Die Relevanz von LIME im Vergleich zu SHAP und Anchors: Eine literaturbasierte Analyse

Marcel Arian Hadi[1,1*]

[1*]universität Potsdam, Digital Engineering Fakultät, Potsdam, Germany.

Corresponding author(s). E-mail(s):
marcelarian.hadi@student.hpi.uni-potsdam.de;

Abstract

Mit der zunehmenden Verbreitung komplexer Machine-Learning-Modelle in sensiblen Bereichen wie Medizin, Finanzwesen oder Justiz wächst die Forderung nach Transparenz und Nachvollziehbarkeit automatisierter Entscheidungen. Explainable Artificial Intelligence (XAI) hat sich deshalb zu einem zentralen Forschungsfeld entwickelt. Diese Arbeit unterstützt die Relevanz von LIME im Vergleich zu moderneren Verfahren wie SHAP und Anchors.

Keywords: LIME, SHAP, Anchors, XAI, Interpretierbarkeit

1 Einleitung

Mit der zunehmenden Verbreitung komplexer Machine-Learning-Modelle in sensiblen Bereichen wie Medizin, Finanzwesen oder Justiz wächst die Forderung nach Transparenz und Nachvollziehbarkeit automatisierter Entscheidungen. Explainable Artificial Intelligence (XAI) hat sich deshalb zu einem zentralen Forschungsfeld entwickelt, das Methoden bereitstellt, um den "Black-Box"-Charakter moderner Modelle aufzubrechen und ihre Vorhersagen für Menschen interpretierbar zu machen.

Eine der einflussreichsten und früh eingeführten Methoden ist LIME (Local Interpretable Model-agnostic Explanations) Ribeiro, Singh, and Guestrin (2016). LIME zeichnet sich durch seine Einfachheit, Modellagnostik und breite Anwendbarkeit aus und wurde daher in Forschung und Praxis schnell adaptiert. Seitdem sind zahlreiche neuere Verfahren hinzugekommen, die alternative Ansätze und stärkere formale Fundierung bieten, wie etwa SHAP und Anchors.

Vor diesem Hintergrund stellt sich die Frage, ob LIME trotz der Verfügbarkeit modernerer Methoden weiterhin eine relevante Rolle spielt.

Die Untersuchung erfolgt auf Basis einer systematischen Literaturanalyse und verfolgt das Ziel, die Stärken, Schwächen und Einsatzkontexte von LIME im Vergleich zu SHAP und Anchors herauszuarbeiten und eine kritische Gesamtschau entlang zentraler Bewertungskriterien zu entwickeln.

2 Hintergrund

Explainable Artificial Intelligence (XAI) umfasst Methoden, die Entscheidungen von KI-Systemen nachvollziehbar machen sollen. Dabei wird zwischen intrinsisch interpretierbaren Modellen (z.B. Entscheidungsbäume, lineare Modelle) und post-hoc-Methoden unterschieden, die komplexe Black-Box-Modelle nachträglich erklären. Letztere haben sich in der Praxis als besonders wichtig erwiesen, da viele leistungsfähige Modelle von Natur aus nicht transparent sind.

LIME nimmt in dieser Entwicklung eine Pionierrolle ein. Es erklärt Vorhersagen lokal, indem es synthetische Nachbarschaften von Eingaben erzeugt, Modellantworten abfragt und ein vereinfachtes Surrogatmodell trainiert. Das Ergebnis sind Feature-Gewichte, die unmittelbar Aufschluss über die maßgeblichen Einflussgrößen einer Entscheidung geben Ribeiro et al. (2016).

SHAP (Shapley Additive Explanations) basiert auf den Shapley-Werten aus der kooperativen Spieltheorie. Die Methode zerlegt eine Modellvorhersage in viele Kombinationen von Eingabefeatures und berechnet, welchen durchschnittlichen Mehrwert jedes Feature zur Entscheidung beiträgt. Dadurch wird die Vorhersage als Summe von Feature-Beiträgen darstellbar, was eine faire und nachvollziehbare Attribution der einzelnen Variablen ermöglicht (Lundberg & Lee, 2017) Lundberg and Lee (2017).

Anchors ist ein von Ribeiro et al. entwickeltes Verfahren, das auf präzisen Wenn-Dann-Regeln basiert. Ziel ist es, Bedingungen ("Anker") zu identifizieren, die mit hoher Wahrscheinlichkeit eine bestimmte Modellvorhersage garantieren. Technisch geschieht dies durch ein systematisches Sampling von Eingaben. Dabei sucht das Verfahren nach minimalen Feature-Kombinationen, die ausreichen, damit das Modell fast immer dieselbe Entscheidung trifft Ribeiro, Singh, and Guestrin (2018).

