INTERPRETOVATELNOST A VYSVĚTLITELNOST SYSTÉMŮ UMĚLÉ INTELIGENCE

Jan Hendl

Abstrakt

Příspěvek podává přehled o tom, co je vysvětlitelná a interpretovatelná umělá inteligence (AI) a uvádí doporučení, která přispívají k zvýšení důvěry v AI a mohou vývojovým pracovníkům pomoci při vytváření důvěryhodných AI systémů.

Stručně se zabýváme modelově agnostickými metodami se zvláštním zaměřením na nejznámější z nich LIME a SHAP.

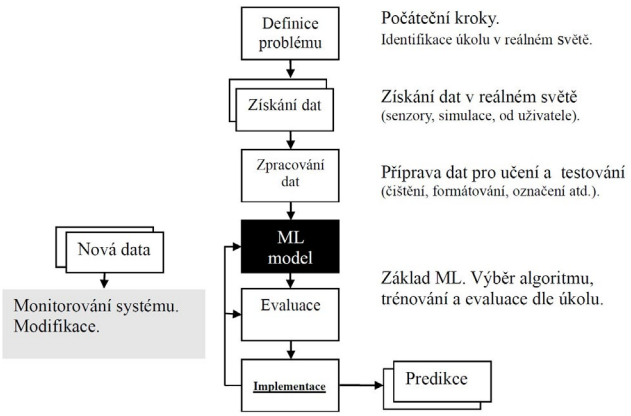
Klíčová slova

umělá inteligence, LIME, SHAP, XAI

Úvod

Umělá inteligence (AI, Artificial Intelligence) proniká do našeho života způsobem, který se před deseti či patnácti lety nezdál být možný [1]. V posledních letech se proměnila v sílu přetvářející profesní a osobní život lidí. Moderní výkonné aplikace založené na strojovém učení se staly téměř všudypřítomnou realitou. Počínaje expertními systémy první generace, které se pokusily řadou pravidel typu „IF… THAN …" zachytit a replikovat specifické znalosti expertů, se umělá inteligence velmi rozvinula. Vznikly třeba velké jazykové modely (LLM) a jejich aplikace, které jsou schopné odpovídat na rozličné otázky, konverzovat a generovat text, hudbu, obrázky a videa. V medicíně dochází k nárůstu aplikací umělé inteligence zejména v oblastech, jakými jsou lékařské zobrazování nebo kritická péče, kde se neustále generují rozsáhlé objemy dat. Mnoho odborníků však nesouhlasí s tím, že současnou umělou inteligenci lze spolehlivě používat při lékařském rozhodování. Pouhá nedůvěra v novinku nemusí být zodpovědná za tento stav.

Nejdříve přibližme pomocí obrázku proces tvorby hlavní části AI systému pomocí modelování založeném na strojovém učení (ML, Machine Learning), což je přístup současnosti. Obrázek 1 znázorňuje jednotlivé kroky od definice problému, který je třeba řešit, až po validaci a implementaci modelu. Každá aplikace ML podstatně závisí na tom, jak jsou definovány problémy v reálném světě, přičemž kroky získání dat a předběžného zpracování představují faktory, které významně ovlivňují přesnost i efektivitu modelů. Vygenerovaný model však nepředstavuje skutečný svět, ale pouze realitu dat.



Obrázek 1 – Vývoj a implementace modelu ML jako hlavní části AI systému [14].

AI systém předpokládá při aplikaci v reálném světě určité znalosti uživatelů. Opominutí splnění tohoto předpokladu může vést k odmítnutí AI systému. V tabulce uvádíme požadavky na potřebné znalosti uživatelů (lékařů), jak je formulovali Abgrall et al. [2].

