wartung benötigt oder dass ihre Einstellungen nicht optimal sind.

Erklärbare KI-Methoden können in modellspezifische und modellagnostische Methoden unterteilt werden. Modellspezifische Methoden nutzen die inneren Abläufe des Modells, das sie erklären sollen, z.B. die Parameter eines neuronalen Netzwerks. Modellagnostische Methoden hingegen benötigen nur die Ein- und Ausgabepaare des Modells, das sie erklären sollen, sodass sie mit jedem Vorhersagemodell verwendet werden können. Die

Erklärungs-Engine besteht aus modellagnostischen Methoden, um eine bessere Integration und Modularität des endgültigen Produkts zu gewährleisten und um sicherzustellen, dass unsere Erklärungen auch dann relevant sind, wenn sich das Fehlervorhersagemodell in Zukunft ändert. Dies könnte aufgrund neuer Fortschritte im maschinellen Lernen oder durch Änderungen in der Quantität oder Qualität der verfügbaren Daten geschehen. Zudem erlaubt dieses Vorgehen das Verwenden von großen Datenmengen und komplizierten Abhängigkeiten zwischen den Messungen.

Die Erklärungs-Engine besteht aus den folgenden Komponenten:

LIME: Local Interpretable Model-Agnostic Explanations (LIME) gehört zur Gruppe der Methoden, die interpretierbare Ersatzmodelle (surrogate models) trainieren, um die Vorhersagen des zugrunde liegenden Black-Box-Modells anzunähern. Die Methode wurde erstmals 2016 vorgestellt und wird auch heute noch häufig für Text- und Bildklassifizierungsprobleme verwendet.

ICE: Ein Individual Conditional Expectation (ICE) Graph zeigt, wie sich die Vorhersage der Instanz ändert, wenn sich eine Eingabegröße ändert. Die Werte der Grafik werden berechnet, indem alle anderen Eingabegrößen unverändert bleiben und künstliche Instanzen geschaffen werden, indem der Wert der interessierenden Eingabegröße durch alle möglichen Werte ersetzt wird, die es im gesamten Datensatz annimmt. Dann werden die Vorhersagen des Fehlervorhersagemodells für diese neu erstellten künstlichen Instanzen generiert. Das Diagramm ist für den kleinsten Merkmalswert bei 0 zentriert und zeigt die relative Änderung der vorhergesagten Wahrscheinlichkeit. Dies hilft, die Trends der ICE-Kurven für verschiedene Instanzen zu vergleichen.

PDP: Partial Dependence Plots (PDP) zeigen die Abhängigkeit zwischen der Zielfunktion und einer Eingabegröße von Interesse, wobei die Werte aller anderen Eingabegrößen im Datensatz marginalisiert werden. Eine solche Darstellung kann zeigen, ob die Beziehung zwischen dem Ziel und einer Eingabe linear, monoton oder komplexer ist. Das Diagramm berücksichtigt alle Instanzen des Datensatzes, um die globale Beziehung einer Eingabegröße mit dem vorhergesagten Label zu zeigen.

Das Diagramm wird folgendermaßen erstellt: Für die Eingabegröße von Interesse wird die Liste aller eindeutigen Werte im Datensatz gesammelt. Dann wird für jede Instanz im Datensatz jeder mögliche Wert der Eingabegröße durchlaufen, eine Vorhersage für diese künstlichen Instanzen gemacht und geprüft, wie sich die Vorhersage des Modells ändert, wenn wir die Eingabegröße auf einen bestimmten Wert ändern. Wenn man diese Wahrscheinlichkeiten gegen die Werte der Eingabegröße als Liniendiagramm aufträgt, erhält man das oben beschriebene ICE-Diagramm. Das PD-Diagramm erhält man dann durch einfache Mittelung der ICE-Diagramme für jede Instanz im Datensatz.

Counterfactuals: Eine Counterfactual Erklärung beschreibt eine kausale Situation: “Wenn X nicht eingetreten wäre, wäre Y nicht eingetreten”. Beim interpretierbaren maschinellen Lernen können Counterfactuals verwendet werden, um Vorhersagen für einzelne Instanzen zu erklären. Das *Ereignis* ist das vorhergesagte Ergebnis einer Instanz, die *Ursachen* sind die jeweiligen Eingabegrößen dieser Instanz, die in das Modell eingegeben wurden und eine bestimmte Vorhersage kausal *verursacht* haben.

Eine Counterfactual Erklärung einer Vorhersage beschreibt die kleinste Änderung der Eingabegrößen, die die Vorhersage zu einem vordefinierten Ergebnis ändert. In unserem Fall ergibt sich daraus eine Rückmeldung für den Benutzer in der Art: "Dieser vorhergesagte Fehler könnte vermieden werden, wenn die Variable X diesen anderen Wert hätte".