3 Verwandte Arbeiten

Die Veröffentlichung von LIME durch Ribeiro et al. Ribeiro et al. (2016) markierte einen Meilenstein in der Forschung zu erklärbarer künstlicher Intelligenz. Das Verfahren war eines der ersten, das konsequent modellagnostisch konzipiert wurde, und eröffnete damit die Möglichkeit, komplexe Modelle unabhängig von ihrer Architektur lokal zu erklären. Mit SHAP legten Lundberg und Lee Lundberg and Lee (2017) eine Methode vor, die auf Shapley-Werten basiert und damit eine stärkere mathematische

Fundierung sowie bestimmte Fairness- und Konsistenzaxiome. Anchors wiederum erweitert den Ansatz um präzise Wenn-Dann-Regeln, die in ihrer Lesbarkeit Vorteile bieten, jedoch in der Abdeckung eingeschränkt sind.

Neben diesen Kernbeiträgen wurden in den letzten Jahren mehrere Übersichtsarbeiten veröffentlicht, die die Vielfalt der Methoden systematisieren und deren Vor- und Nachteile im Überblick darstellen. Hervorzuheben sind etwa die Arbeiten von Carvalho et al. Carvalho, Pereira, and Cardoso (2019), die verschiedene Metriken zu Bewertung von Interpretierbarkeit zusammentragen, sowie Holzinger et al. Carvalho et al. (2019), die den Begriff der "Causability" einführen, um die Qualität von Erklärungen aus einer nutzerzentrierten Perspektive zu bewerten. Diese Übersichtsarbeiten verdeutlichen, dass es kein universelles "bestes" Verfahren gibt, sondern dass die Eignung stark vom Anwendungs- und Nutzungskontext abhängt.

Darüber hinaus existieren eine Vielzahl von Fallstudien, die die praktische Anwendung von LIME in verschiedenen Domänen dokumentieren. Beispiele finden sich sowohl in der Medizin, etwa bei der Risikoprognose für COVID-19-Patienten Gabbay, Tsur, and Avram (2021), als auch in industriellen Kontexten wie der Prozessautomatisierung Upadhyay, Chakraborty, and Rajamani (2021). Solche Arbeiten zeigen, dass LIME trotz bekannter Limitierungen weiterhin in realen Szenarien genutzt wird. Parallel dazu wurden in jüngerer Zeit auch methodische Erweiterungen entwickelt, die spezifische Schwächen adressieren, etwa im Hinblick auf die Stabilität visueller Erklärungen Rahimiaghdam and Alemdar (2025).

Zusammenfassend lässt sich feststellen, dass die Forschung zu erklärbaren Methoden einerseits eine große methodische Diversifizierung hervorgebracht hat, andererseits aber zugleich zahlreiche Belege für die anhaltende Relevanz von LIME liefert. Eine detaillierte Analyse dieser Arbeiten und ihrer Befunde erfolgt in den folgenden Kapiteln entlang der definierten Vergleichskriterien.

4 Methodik

Die vorliegende Arbeit verfolgt ein qualitatives, literaturbasiertes Vorgehen. Anstelle von Implementierungen oder experimentellen Vergleichen wird eine systematische Analyse relevanter wissenschaftlicher Arbeiten durchgeführt. Dieses Vorgehen erlaubt, die Entwicklung, Anwendung und Bewertung von LIME im Kontext anderer Verfahren wie SHAP und Anchors aus einer breiten empirischen und theoretischen Basis heraus zu beleuchten.

Die Untersuchung orientiert sich an drei Forschungsfragen. Erstens, welche Stärken und Schwächen von LIME werden in der Forschung identifiziert? Zweitens, in welchen praktischen Domänen wird LIME eingesetzt und mit welchen Resultaten? Drittens, wie positioniert sich LIME im Vergleich zu SHAP und Anchors entlang zentraler Bewertungskriterien wie Verständlichkeit, lokale Treue, Stabilität, Effizienz, Abdeckung, Handlungsorientierung, medizinische und industrielle Nutzung sowie

Sicherheitsaspekte?

Zur Beantwortung dieser Fragen wurde eine systematische Literaturrecherche durchgeführt. Dabei wurden einschlägige Datenbanken wie ACM Digital Library, IEEE Xplore, SpringerLink, ScienceDirect, PubMed und arXiv Berücksichtigt. Der betrachtete Zeitraum erstreckt sich von 2014, also den ersten Vorläufern von LIME, bis 2025. Die Suchstrategie kombiniert Begriffe wie "LIME explainability", "SHAP interpretability" oder "Anchors rules" mit spezifischen Anwendungskontexten wie "medical", "industry" oder "finance".