1. Cíl a rozsah	
Účel	Primární cíl AI systému (např. predikce, diagnóza, doporučení).
Cílová populace	Demografická skupina pacientů, pro kterou je model určen.
2. Poznatky o systému	
Struktura	Stručný popis designu AI systému.
Vysvětlitelnost	Srozumitelnost výstupů modelu pro lékaře a pacienty.
Klíčové proměnné	Hlavní vstupní proměnné, které model používá, a jejich lékařský význam.
3. Zdroj dat	
Původ dat	Odkud pocházejí data pro učení systému a jeho validaci, což zajišťuje relevanci pro pacienty a lékaře.
Přizpůsobivost	Schopnost přetrénovat model pomocí lokálních datových souborů.
Otevřený přístup	Přístupnost k datům/kódu pro replikaci (např. na platformách jako GitHub).
4. Hodnocení a validace	
Metriky výkonu	Měřítka přesnosti modelu.
Benchmarking	Srovnání s jednoduššími a lépe interpretovatelnými systémy a modely.
Praktická validace	Testování v reálném klinickém prostředí, nejen retrospektivními daty.
5. Omezení modelu	
Problémy s výkonem	Situace nebo podmínky, kdy se účinnost systému může snížit.
Spolehlivost	Vyjádření spolehlivosti a nejistoty ve výsledcích systému.
Řízení chyb	Přístupy pro zpracování a opravu nepřesných výstupů.
6. Klinická integrace	
Lidský dohled	Zapojení člověka do rozhodování založeného na AI systému.
Integrace pracovních postupů	Přizpůsobení modelu stávajícím klinickým procesům
Uživatelská zkušenost	Návrh rozhraní a přehlednost informací.
Školení a vzdělávání	Vzdělávací zdroje poskytované pro zaměstnance a lékaře.
7. Etické aspekty	
Demografická spravedlnost	Konzistence výkonu napříč různými skupinami pacientů.
Audit spravedlnosti	Snaha o identifikaci a nápravu potenciálních bias a předsudků.

8. Regulační aspekty	
Ochrana osobních údajů a bezpečnost	Protokoly pro správu a ochranu údajů pacientů.
Dodržování právních předpisů	Soulad s předpisy jako GDPR, AI Act.
Odpovědnost lékaře	Odpovědnost při používání systému.
9. Údržba a audit	
Bezpečnostní kontroly	Sledování bezpečnosti a efektivity systému.
Aktualizace a vývoj	Udržování aktuálního modelu novými daty a poznatky.
10. Zpětná vazba a podávání zpráv	
Kanály zpětné vazby	Systémy pro sběr a řešení zpětné vazby od uživatelů.
Nežádoucí události	Postupy pro zpracování a hlášení jakýchkoli negativních výsledků spojených s nasazením systému.

podle Abgrall et al. 2024

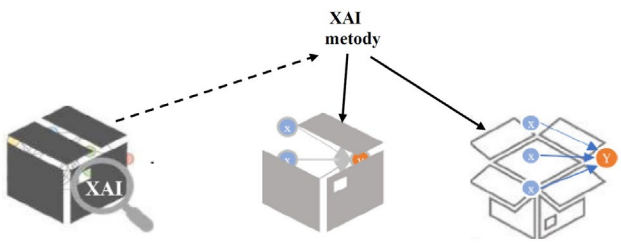
V tomto pojednání se věnujeme přednostně problému vysvětlitelnosti, interpretovatelnosti a souvisejících vlastností AI systémů [3].

Základní definice

Určitá část nedůvěry k rozhodnutím založených na AI ML algoritmech pramení z toho, že uživatelům připomínají tzv, „černé skříňky" (black box). Proto se zvyšuje význam kvality vysvětlitelnosti algoritmu v systému AI a význam přítomnosti relevantních prostředků k jejímu dosažení.

Poznamenejme, že již v první generaci expertních systémů (př. MYCIN) založených na pravidlech [4] se vypisovala použitá pravidla typu „IF… THAN …" pro daný případ, aby se uživateli poskytl vhled do práce expertního systému. Schéma expertního systému této generace obsahovalo vždy komponentu pro vysvětlení.

O vysvětlitelné umělé inteligenci (XAI, Explainable Artificial Intelligence) se dnes diskutuje v mnoha oblastech aplikací AI v reakci na vývoj moderních systémů AI a ML. Vznikající AI ML programy v současnosti jsou mimořádně složité (mají i miliardy parametrů) a někdy jsou dokonce zkreslené. Přesto činí rozhodnutí, která mohou rozhodujícím způsobem ovlivnit životy lidí. Příkladem takových složitých a komplexních AI systémů jsou prostředky založené na neurálních sítích v medicíně a aplikace pro automatické řízení automobilů. Pro uživatele představují neproniknutelnou tzv. černou skříňku (black box) a úkol představuje tuto černou skříňku trochu osvětlit a zprůhlednit (obrázek 2), v ideálním případě vytvořit bílou skříňku (white box).