Dashboard

In AP3 (Entscheidungsunterstützung und Benutzerschnittstelle), speziell in 3.1 (User Interface Design für Vorwissenserfassung und für Erklärungen) und in AP4 (Demonstrative Anwendungen), speziell in 4.3 (Demonstratoren) wurde anhand der in Abschnitt II.3 detaillierten möglichen Anwendungsfälle ein umfassendes Dashboard als eine zentrale benutzerfreundliche Oberfläche erstellt.

Dieses Dashboard erlaubt zunächst die Navigation durch die Datenbasis der Produktionsanlage, aber ermöglicht vor allem die Verwendung der anderen Ergebnis-Bestandteile. Die Modellanalyse wird hierbei direkt so eingebunden, wie in den Abbildungen 3, 4, 5 und 6 zu sehen ist. Das Dashboard bietet Optionen zur Auswahl eines von mehreren Fehlervorhersagemodellen an, inklusive grundsätzlicher Informationen dieser Modelle, wie z.B. den zugrundeliegenden Eingabegrößen und der auf historischen Daten geschätzten Güte des jeweiligen Modells. Für das ausgewählte Modell werden einzelne Vorhersagen inklusive der dazugehörigen Konfidenzen für die entsprechenden Teile im Dashboard direkt angezeigt. Zudem wird das Fehlervorhersagemodell indirekt über die Erklärungs-Engine eingebunden. Erklärungsmethoden der Erklärungs-Engine wie z.B. LIME und ICE werden grafisch in dem Dashboard eingebunden, wie in den folgenden anonymisierten Screenshots zu sehen ist.