Für die Auswahl wurden klare Kriterien angelegt, wie das Berücksichtigen von Beiträgen, die entweder systematische Übersichten, reproduzierbare Fallstudien oder empirische Evaluationen vorlegen. Rein theoretisch Positionspapiere oder Arbeiten ohne methodische Transparenz wurden ausgeschlossen. Die einbezogenen Publikationen decken sowohl methodische Grundlagenarbeiten als auch aktuelle Anwendungsstudien und kritische Untersuchungen ab.

Die Auswertung erfolgt in Form einer narrativen Synthese. Für jedes Vergleichskriterium wird zunächst die begriffliche Bedeutung umgerissen, anschließend werden relevante Befunde aus der Literatur zusammengetragen und kritisch diskutiert. So entsteht eine kohärente Vergleichsanalyse, die aufzeigt, in welchen Aspekten LIME weiterhin eine relevante Rolle spielt und wo andere Verfahren überlegen oder notwendige Ergänzungen darstellen. Dieses Vorgehen erlaubt es, trotz begrenzten Umfangs eine differenzierte und evidenzbasierte Bewertung vorzunehmen, die sich eng an den praktischen theoretischen Diskurs in der XAI-Forschung anlehnt.

5 Vergleichsanalyse

Die folgende Analyse vergleicht LIME, SHAP und Anchors entlang der neun Kriterien Verständlichkeit, lokale Treue, Stabilität/Robustheit, Laufzeit/Effizienz, Abdeckung, Handlungsorientierung, Einsatz in der Medizin, Einsatz in der Industrie sowie Sicherheit/Manipulationsrisiko.

5.1 Verständlichkeit

Verständlichkeit bezeichnet die Lesbarkeit und intuitive Nachvollziehbarkeit der Erklärung durch Nutzer, etwa Domänenexpert:innen oder Laien. LIME erzeugt lokale lineare Surrogatmodelle mit wenigen Features, so dass sich die wichtigsten Einflussfaktoren unmittelbar ablesen lassen. Tatsächlich dokumentieren Ribeiro et al., dass Nicht-Expert:innen mithilfe von LIME-Visualisierungen Modellfehler identifizieren und Modelle vergleichen konnten Ribeiro et al. (2016). Dies spricht für eine gute kognitive Zugänglichkeit, da Anwender:innen nur wenige gewichtete Features benötigen, um Schlussfolgerungen zu ziehen. Christoph Molnar weist ebenfalls darauf hin, dass man bei LIME typischerweise einen Parameter K (Anzahl der Features) wählt, und dass geringes K zu einer leichter Interpretierbaren Erklärung führt. Dieses Top-K Konzept

halt die Erklärungen übersichtlich Molnar (2024). Im Vergleich erzeugt SHAP Attributionen, die auf Spieltheorie basieren. Die SHAP-Werte besitzen intuitive Bedeutungen, etwa als positive oder negative Beiträge einzelner Features zur Vorhersage. Aufgrund der mathematischen Fundierung (additiv und konsistent, vgl. Abs. "Lokale Treue") können die Erklärungen jedoch für Laien zunächst abstrakt wirken. Obwohl es bisher keine Studie gibt, die eindeutig belegt, dass LIME-Erklärungen grundsätzlich verständlicher sind als SHAP, zeigen Nutzerstudien, dass die Verständlichkeit von SHAP-Erklärungen signifikant abnimmt, wenn Erklärungen für Proben nahe der Entscheidungsgrenze eines Modells gegeben werden Jalali, Haslhofer, Kriglstein, and Rauber (2023). Anchors stellt Erklärungen als präzise Wenn-Dann-Regeln dar. Diese Regeln sind semantisch sehr eingängig, und Molnar betont, dass Anchors leichtverständliche Wenn-Dann-Regeln liefert. Ebenfalls schreibt Molnar, dass Anchors intuitiv und leicht verständlich ist und eine klare Definition ihrer Gültigkeit hat Molnar (2024). Zusammengefasst zeigen die Quellen, dass LIME, aufgrund seiner reduzierten Feature-Darstellung, und Anchors, durch klare Regeln, oft schneller erfassbar sind als formellere Erklärungen. Dieses Befundmuster stützt die These, dass Lime- und Anchor-Erklärungen in der Regel weniger kognitive Überlastung verursachen als etwa komplex verschränkte Attributionen wie SHAP, solange die Nutzer die Repräsentation verstehen.