Obrázek 2 – Proces tvorby XAI znázorněné pomocí metafory skříňky.

Vysvětlitelnou umělou inteligenci (XAI) můžeme vymezit jako soubor procesů a metod, které umožňují lidským uživatelům pochopit a důvěřovat výstupům vytvořených algoritmy strojového učení. Vysvětlitelná AI se používá k popisu užitého algoritmu AI, očekávaných výsledků a možných zkreslení. Má pomoci charakterizovat správnost (accuracy), průhlednost (transparency), spravedlnost (fairness), robustnost a výstupy/výsledky (outcomes) rozhodnutí doporučených/provedených AI systémem [3].

Zmínili jsme vlastnosti, které souvisí s XAI. Jejich cílem je posílit důvěru v navržený systém. Standardně se klade důraz na kvalitu systému s ohledem na jeho klasifikační a predikční schopnosti. V této souvislosti se používají různé statistické koeficienty, jejichž hodnoty se získaly v průběhu testování systému. Některé vlastnosti však nelze jednoduše kvantifikovat.

Například „spravedlnost" (férovost) vychází z předpokladu, že v důsledku zkreslení v učebních souborech dat a v navržených algoritmech může být s některými skupinami jednotlivců zacházeno nespravedlivě, mohou být třeba diskriminovány. Spravedlnost se týká schopnosti modelu činit nezaujatá rozhodnutí, aniž by upřednostňoval některou z populací zastoupených ve vstupní distribuci dat. Systémy umělé inteligence mohou být ovlivněny různými způsoby. V modelech umělé inteligence by neměla hrát roli zkreslení, jako jsou místo narození, pohlaví, rasa a socioekonomický status.

Transparentnost (transparency) znamená jasnost a otevřenost, jak pracuje a dělá rozhodnutí AI systém. Citlivost výstupu systému na změnu vstupu se zachycuje robustností. Posuzujeme schopnost modelu správně fungovat v případě určité nejistoty ve vstupních datech. Chování systému by nemělo být ovlivněno malými změnami ve vstupech.

Literatura o umělé inteligenci nabízí různé definice vysvětlitelnosti, zdůrazňuje se v ní, že formální definice chybí. Platí to i pro interpretovatelnost. Někdy se vysvětlitelnost ztotožňuje s interpretovatelností. Specialisté však rozlišují mezi oběma termíny. Správnější je podle nich považovat interpretovatelnost za podřazenou vysvětlitelnosti nebo dokonce dělat mezi těmito termíny rozdíl.

Lze definovat vysvětlitelnost a interpretovatelnost AI systému také takto:

Vysvětlitelnost se týká schopnosti systému nabídnout vysvětlení svých rozhodnutí způsobem, který je pro člověka srozumitelný. V důsledku toho se vysvětlitelná umělá inteligence zaměřuje na objasnění aspektu **„proč"** rozhodování s pomocí umělé inteligence s cílem zvýšit transparentnost a srozumitelnost pro člověka.

Na druhou stranu **interpretovatelnost** zdůrazňuje aspekt **„jak"** rozhodování. Odkazuje na schopnost systému odhalit proces, který se použil k dosažení rozhodnutí, a ukázat důvody svých rozhodnutí.

To znamená zabývat se konkrétními výpočty, matematickými funkcemi a příspěvky jednotlivých vstupních proměnných, které ovlivnily predikci výstupu systému.

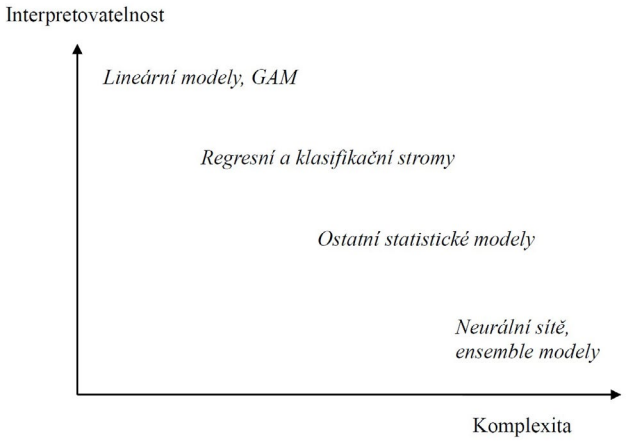
Interpretovatelné systémy jsou považovány za vysvětlitelné pro cílové uživatele, pokud jsou jejich procesy pro tyto uživatele snadno srozumitelné. Na druhou stranu člověk nemusí nutně rozumět mechanistickému **„jak"**, znát přesně vzorce, které objasňují predikce modelu.