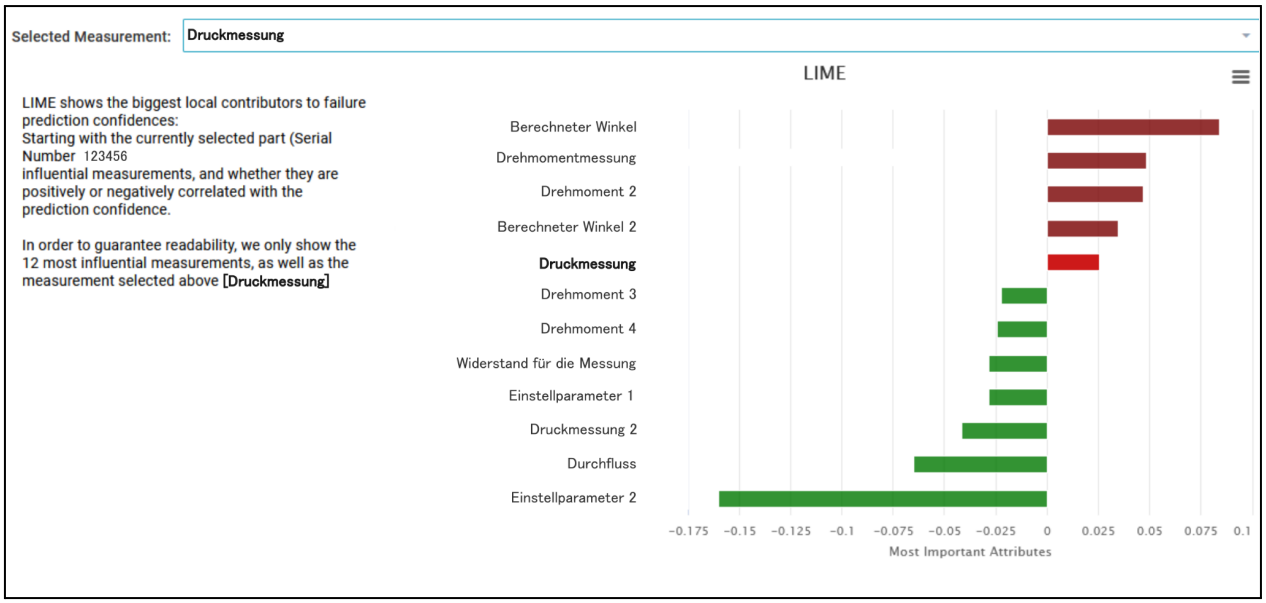

Abbildung 8: LIME und Auswahl von Messungen im Dashboard

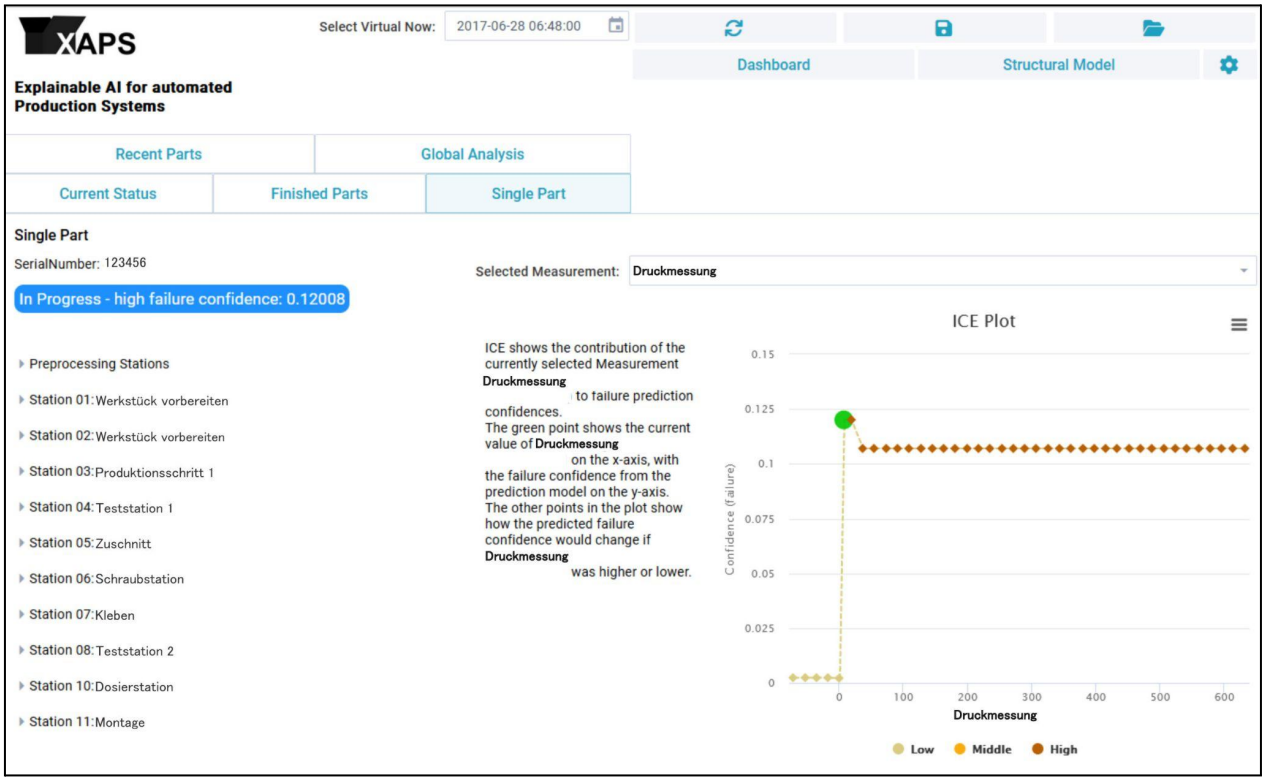

Abbildung 9: Stationsübersicht und ICE im Dashboard

Indem wir interaktive Implementierungen in den Vordergrund stellen, können wir dem Benutzer die Möglichkeit geben, einzelne Teile, einzelne Vorhersagen, oder Teilmengen von Interesse auszuwählen und zu erkunden. Diese Interaktivität bietet einen Einblick in die datenerzeugenden Prozesse und erläutert die den Vorhersagen zugrunde liegenden Überlegungen, was zu einem besseren Verständnis der Systemdynamik beiträgt.

Zur Erstellung des Dashboards wurde bereits früh in der Planung eine gute Skalierbarkeit durch eine sehr direkte Integration der Benutzeroberfläche mit den auf die Betriebsdaten zugreifenden Prozessen mitgedacht, wodurch das Dashboard auch für große Datenmengen echtzeitfähig ist.

Durch Betrachtung all dieser Informationen des Dashboards und Interaktionen mit ihnen können potenzielle Verbesserungen der Produktionsanlage identifiziert werden, und bei einer Echtzeit-Einbindung des Dashboards kann im Betrieb rechtzeitig reagiert werden, um Fehler vorzubeugen. Zudem können die Erklärungen genutzt werden, um Akzeptanz für das Dashboard zu schaffen und generelle Fehlerbilder zu analysieren und zu beheben.

2. Die wichtigsten Positionen des zahlenmäßigen Nachweises

Der Großteil der im Forschungsprojekt XAPS entstandenen Kosten waren für alle Konsortialpartner die jeweiligen Personalaufwendungen. Da aufgrund der Corona-Pandemie speziell in 2020 und 2021 nahezu komplett auf die Durchführung von persönlichen Meetings verzichtet wurde, und auch in 2022 grundlegend auf größere persönliche Treffen mehrerer Personen verzichtet wurde, gab es keine größeren Reisekosten.

Vorhandene Infrastruktur, wie z.B. Server von OWC konnten für eine kollaborative Plattform zum Daten- und Informationsaustausch zwischen Konsortialpartnern, sowie für die Datenaufbereitung und Speicherung von Daten, und nicht zuletzt für das Trainieren und Optimieren von Modellen des maschinellen Lernens wiederverwendet werden, wodurch keine großen Materialkosten zustande kamen.

Durch die Ausscheidung des Projektpartners iTAC aus dem Konsortium wurde eine Budgetumverteilung ausgeführt, wodurch die Bestrebungen der Inhalte weiterverfolgt werden konnten und das Fraunhofer ISST ihre Arbeiten in AP4 intensivieren konnte.

Insgesamt konnten so trotz Verlängerung des Projektzeitraums auf Januar 2020 - Dezember 2022 die Projektkosten plangemäß eingehalten werden, wodurch zuzüglich zu den oben

beschriebenen Umverteilungen am ursprünglichen Ausgabenplan keine Änderungen notwendig geworden sind.