5.2 Lokale Treue

Unter lokaler Treue versteht man, wie gut eine Erklärung das Verhalten des zugrundeliegenden Modells in der kleinen Umgebung der betrachteten Instanz abbildet. LIME optimiert direkt auf lokale Surrogate, indem es zufällig veränderte Instanzen um den betrachteten Punkt erzeugt und ein lineares Modell daran anpasst. Ribeiro et al. selbst zeigen, dass die Qualität dieser lokalen Approximation stark von der Streuung der Samplingpunkte abhängt Ribeiro et al. (2016). Die lokale Treue von SHAP bedeutet, dass die Vorhersage eines Modells für eine bestimmte Instanz exakt durch die Summe der Shapley-Werte dieser Instanz wiedergegeben wird. Die Stabilität und Genauigkeit dieser lokalen Erklärungen hängen dabei von der genauen Berechnung der Shapley-Werte ab. Bei bestimmten Modelltypen, wie z.B. baumbasierten Modellen, erlaubt SHAP durch spezialisierte Algorithmen wie TreeSHAP eine effiziente Berechnung der Werte, ohne dass die theoretische Konsistenz verloren geht. In empirischen Vergleichen ergeben sich oft stabilere Rangordnungen der wichtigsten Merkmale mit SHAP als mit LIME, gerade wenn die Daten stark strukturiert sind Goldwasser and Hooker (2024). Bzgl. Anchors ist ein Anker lokal treu, wenn die Modellentscheidung für fast alle Instanzen, die die Bedingungen des Ankers erfüllen, gleich bleibt Ribeiro et al. (2018). Die lokale Treue beschreibt also, wie zuverlässig die Ankerregel die Vorhersage des Modells innerhalb ihres Gültigkeitsbereichs widerspiegelt. Gleichzeitig ist die Deckung des Ankers über alle Daten begrenzt, sodass die Regel nur auf einen Teil der Gesamtmenge zutrifft. Insgesamt sind die Quellen einhellig, LIME ist sachgerecht für lokale Surrogat-Erklärungen, aber seine Treue hängt stark von Sampling und Modellkomplexität ab. SHAP bietet mathematisch saubere Attributionswerte, die besonders bei Baum-Modellen sehr akkurat und reproduzierbar sind [10]. Anchors wiederum ist grundsätzlich lokal präzise, weil es nur dann gilt, wenn das Modell wirklich in der

beschriebenen Region stabil bleibt. Welche Methode die beste lokale Treue liefert, hängt stark vom Anwendungsbereich und dem zugrunde liegenden Modell ab.

5.3 Stabilität/Robustheit

Stabilität beschreibt, ob Erklärungen bei minimalen Änderungen wie anderen Perturbationen oder durch zufälliges Sampling konsistent bleiben. Zahlreiche Studien zeigen, dass LIME-Erklärungen sehr empfindlich auf Sample-Rauschen reagieren Burger, Chen, and Le (2023). Ghorbani et al. Ghorbani, Abid, and Zou (2019) belegten allgemein, dass Interpretationen bei kleinen Bildveränderungen stark schwanken können. Slack et al. Slack, Hilgard, Kamar, Singh, and Lakkaraju (2020) demonstrieren explizit, dass post hoc explanation techniques, die auf Eingabe-Perturbationen basieren, wie LIME und SHAP, nicht zuverlässig sind und gezielte Scaffolding-Angriffe es erlauben, die Erklärungen zu manipulieren, ohne Modellfehler zu korrigieren. In der Praxis bedeutet das, dass Standard-LIME bei identischen Bedingungen bei jedem Lauf leicht unterschiedliche Feature-Gewichte liefert. Anchors erzeugt lokale Regeln, die nur dann gelten, wenn das Modell in der beschriebenen Region stabil ist, und liefern somit eine intrinsisch hohe lokale Treue. Die Stabilität von Anchors ist innerhalb der Ankerregion hoch, da die Vorhersage für fast alle Instanzen in dieser Region gleich bleibt. Sie hängt weniger von zufälligen Perturbationen ab, sondern direkt von der Gültigkeit der definierten Ankerregion. Allerdings entstanden Gegenmaßnahmen für LIME. MindfulLIME Rahimiaghdam and Alemdar (2025) ist ein aktueller Ansatz, der gezielt alternative Samples erzeugt, beispielsweise graphbasiert oder über Unsicherheits-Sampling. Der Artikel berichtet, dass MindfulLIME die Konsistenz visueller Erklärungen im Vergleich zu zufallsbasierten Ansätzen erheblich verbessert. In einem medizinischen Bildvergleich erreichten die Autoren eine hundertprozentige Stabilität bei gleichbleibenden Bedingungen. Diese neueren Varianten zeigen, dass die Stabilitätsproblematik bekannt ist und adressiert wird.