Poznamenejme, že často se spíše než termín „systém AI" používá v odborné literatuře termín „model AI", který je specifičtější. AI modely neboli modely umělé inteligence jsou programy, které detekují specifické vzorce dat. Jedná se o realizaci algoritmu, který může přijímat datové vstupy a vyvozovat závěry nebo provádět akce v závislosti na těchto závěrech.

Koncepty interpretovatelnosti a vysvětlitelnosti lze přiblížit třemi rozdíly:

1. Úroveň detailů: **Interpretovatelnost** se zaměřuje na pochopení vnitřního fungování modelů, zatímco **vysvětlitelnost** se zaměřuje na vysvětlení učiněných rozhodnutí. V důsledku toho vyžaduje interpretovatelnost větší míru podrobnosti o systému a při popisu než vysvětlitelnost.
2. Složitost modelu: Složitější modely umělé inteligence, jako jsou hluboké neuronové sítě, mohou být obtížně interpretovatelné kvůli své složité struktuře a interakcím mezi různými částmi modelu. V těchto případech může být vysvětlitelnost schůdnější, neboť se zaměřuje spíše na vysvětlování rozhodnutí než na pochopení samotného modelu.
3. Komunikace: Interpretovatelnost se týká chápání modelu odborníky a výzkumnými pracovníky zabývajícími se umělou inteligencí, zatímco vysvětlitelnost je více zaměřena na sdělování rozhodnutí o modelu konečným uživatelům. Vysvětlitelnost proto vyžaduje jednodušší a intuitivnější prezentaci informací.

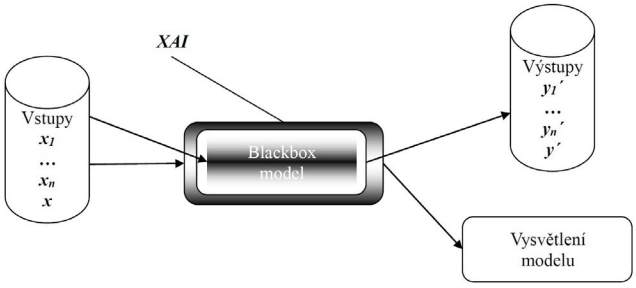
Některé algoritmy použité v AI modelech (viz obrázek 3) vykazují ze své podstaty vysokou interpretovatelnost (transparentní modely nebo „white-box" modely), zatímco jiné nejsou okamžitě srozumitelné („black-box" modely), což vyžaduje doplnění o lokální interpretaci pro daný případ (Post-hoc vysvětlení) nebo globálně pro všechny případy. Tato vysvětlení mohou být použitelná pro konkrétní model nebo univerzálně pro všechny použité modely (agnostické metody).



Obrázek 3 – Vztah mezi interpretovatelností a komplexitou ML modelu.

Globální vysvětlení: „Proč se model obecně rozhoduje tak, jak se rozhoduje?"

Globální vysvětlení se zabývá rozhodovacím procesem na makroúrovni a nabízí vhled do jeho základních mechanismů a obecných strategií (viz obrázek 4).

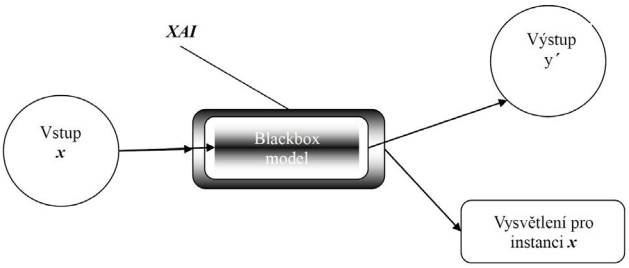


Obrázek 4 – Globální vysvětlení se zaměřuje na hodnocení výstupů pro učební soubor a budoucí instance.