Der zahlenmäßige Nachweis selbst wird in einem separaten Dokument dargestellt.

3. Notwendigkeit und Angemessenheit der geleisteten Arbeit

Um einen Überblick über die Anforderungen und Use Cases des Projekts zu erhalten, wurden mehrere Workshops und Treffen mit den Projektpartnern vom Fraunhofer ISST organisiert und durchgeführt. Die Ergebnisse dieser Besprechungen bilden die Grundlage für die in der folgenden Tabelle beschriebenen Anwendungsfalldefinitionen.

Im Detail werden die Basisanwendungsfälle spezifiziert, um eine Roadmap für die Generierung von Anforderungen in Form von User Stories bereitzustellen. Das folgende Use Case-Diagramm gibt einen ersten Überblick. Weitere Details finden Sie in der folgenden Tabelle mit zusätzlichen Informationen zu den Anwendungsfällen und ihren Beziehungen.

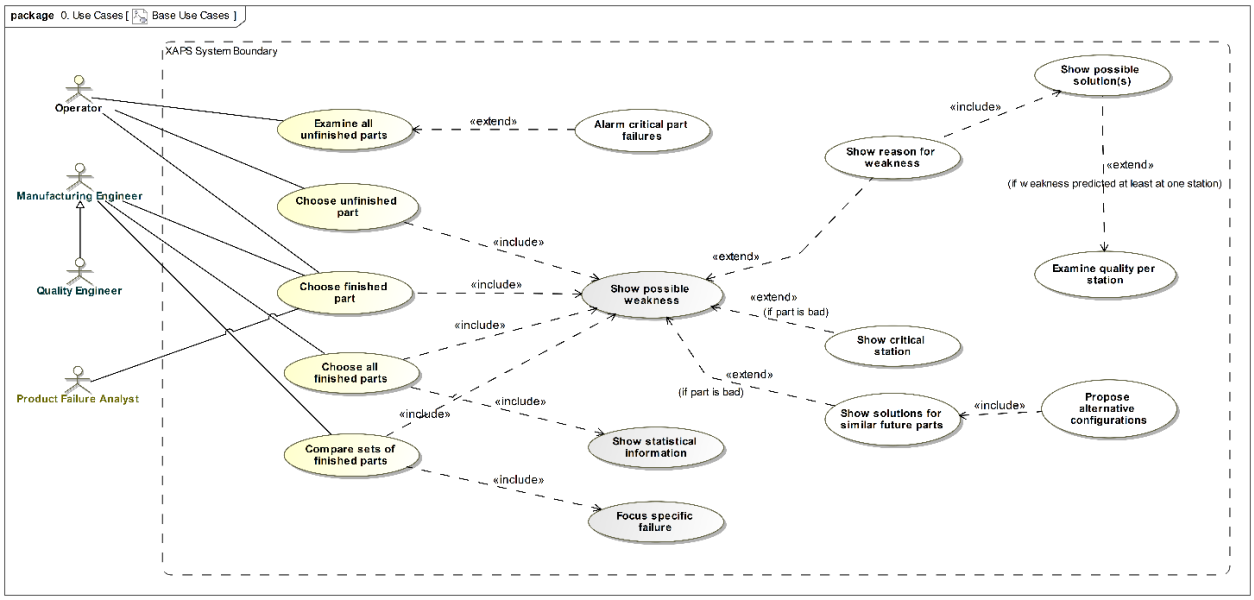

Abbildung 10: Use Case Diagram

UC ID	Use Case	Roles	Constraints	Comments
UC10	Examine all unfinished parts	Operator	-	All unfinished parts are currently available on the production line, and additional functionality to examine them further.
UC11	Alarm critical part failure	-	Extends UC10	If a critical part failure is assumed, it will be shown to the operator with more details.
UC20	Choose unfinished part	Operator	-	Similar to UC10 but with a single part to choose from that are currently on the production line.
UC30	Choose finished part	Operator, Manufacturing Engineer, Quality Engineer	Includes UC60	Similar to UC20 but with finished parts that ultimately passed the production line.
UC40	Choose all finished parts	Manufacturing Engineer, Quality Engineer	Includes UC60, includes UC62	Similar to UC30 but for all parts that passed the production line ultimately.
UC50	Compare Sets of finished parts	Manufacturing Engineer, Quality Engineer	Includes UC60, includes UC64	The actor can choose different amounts of finished parts assets to examine further and compare them.
UC60	Show possible weakness	-	-	It is included in the prominent use cases and predicts or calculates possible weaknesses in finished or unfinished parts.
UC62	Show statistical information	-	-	Invokes statistical information in the use case for all finished production line parts.