5.4 Laufzeit/Effizienz

Die Effizienz umfasst die benötigte Rechenzeit und die Skalierbarkeit einer Methode. LIME ist bewusst leichtgewichtig gestaltet und kann einzelne Erklärungen in interaktiven Anwendungen innerhalb weniger Sekunden liefern. Im Gegensatz zu einigen tief lernenden Verfahren erfordert LIME nur so viele Modellabfragen, wie es die gewählte Stichprobengröße vorgibt. SHAP bietet verschiedene Varianten. KernelSHAP approximiert Shapley-Werte über zufälliges Sampling, was bei vielen Merkmalen sehr aufwendig sein kann. TreeSHAP arbeitet speziell für Baum-Modelle und liefert Ergebnisse in polynomieller Zeit, wodurch es für Entscheidungsbäume sehr schnell ist. Exakte Shapley-Berechnungen sind allgemein rechenintensiv, liefern dafür aber konsistente und mathematisch saubere Ergebnisse. Anchors nutzt eine explorative Suche, um Regeln zu finden. Dies kann aufwendig sein, weil viele Stichproben nötig sind, um Regeln mit hoher Präzision und ausreichender Abdeckung zu bewerten.

LIME ist für interaktive Nutzung und Prototyping meist besser geeignet, da nur ein schnelles Training eines lokalen Modells erforderlich ist. Für Batch-Analysen

oder Prüfungen mit vielen Datenpunkten lohnt sich oft der Mehraufwand, SHAP zu berechnen, insbesondere TreeSHAP bei Entscheidungsbäumen, da die Erklärungen dann konsistenter ausfallen. Anchors sollte eher für ausgewählte Instanzen oder Situationen verwendet werden, in denen leicht verständliche Regeln explizit benötigt werden, da die Suche nach maximal abdeckenden Regeln aufwendiger sein kann, besonders bei hochdimensionalen Datensätzen.

5.5 Abdeckung

Abdeckung meint hier, für wie viele Instanzen und Datentypen eine Methode im Alltag sinnvolle Erklärungen liefert. LIME und SHAP sind beide grundsätzlich datentypagnostisch und können für Tabellen-, Text- und Bilddaten eingesetzt werden, solange man eine entsprechende Repräsentation vorgibt. Sie liefern in der Regel für jede neue Instanz ein Erklärungsresultat. Anchors gibt nur dann eine Erklärung aus, wenn tatsächlich eine hochpräzise Regel gefunden wird. Für manche Instanzen kann es passieren, dass kein klarer Anker existiert, dann gibt es keine Erklärung. In Fallstudien zeigt sich dass LIME und SHAP auf den meisten Instanzen durchlaufen, auch wenn die Erklärungsgüte variieren kann. Anchors dagegen hat eine eingeschränktere Abdeckung. Es garantiert hohe Präzision, bricht aber bei Instanzen ab, für die sich keine präzise Bedingung formulieren lässt. Damit eignet sich Anchors vor allem, wenn man wenige, sehr robuste Regeln mit hoher Sicherheit benötigt. Soll man aber grundsätzlich für jeden Datenpunkt eine Erklärung ausgeben, sind LIME und SHAP die verlässlicheren Werkzeuge.

5.6 Handlungsorientierung

Handlungsorientierung bewertet, ob eine Erklärung konkrete Vorschläge liefert, wie sich das Ergebnis ändern ließe. LIME und SHAP liefern hauptsächlich Attributionen, also wie stark einzelne Merkmale zur Vorhersage beitragen, und sind damit eher diagnostisch. Sie können indirekt Hinweise für Handlungsmöglichkeiten geben, etwa dass ein bestimmtes Merkmal ausschlaggebend ist und man dort ansetzen könnte, liefern aber keine direkten Aktionspläne. Anchors kann unter Umständen operative Regeln anbieten wie “wenn A und B, dann Klasse”. Diese Regeln sind zwar nicht explizit handlungsweisend, da sie nur den Gültigkeitsbereich beschreiben, enthalten jedoch implizit Empfehlungen, etwa, “Wenn Feature X hoch ist, steigt die Wahrscheinlichkeit für Y, was Handlung Z nahelegen könnte”. Insgesamt zeigt sich, dass LIME und SHAP vor allem als Diagnosewerkzeug nützlich sind, aus deren Ergebnissen in Kombination mit Domänenwissen Hypothesen über relevante Variablen abgeleitet werden können. Anchors erzeugt verständliche Regeln, die lokal gültig sind und implizit Hinweise geben können, welche Bedingungen das Modellverhalten bestimmen. Damit liefern alle drei Methoden nützliche Einsichten, wobei LIME und SHAP stärker diagnostisch wirken und Anchors zusätzlich interpretierbare lokale Regeln bereitstellt.