Lokální vysvětlení: „Proč model učinil toto konkrétní rozhodnutí pro tento aktuální případ?"

Lokální vysvětlení se zaměřuje na zdůvodnění a objasnění fungování systému pro jednotlivý případ (viz obrázek 5).

Postupy pro XAI mají několik možností zaměření.



Obrázek 5 – Lokální vysvětlení se zaměřuje na hodnocení jedné instance.

Vizualizace

Vizualizace dat a modelů může pomoci zjednodušit pochopení fungování modelů umělé inteligence. Tepelné mapy (heat maps) lze použít k vizualizaci důležitosti různých proměnných (vlastností) v rozhodovacím procesu modelu.

Dekompozice

Rozložení modelu na jednodušší komponenty může usnadnit pochopení jeho fungování. Rozložení klasifikačního modelu na jednotlivé binární klasifikátory může například usnadnit pochopení rozhodovacího procesu modelu.

Vysvětlení pomocí příkladu

Dalším přístupem ke zlepšení vysvětlitelnosti je poskytnout vysvětlení založená na příkladech. To znamená ukázat uživatelům příklady vstupů podobné tomu, který je předmětem úvah, a vysvětlit, jak daný model v těchto případech rozhodoval.

Kdy není vhodné vysvětlení?

Ne všechny predikce AI by měly být vysvětleny. Generování vysvětlení přináší další složitost softwaru a výpočetní nároky. Implementace XAI může být v určitých scénářích neopodstatněná. Uvedeme příklady.

Nerizikové scénáře: V situacích, kdy jsou důsledky chyb minimální, jako jsou například doporučení na streamovací platformě. Uživatelé mají obvykle malý zájem o pochopení vnitřního fungování toho, jak jsou tyto návrhy generovány.

Dobře pochopené problémy: U dobře pochopených problémů, jako je filtrování spamu v e-mailových systémech, nemusí být vysvětlení každého rozhodnutí zásadní.

Rizika manipulace: V některých případech může poskytnutí podrobného vysvětlení neúmyslně pomoci těm, kteří se snaží systém zneužít. Hackeři mohou například zneužít podrobné informace o vyhledávacích algoritmech k manipulaci a narušení integrity systému.

Ante-hoc vs Post-hoc interpretovatelnost

Ante-hoc (většinou globální interpretovatelnost): Tento přístup se zaměřuje na používání modelů, které jsou ze své podstaty interpretovatelné díky jejich jednoduché struktuře a transparentním rozhodovacím procesům. Příklady zahrnují, jak jsme zmínili, lineární modely, rozhodovací stromy a systémy založené na pravidlech (viz také obrázek 3). I když tyto modely nabízejí jasnost od samého počátku, jejich interpretovatelnost se může snižovat s rostoucí složitostí nebo počtem parametrů.

Post-hoc (většinou lokální interpretovatelnost): Post-hoc metody mají za cíl vysvětlit rozhodnutí učiněná složitými modely poté, co byly naučeny daty. Tato vysvětlení nejsou zabudována do modelu, ale jsou odvozena z analýzy chování modelu. Tento přístup je zvláště užitečný pro modely černé skříňky, jako jsou hluboké neuronové sítě. Post-hoc vysvětlení se snaží být lokálně věrná, což znamená, že přesně odrážejí rozhodování modelu pro konkrétní případy nebo rozhodnutí.

Navržené postupy a jejich klasifikace

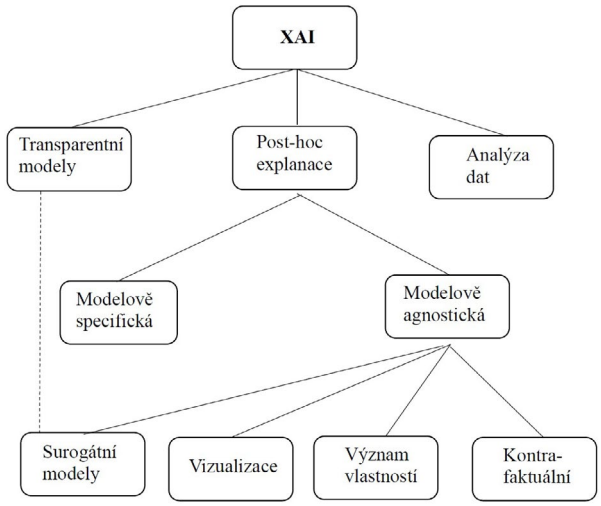
Výzkumníci a vědci navrhli různé taxonomie a seznamy druhů vysvětlení. Nejdříve vyjmenujeme formální prostředky vysvětlení.