UC64	Focus specific failure	-	-	Included in the comparison of part sets use case to add the possibility of focusing on a specific type of failure.
UC66	Show reason for the weakness	-	Extends UC60, includes UC72	Extends the use case UC60 when a possible weakness has been found to find reasons for the suggestion.
UC68	Show critical station	-	Extends UC60	If a part is declared as "bad," the corresponding station can be shown.
UC70	Show solutions for similar future parts	-	Extends UC60, includes UC76	If a part is declared as "bad," possible solutions for details of a similar type can be shown.
UC72	Show possible solution(s)	-	Extends UC74	This use case is included in UC66 to examine possible weaknesses further.
UC74	Examine quality per station	-	-	A prediction of part quality for all stations of the whole production line.
UC76	Propose alternative configuration(s)	-	-	If there has been a possible solution calculated in UC70, this use case is included to propose configurations of the stations to avoid a specific weakness.

Tabelle 3: Mögliche Anwendungsfälle

Qualitätsziele

Die Qualitätsziele für dieses Projekt werden unterteilt in die allgemeinen Ziele, die für das gesamte Projekt definiert werden, und die Qualitätsziele, die speziell für die Komponenten gelten, die sich mit Algorithmen des maschinellen Lernens befassen.

Allgemeine Qualitätsziele

Die folgenden fünf wichtigsten Qualitätsziele sind bei der Entwicklung der Systemarchitektur zu berücksichtigen (ISO25010).

Qualitätsziele	Spezifisches Qualitätsziel	Kommentar	Priorität
Leistungseffizienz	Zeitverhalten	Kurze Wartezeiten bei Use Cases und Operationen, die mit der hohen Datenmenge arbeiten.	Hoch
Kompatibilität	Interoperabilität	Alle Komponenten des Systems müssen zusammenarbeiten. Die externen Schnittstellen müssen klar spezifiziert sein, um eine einfache Handhabung zu ermöglichen.	Hoch
Zuverlässigkeit	Wiederherstellbarkeit	Das System sollte nach einem kritischen Problem wiederhergestellt und neu gestartet werden. Es muss vermieden werden, dass ein technischer Ingenieur notwendig ist, um das System in der Betriebsumgebung neu zu starten.	Hoch
Wartbarkeit	Modularität	Das System sollte klar definierte Funktionalitäten enthalten, die in verschiedene Komponenten mit klaren Anliegen gekapselt sind.	Hoch
Übertragbarkeit	Anpassungsfähigkeit	Das System sollte leicht auf andere Umgebungen (Produktionslinien-Setups) oder Anwendungsfälle (Unternehmen) übertragbar sein.	Hoch

Tabelle 4: Allgemeine Qualitätsziele

Qualitätsziele für maschinelles Lernen

Die folgenden Qualitätsziele für den Teil des maschinellen Lernens (ML) sind bei der Entwicklung der Systemarchitektur zu berücksichtigen.

Qualitätsziele	Spezifisches Qualitätsziel	Kommentar	Priorität
Funktionelle Eignung	Funktionale Korrektheit	Das System sollte eine hohe Zuverlässigkeit in Bezug auf die Richtigkeit der Vorhersagen und Ergebnisse der ML "Black Box" haben.	Hoch
	Interpretierbarkeit	Alle Ergebnisse und Entscheidungen des ML-Teils sollten in hohem Maße interpretierbar sein.	Hoch
Zuverlässigkeit	Funktionsfähigkeit	Das ML-System und seine Ergebnisse sollten von relevanten Stakeholdern nutzbar sein.	Hoch
Wartbarkeit	Modularität	Das ML-System sollte geringe Abhängigkeiten zu anderen Teilen des Systems haben, die nicht mit ML zu tun haben.	Hoch
	Anpassungsfähigkeit	Das ML-System sollte an neue Situationen, z.B. Änderungen in der Produktionslinie, anpassbar sein.	Hoch

Tabelle 5: Qualitätsziele für maschinelles Lernen

Die Umsetzung von Industrie 4.0-Systemen erfordert die Einbeziehung von erklärbaren maschinellem Lernen (XAI), um Transparenz, Vertrauen und ethische Entscheidungsfindung in automatisierten Prozessen zu gewährleisten. Die komplexe Natur von maschinellen Lernmodellen, insbesondere Deep Learning, führt oft zu Blackbox-Entscheidungen, die es den Interessengruppen unmöglich machen, die Gründe für diese Entscheidungen zu verstehen und zu validieren. XAI zielt darauf ab, diese Lücke zu schließen, indem sie menschlich interpretierbare Erklärungen für die Ergebnisse des maschinellen Lernmodells liefert. Darüber hinaus kann das Problem der Anpassung von

KI-Systemen an die menschlichen Werte und Absichten durch XAI gelöst werden. Indem maschinelle Lernmodelle erklärbar gemacht werden, können Interessengruppen die Ausrichtung von KI-Systemen an gewünschten Zielen und ethischen Erwägungen prüfen und validieren und so eine verantwortungsvollere Integration von KI in die Industrie 4.0 fördern.