5.7 Einsatz in der Medizin

Medizinische Anwendungen stellen hohe Anforderungen. Erklärungen müssen plausibel, robust und überprüfbar sein. Zahlreiche Studien setzen LIME in medizinischen

Pilotprojekten ein. Gabbay et al. Gabbay et al. (2021) beschreiben ein LIME-basiertes Modell zur Abschätzung der COVID-19-Schwere, bei dem LIME Ärzten schnell zeigte, welche Faktoren wie Symptome und Laborwerte zur Prognose führten. Solche Fallberichte belegen, dass LIME häufig als exploratives Unterstützungstool genutzt wird, um Behandlungshypothesen zu entwickeln. Gleichzeitig warnen Übersichtsarbeiten, dass LIME-Ausgaben in produktiven, regulierten Kontexten sorgfältig validiert werden müssen und zusätzliche Absicherungen wie Sensitivitätsanalysen oder Bootstrap-Intervalle erforderlich sind. Rahimiaghdam und Alemdar Rahimiaghdam and Alemdar (2025) adressieren die Stabilität in der medizinischen Bildanalyse am Beispiel von Chest-Xray und berichten, dass MindfulLIME im Vergleich zu Standard-LIME deutlich konsistentere Visualisierungen liefert. SHAP wird ebenfalls in medizinischen Szenarien eingesetzt. Kim et al. Kim (2025) entwickelten ein Modell zur Vorhersage von Appendixkrebs, indem sie SHAP zur Merkmalsauswahl und -gewichtung einsetzten. Dies verbesserte die Vorhersagegenauigkeit und die klinische Relevanz des Modells. In der Pharmakologie wird SHAP verwendet, um die Beziehungen zwischen Risikofaktoren zu verstehen und individuelle Risikofaktoren-Rangordnungen zu erstellen. Dies hilft, die Transparenz von maschinellen Lernmodellen in der Pharmakologie zu erhöhen REPROCELL (2024). Doch auch Anchors wird in der Medizin genutzt. Halpern et al. Halpern, Horng, Choi, and Sontag (2016) führten eine Methode ein, um klinische Zustände ohne gelabelte Daten zu schätzen, indem sie Anchors verwendeten. Diese Methode ermöglichte eine effiziente Identifizierung von statistisch getriebenen Phänotypen mit minimalem manuellen Aufwand. Zusammengefasst können LIME, SHAP und Anchors im Medizinbereich Hilfestellung leisten. LIME eignet sich besonders für erste Analysen und Prototypen, bei denen Schnelligkeit und Verständlichkeit wichtig sind. SHAP bietet mathematisch saubere Attributionswerte und erleichtert die Einschätzung wichtiger Merkmale. Anchors liefert lokal gültige Regeln, die das Modellverhalten interpretierbar machen.

5.8 Einsatz in der Industrie

In industriellen Umgebungen sind die Anforderungen an Erklärungen sehr unterschiedlich. Einerseits benötigt man schnelle Analysen und Fehlerdiagnosen, andererseits nachvollziehbare Aufzeichnungen für die Verwaltung. Industrielle Fallstudien zeigen ein gemischtes Bild. LIME wird häufig für Prototypen, Fehlersuche und schnelles Feedback an Beteiligte verwendet. Upadhyay und Kollegen entwickelten eine Erweiterung von LIME, um Geschäftsprozessmodelle zu erklären, da Standard-LIME wegen Reihenfolge- und Abhängigkeitsbeschränkungen oft versagte Upadhyay et al. (2021). In vielen Berichten heißt es, dass LIME gut für erste Iterationen und Gespräche ist. Es lässt sich leicht in bestehende maschinelle Lernprozesse integrieren und erfordert wenig Expertenaufwand. Für formelle Prüfungen und Regelkontrollen wird dagegen eher SHAP eingesetzt, insbesondere bei Baum- oder Regressionsmodellen. Manche Unternehmen kombinieren beide Methoden. LIME dient dem Erkunden und Kommunizieren im Team, SHAP dem Berichtswesen und Qualitätsmonitoring. Insgesamt zeigen die Quellen einen hybriden Ansatz. LIME ist beliebt für schnelle Analysen in dynamischen Umgebungen und seine Anpassbarkeit, zum Beispiel prozessbewusste Varianten, macht es pragmatisch. SHAP wird eingesetzt, wenn die Sicherheit

formaler Garantien wichtiger ist. Anchors spielt in der Industrie bisher eine kleinere Rolle, da echte Wenn-Dann-Regeln oft zu unflexibel sind, außer bei klar definierten transaktionalen Daten.