Wie die in Tabelle 3 angegebenen Anwendungsfälle und die in Tabellen 4 und 5 angegebenen Qualitätsziele zeigen, gibt es einen klaren Bedarf für die Einbeziehung von XAI, um Transparenz, Vertrauen und ethische Entscheidungsfindung in automatisierten Produktionsanlagen sicherzustellen. Vorgefertigte Lösungen sind jedoch nicht ausreichend, um die spezifischen zuvor dargestellten Anforderungen zu erfüllen, was die Notwendigkeit von Forschung auf diesem Gebiet unterstreicht.

Die in AP4.4 auf der Seite von HELLA durchgeführte Evaluation des Demonstrators hat eindeutig die Angemessenheit der im Forschungsprojekt XAPS geleisteten Arbeiten dargestellt. Die Evaluation wurde in einem diversen Team von Experten für automatisierte Produktionsanlagen durchgeführt (Fertigungsplaner, Linienbediener, Qualitätsingenieur und Analysemitarbeiter), und hat dargelegt, dass 8 der zuvor definierten möglichen Anwendungsfälle (UC10, UC20, UC30, UC40, UC50, UC64, UC66, UC68) bereits vollständig erfüllt sind, und 4 weitere Anwendungsfälle (UC11, UC60, UC62, UC70) erfüllt sind, aber weitere Wünsche offen bleiben. Für eine vollständige Erfüllung der letzteren 4 Anwendungsfälle wären zwei mögliche Szenarien zu verfolgen. Erstens könnten durch kleine Ergänzungen der geleisteten Arbeiten z.B. eine Alarmfunktion und Reporting, die für UC11 nötig wären, integriert werden. Das zweite Thema ist durchaus komplexer und wäre durch Involvierung von Anschlussarbeiten mit Large Language Models erforschbar, wie in dem *Ausblick* in II.5 beschrieben, um z.B. für UC60, UC62, und UC70 besseres textbasiertes Feedback zu erzeugen.

4. Voraussichtlicher Nutzen und Verwertbarkeit

Auf Seiten von der Universität Koblenz-Landau und der Universität Stuttgart wurden und werden die Ergebnisse im Forschungs- und Lehrbetrieb verwendet. Teile der erarbeiteten Ergebnisse sind in Vorlesungen im Bereich des sicheren Software Engineering in die Lehre der Universität Koblenz-Landau eingeflossen. Weiterhin waren Themen des Projekts Grundlage verschiedener Abschlussarbeiten, die am Institut für Software Engineering der Universität Koblenz-Landau geschrieben wurden.

Das erstellte Dashboard wird bei HELLA prototypisch eingesetzt, um den Ansatz zu evaluieren. Durch parallele Entwicklungen im Unternehmen wurden Plattformen etabliert, durch die ähnliche Funktionen realisierbar sind. Daher entwickelt HELLA gerade auf Basis der im Forschungsprojekt generierten Kenntnisse weitere Reports und Dashboards, die auf dem Forschungsprojekt aufbauen. Diese Entwicklungen werden nun weiter im Unternehmen evaluiert. Zunehmend ist ein steigendes Interesse der Stakeholder zu erkennen. Aus HELLA Sicht wurden wichtige Erkenntnisse erzielt, die in den kommenden Jahren zu einer stärkeren Einführung und Nutzung von Maschinelles Lernen und KI führen werden.
Durch die vom Beginn des Projektes an mitgedachte hohe Modularität, und die flexible Anpassungsfähigkeit der einzelnen Komponenten, könnten diese auch für andere Produktionsanlagen wiederverwendet werden, was sowohl deutlich geringere einmalige Kosten für jede weitere Inbetriebnahme, als auch einen insgesamt hohen wirtschaftlichen Nutzen bedeutet. Gerade bei der langfristigen Nutzung ist für HELLA eine Erreichung der in Tabelle 4 und Tabelle 5 detaillierten Qualitätsziele sehr wichtig. Hierbei wird das Hauptaugenmerk auf funktionelle Eignung und Wartbarkeit gelegt. Im Industrieumfeld ist wichtig, dass die prognostizierten Ergebnisse eine sehr hohe “Funktionale Korrektheit” aufweisen, da darauf direkt Schlussfolgerungen aufgebaut werden sollen und diese einen hohen finanziellen Einfluss auf das Geschäftsergebnis haben. Der zweite wichtige Punkt ist die Interpretierbarkeit, um eine hohe Akzeptanz innerhalb der verschiedenen Funktionsbereiche zu erhalten. Hierbei ist besonders wichtig, die Themen für die verschiedenen Qualifikationen entsprechend verständlich verfügbar zu machen. Wenn dies nicht ausreichend gegeben ist, wird eine Nutzung sehr schnell abgelehnt. Im aktuellen Dashboard wurde die Zuverlässigkeit sehr gut erfüllt. Eine Web-basierte Ansicht ermöglicht einen schnellen und einfachen Zugriff von allen Stakeholdern. Innerhalb der Qualitätsziele ist zudem auch Wartung als Ziel definiert. Dies ist sowohl für die Betreibung des Systems (Modularität und Anpassbarkeit) wichtig, aber gerade diese Modularität führt auch zu einer schnellen Skalierbarkeit, die zusätzlich entscheidend für den Erfolg innerhalb der Organisation ist. Durch Standards kann hier eine schnelle Erweiterung auf neue Produktionslinien ermöglicht werden. Im Projekt wurde deutlich, dass gerade hier noch Weiterentwicklungen notwendig sind.

Zusätzlich zu den im Dashboard gewonnenen Erkenntnissen sind von HELLA Seite die Modelle der statischen und dynamischen Modellanalyse als sehr interessantes Mittel zur Visualisierung empfunden worden. Seitens HELLA sind in diese Richtung vorher keine Aktivitäten geplant gewesen. Durch das Projekt sind die Vorzüge solcher Visualisierungen klar geworden. HELLA prüft nun auch interne Verwendung dieser Methoden und automatische Erstellung solcher Modelle anhand von Manufacturing Execution System (MES) Daten.

Wie bereits in der Vorhabensbeschreibung geplant, können Ergebnisse des Projektes durch Integration in die Produktpalette von Old World Computing GmbH verwertet werden. Arbeiten zur Erstellung eines Benutzerinterfaces in AP3 spiegeln sich in Form des WebAppBuilders als Erweiterung für die Plattform RapidMiner wider. Dieser dient als ein Werkzeug zur Erstellung von benutzerfreundlichen Oberflächen mit Verbindung zu komplexen datenbasierten Prozessen, und erschließt damit zum einen neue Anwendungsfelder und ermöglicht zudem eine signifikant bessere soziale Akzeptanz von technischen Lösungen. Zusätzlich hat sich bei der Zusammenführung der Komponenten im zentralen Dashboard gezeigt, dass auch die statische und dynamische Modellanalyse, sowie die Methoden zur Erklärbarkeit von KI über RapidMiner und dem WebAppBuilder einbindbar sind. Diese Resultate können in späteren Projekten verwendet werden und ermöglichen eine größere Transparenz und damit einhergehend mehr Vertrauen in die KI-Lösungen von OWC und in die Möglichkeiten von Methoden der künstlichen Intelligenz im Allgemeinen.

5. Fortschritte bei anderen Stellen

Erklärbare Künstliche Intelligenz (XAI) ist ein neues und wichtiges Forschungsfeld, da es die Akzeptanz von Entscheidung von KI-Modellen fördert. XAI-Methoden ermöglichen ein tiefes Verständnis der Entscheidungsprozesse von KI-Modellen, denn sie stellen die Faktoren und Regeln, die im Entscheidungsprozess berücksichtigt wurden, transparent dar.

Das Thema ist weiterhin hochaktuell, und gerade in den vergangenen 1-2 Jahren gab es signifikante Fortschritte bei KIs für Bild- und Spracherkennung, und -synthese durch Modelle, die auf Generative Pre-Trained Transformer basieren. Diese Fortschritte haben sich auch über den akademischen Kontext hinaus in der Öffentlichkeit durch von vielen Nutzern verwendeten generativen Text-zu-Text oder Text-zu-Bild Modellen wie ChatGPT[1], Stable Diffusion[2], und Midjourney[3] geäußert. Damit einhergehend wird der größte Teil der Arbeiten zum Thema Erklärbarer KI in den Bereichen Computer Vision und Verarbeitung natürlicher Sprache durchgeführt, in denen die Eingabedaten der Modelle des maschinellen Lernens aus Videos, Bildern oder Texten bestehen. So befassten sich beispielsweise beim diesjährigen Explainable ML Workshop in Tübingen[4] alle vorgestellten Arbeiten mit der Erklärbarkeit bei Computer-Vision-Aufgaben. Diese Bild-, Video- oder auch Textdaten ähneln sich von Domäne zu Domäne so sehr, dass riesige Datenmengen wie z.B. große Teile des Internets zum Training von Modellen verwendet werden können. Im Vergleich dazu

[1] https://openai.com/blog/chatgpt
[2] https://github.com/CompVis/stable-diffusion
[3] https://www.midjourney.com
[4] https://www.eml-unitue.de/eml-workshop

sind die in der Industrie 4.0, und speziell in Produktionsanlagen aufgezeichneten Daten wie Zeitreihen, sequentielle und tabellarische Daten komplexer, meist nicht öffentlich zugänglich, spezieller und für jede Art von Anlage einzigartig, wodurch in der Literatur spezifische Methoden um KI-Modelle für diese Arten von Datenstrukturen zu erklären, weiterhin fehlen.

Dennoch gibt es einzelne eingereichte Arbeiten zur Erklärbarkeit von zeitreihenbasierten Modellen, wie z.B. "*Explainable AI for Time Series via Virtual Inspection Layers*"[5] und "*XAI-based Comparison of Input Representations for Audio Event Classification*"[6]. Bei Zeitreihen besteht die größte Herausforderung darin, dass sie aufgrund der zeitabhängigen Dynamik schwer zu kodieren sind, was bedeutet, dass die Erklärungen nicht nur mit dem Blackbox-Fehlervorhersagemodell Schritt halten müssen, sondern sich auch jedes Mal an den spezifischen Kontext anpassen müssen.