5.9 Sicherheit/Manipulationsrisiko

Sicherheit betrachtet, wie robust Erklärungen gegen Manipulationen oder unbeabsichtigte Fehlinterpretationen sind. Slack und Kollegen zeigen eindrücklich, dass man LIME- und SHAP-Erklärungen täuschen kann, indem man ein manipuliertes Gerüst im Modell einbaut Slack et al. (2020). Dabei wird zum Beispiel ein rassistisches Modell absichtlich so verändert, dass die Erklärungen harmlos erscheinen. Ghorbani und Kollegen warnten bereits, dass kleine Störungen in den Eingabedaten zu stark unterschiedlichen Erklärungen führen können, was die Verlässlichkeit generell in Frage stellt Ghorbani et al. (2019). Die Schlussfolgerung aus der Literatur ist eindeutig. Erklärungen allein können trügerisch sein. Man darf einer einzelnen LIME- oder SHAP-Erklärung niemals blind vertrauen. Forscher empfehlen stattdessen, mehrere Methoden zu kombinieren und die Konsistenz der Erklärungen zu prüfen. Nur so kann festgestellt werden, ob die Erklärung wirklich ein robustes Bild liefert oder nur zufällig wirkt. Sicherheit bleibt daher ein zentrales Risiko für alle nachträglichen Erklärmethoden, auch für LIME. Die Quellen raten daher zu einer Überprüfung der Erklärungen und zu einem gesunden Misstrauen gegenüber einzelnen Resultaten.

6 Diskussion

Die vergleichende Analyse zeigt, dass die Debatte um die Relevanz von LIME weniger in der Frage “ob” als vielmehr “wo” und “wie” es eingesetzt wird zu verorten ist. Einerseits weisen die Studien deutlich auf zentrale Einschränkungen hin. Instabilität bei wiederholten Läufen, Angreifbarkeit durch Manipulationen und fehlende Garantien in Bezug auf Konsistenz und Fairness. Diese Kritikpunkte sind weder marginal noch trivial sondern berühren das Kernproblem vieler nachträglicher Erklärmethoden. Die Gefahr besteht, dass die erzeugte Erklärung mehr suggeriert als sie tatsächlich abbildet. SHAP konnte hier durch seine mathematische Fundierung klare Vorteile demonstrieren während Anchors punktuell mit hoher Präzision überzeugt.

Gleichzeitig zeigt sich, dass LIME gerade wegen seiner pragmatischen Einfachheit, seiner breiten Anwendbarkeit und der niedrigen Einstiegshürden in Praxisprojekten weiterhin eine bedeutende Rolle spielt. In medizinischen und industriellen Fallstudien wird es regelmäßig genutzt nicht als alleinige Entscheidungsgrundlage sondern als exploratives Werkzeug zur Generierung erster Hypothesen oder zur schnellen Kommunikation mit Beteiligten. Dieser Befund deutet darauf hin, dass die Stärke von LIME weniger in formal abgesicherten Garantien liegt sondern in seiner Funktion als Brückenmethode zwischen komplexer Modelllogik und menschlichem Verständnis.

Interessant ist zudem, dass methodische Erweiterungen wie MindfulLIME zeigen, dass die Community die bekannten Schwächen adressiert. Dies deutet auf eine Verschiebung hin. Während SHAP die formale Konsistenz stärkt und Anchors eine hohe Präzision für Teilbereiche liefert, wird LIME durch iterative Weiterentwicklungen zunehmend stabiler und zuverlässiger. Das spricht dafür, dass sich der Diskurs von

einem Gegensatz zwischen LIME und modernen Verfahren zu einer Ergänzungslogik bewegt in der LIME seinen Platz als flexibles, niedrigschwelliges Werkzeug behauptet.

7 Fazit und Ausblick

Die Arbeit hat gezeigt, dass LIME aufgrund seiner Einfachheit, Modellagnostik und Flexibilität weiterhin ein wichtiges Werkzeug im Methodenspektrum der XAI bleibt. Auch wenn modernere Ansätze wie SHAP oder Anchors in einzelnen Dimensionen etwa mathematische Fundierung, Konsistenz oder Präzision überlegen sind, erfüllt LIME eine unverzichtbare Funktion. Es macht komplexe Modelle in kürzester Zeit zugänglich und unterstützt die Kommunikation zwischen Entwicklern, Fachanwendern und Entscheidungsträgern.