Ausblick

Potenziell könnten die großen Fortschritte bei generativen Sprachmodellen wie z.B. ChatGPT in Kombination mit unseren Arbeiten und eventuell einem ergänzenden Korpus von Domänenwissen genutzt werden, um eine intelligente textbasierte Schnittstelle in das Dashboard zu integrieren. Diese könnte genutzt werden, um Zusammenhänge automatisch zu vertextlichen oder vielfältigere Interaktionen zu erlauben. Dafür ist es jedoch wichtig, dass auch diese Sprachmodelle erklärbar sind, um das Vertrauen der Benutzenden zu fördern und ihre Akzeptanz zu stärken.

6. Veröffentlichungen des Ergebnisses

Wesentliche Informationen zu dem Projekt wurden auf der Webseite https://explainable-ai.de/ veröffentlicht. Diese beinhaltet zum einen grundsätzliche Erklärungen zu dem Hintergrund des Forschungsprojekts, inklusive Erläuterungen zu den Themen KI, erklärbare KI, und automatisierten Produktionssystemen, sowie zum anderen Informationen zu den am Projekt beteiligten Forschungspartnern, Fakten zu dem Projekt und grundlegende Ergebnisse des Forschungsprojekts. Damit bietet die Webseite eine zentrale Anlaufstelle für erklärbare KI in automatisierten Produktionssystemen und ermöglicht zudem bei Interesse die Kontaktaufnahme mit den jeweiligen Konsortialpartnern.

[5] https://arxiv.org/abs/2303.06365
[6] https://arxiv.org/abs/2304.14019

Außerdem flossen Ergebnisse und Erkenntnisse des Projekts bereits in eine Vielzahl von Veröffentlichungen, Vorträgen und Abschlussarbeiten ein:

Veröffentlichungen

* C. Heise, A. Poddey, J. Jürjens. Strong long-term incentive design in GAIA-X based on tokenomics. Position Paper – GAIA-X TaskForce Tokenomics, 2020.

* S. Peldszus, J. Bürger, T. Kehrer, J. Jürjens. Ontology-Driven Evolution of Software Security. In Data & Knowledge Engineering (Elsevier). Special issue on selected publications at RCIS'2020, 2021. Accepted subject to minor revision.

* Q. Ramadan, D. Strüber, M. Salnitri, J. Jürjens, V. Riediger, S. Staab. A Semi-Automated BPMN-based Framework for Detecting Conflicts between Security, Data-Minimization and Fairness Requirements. In Journal of Software and Systems Modeling (SoSyM) (Springer Verlag). 2020, 35 pp.

* J. Bürger, T. Kehrer, J. Jürjens. Ontology Evolution in the Context of Model-based Secure Software Engineering. In 14th International Conference on Research Challenges in Information Science (RCIS 2020), Springer LNBIP, 2020, 16 pp.

* M. Ehl, M. Konersmann
Model-based Monitoring of Integrated UML State Machine Models and Code
EMLS 2021: 8th Collaborative Workshop on Evolution and Maintenance of Long-Living Software Systems

Vorträge

* J. Jürjens. How it all began and where we are going – Developing the IDS architecture from first principles and what IDS means for GAIA-X (Invited Lecture). In Data Sharing Winter School, session on Data Space Architecture, online, 2-3 Dec. 2020, IDSA.

* J. Jürjens, A. Sudhoff: Intelligente Digital Twins in der automobilen Industrie 4.0, TDWI München digital, 21.-23. Juni 2021

* K. Feichtinger, K. Meixner, F. Rinker, I. Koren, H. Eichelberger, T. Heinemann, J. Holtmann, M. Konersmann, J. Michael,. E.- M. Neuman, J. Pfeiffer, R. Rabiser, M. Riebisch, K. Schmid

Industry Voices on Software Engineering Challenges in Cyber-Physical Production Systems Engineering
(2022) IEEE International Conference on Emerging Technologies and Factory Automation, ETFA, 2022-September

* C. Scharpenberg. WebAppBuilder extension for RapidMiner – live introduction. Online webinar, 11 Feb. 2021.

* S. Land. Tell me why: Explainable KI. Session Research Bulletin, Data Science Ruhrgebiet Kongress 2022, 15 Sep. 2022.

Abschlussarbeiten

Eine Bachelorarbeit mit dem Titel *Enhancing the Security Design of Industrial IoT Platforms* (2022, Univ. Koblenz-Landau, Institut für Softwaretechnik) beschäftigte sich mit der Untersuchung von 12 Cloud Plattformen des Industriellen Internet der Dinge (IIoT), davon 3 Plattformen aus Deutschland und eine Plattform aus dem Europäischer Wirtschaftsraum (EWR), auf Sicherheits- und Datenschutzmechanismen.

Die studentische Masterarbeit *Categorizing Heterogeneous Data Sources in Automated Production Systems* (2022, Univ. Koblenz-Landau, Institut für Softwaretechnik) beschäftigte sich mit der Kategorisierung von im Produktionsumfeld eingesetzten Modellen.