Für die Praxis bedeutet dies, dass LIME insbesondere in explorativen, iterativen und didaktischen Kontexten eine hohe Relevanz behält. Gleichzeitig darf es nicht unkritisch eingesetzt werden. Gerade in sensiblen Anwendungsfeldern wie Medizin oder Recht müssen die Grenzen von LIME wie Instabilität, Manipulierbarkeit und eingeschränkte Aussagekraft klar kommuniziert und durch ergänzende Verfahren oder Validierungsmethoden abgesichert werden.

Mit Blick auf die Zukunft zeichnet sich ab, dass sich die Rolle von LIME verändern wird. Einerseits werden Weiterentwicklungen wie MindfulLIME oder hybride Ansätze die Stabilität und Verlässlichkeit erhöhen. Andererseits könnte LIME zunehmend als Teil eines methodischen Ensembles eingesetzt werden. Als erste, schnell einsetzbare Erklärungsmethode wird es durch robustere Verfahren wie SHAP oder kontrafaktische Erklärungen ergänzt. Damit wird LIME weniger als Konkurrent moderner Verfahren verstanden, sondern als pragmatisches Bindeglied in einem Ökosystem erklärbarer Methoden.

Insgesamt verdeutlicht die Analyse, dass die Zukunft der XAI nicht in einer einzigen dominanten Methode liegt, sondern in der Kombination komplementärer Ansätze. LIME wird in diesem Ensemble weiterhin eine Rolle spielen, nicht trotz, sondern gerade wegen seiner Einfachheit.

References

Burger, C., Chen, L., Le, T. (2023). *Are your explanations reliable? investigating the stability of lime in explaining textual classification models via adversarial perturbation.* Retrieved from https://arxiv.org/abs/2301.00001

Carvalho, D.V., Pereira, E.M., Cardoso, J.S. (2019). *Machine learning interpretability: A survey on methods and metrics.* Retrieved from https://doi.org/10.3390/electronics8080832

Gabbay, F., Tsur, N., Avram, A. (2021). *A lime-based explainable ml model for predicting severity level of covid-19 patients.* Retrieved from https://doi.org/10.3390/app112110417

Ghorbani, A., Abid, A., Zou, J. (2019). Interpretation of neural networks is fragile. *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, , Retrieved from https://arxiv.org/abs/1710.10547

Goldwasser, J., & Hooker, G. (2024). *Provably stable rankings with shap and lime.* Retrieved from https://arxiv.org/html/2401.15800v2

Halpern, Y., Horng, S., Choi, Y., Sontag, D. (2016). *Electronic medical record phenotyping using the anchor and learn framework* (Vol. 23) (No. 4). Retrieved from https://pubmed.ncbi.nlm.nih.gov/27107443/

Jalali, A., Haslhofer, B., Kriglstein, S., Rauber, A. (2023). *Predictability and comprehensibility in post-hoc xai methods: A user-centered analysis.* Retrieved from https://doi.org/10.1007/978-3-031-37717-$4_4$6

Kim, J.Y. (2025). *Improving appendix cancer prediction with shap-based feature engineering for machine learning models: a prediction study* (Vol. 48) (No. 2). Retrieved from https://doi.org/10.12771/emj.2025.00297

Lundberg, S., & Lee, S.-I. (2017). *A unified approach to interpreting model predictions.* Retrieved from https://arxiv.org/abs/1705.07874

Molnar, C. (2024). *Interpretable machine learning* (2nd ed.). Leanpub. Retrieved from https://christophm.github.io/interpretable-ml-book/

Rahimiaghdam, S., & Alemdar, H. (2025). *Mindfullime: Improving the stability of lime explanations using graph-based and uncertainty sampling.* Retrieved from https://arxiv.org/abs/2501.01234

REPROCELL (2024, July 11). *Ai in drug discovery: The role of shap in pharmacology.* Retrieved from https://www.reprocell.com/blog/biopta/ai-in-drug-discovery-the-role-of-shap-in-pharmacology

Ribeiro, M.T., Singh, S., Guestrin, C. (2016). *"why should i trust you?": Explaining the predictions of any classifier.* Retrieved from https://arxiv.org/abs/1602.04938

Ribeiro, M.T., Singh, S., Guestrin, C. (2018). *Anchors: High-precision model-agnostic explanations.* Retrieved from https://doi.org/10.1609/aaai.v32i1.11699

Slack, D., Hilgard, S., Kamar, E., Singh, S., Lakkaraju, H. (2020). *Fooling lime and shap: Adversarial attacks on post hoc explanation methods.* Retrieved from https://arxiv.org/abs/1911.02508

Upadhyay, S., Chakraborty, S., Rajamani, K. (2021). *Extending lime for business process automation.* Retrieved from https://arxiv.org/abs/2108.04371