Formát vysvětlení (jak je vyjádřeno): Vizualizace (např. Teplotní mapy) – Text (výroky, vyprávění nebo příběhy, odpovědi na dotazy, dialogy mezi člověkem a strojem) – Formální vyjádření (logické výrazy, matice) – Konceptuální modely procesů (diagramy) – Grafy, sítě – Tabulky – Abstrakce, zobecnění – Časové osy – Hierarchie (stromy).

Reference (čeho se formálně týká vysvětlení): Příklady (zahrnují chybné klasifikace, protipříklady, odlehlé hodnoty, jasné případy, blízcí konkurenti) – Vzory, třídy, ontologie – Vlastnosti, váhy, pravděpodobnosti, pořadí, parametry – Rozhodnutí, strategie, cíle – Algoritmy, výpočetní procesy, důkazy – Příhody, události (zahrnuje vlastní vysvětlení nebo příběhy) – Vztahy příčina–následek.

Je třeba poznamenat, že vysvětlení může zahrnovat smíšené typy jmenovaných technik.

Uvedeme všeobecně uznávanou taxonomii postupů XAI pro zvýšení kvality interpretovatelnosti a vysvětlitelnosti modelu AI a jejich popisy (viz obrázek 6).



Obrázek 6 – Klasifikace XAI podpory [2].

Specifické prostředky patří mezi Post-hoc postupy (pro daný případ), které dělíme na závislé na modelu a modelově agnostické. Transparentní modely se používají také v surogátních postupech.

Dále stručně popíšeme a charakterizujeme hlavní metody používané v rámci podpory vysvětlitelnosti umělé inteligence (SHAP, LIME, náhradní stromy atd.), uvedeme jejich vlastnosti. Nejedná se o všechny metody, které existují. Metodám SHAP a LIME se vzhledem k jejich významu věnujeme poněkud více. Instance v našem výkladu označuje daný případ, o kterém AI rozhoduje.

Fungování nástrojů SHAP a LIME ilustrujeme na příkladu. Použijeme datový soubor o onemocnění cukrovkou od společnosti Kaggle DIABETES2, obsahující hodnoty parametrů od 9000 pacientů, z nichž někteří měli prokázaný diabetes. Nejdříve se sestrojil klasifikátor pro zařazení pacienta na základě jeho dat do třídy s diabetem (label 1) nebo bez něho (label 0) pomocí metody Náhodné lesy. Tato metoda má dobré klasifikační vlastnosti, ale zároveň charakter černé skříňky (upraveno podle DataCamp, www.datacamp.com) .

SHAP (ShapleyAdditiveExplanation)

Procedura SHAP, kterou navrhli Lundberg a Lee [6], je založena na konceptu Shapleyho hodnot z teorie kooperativních her. Hodnoty Shapley poskytují způsob, jak rozdělit „výplatu" mezi hráče kooperativní hry na základě jejich individuálních příspěvků. V kontextu strojového učení bere SHAP každou proměnnou (vlastnost, pixel) jako „hráče" v kooperativní hře a předpověď je „výplata". Hodnota Shapley proměnné (vlastnosti) představuje její příspěvek k predikci při zvážení všech možných podskupin vlastností.

Klíčovými kroky při výpočtu Shapley hodnot jsou:

1. Generování podskupin: SHAP zvažuje všechny možné podskupiny vlastností. Pro model s n vstupními vlastnostmi existuje 2^n možných podmnožin.
2. Výpočet mezního příspěvku: Pro každou podskupinu vypočítá SHAP mezní příspěvek každé proměnné (vlastnosti) porovnáním předpovědi modelu s danou vlastností a bez ní.
3. Výpočet Shapley hodnoty: Hodnota proměnné (vlastnosti) Shapley je průměr jejích příspěvků napříč všemi možnými podskupinami. Tato hodnota představuje celkový příspěvek dané vlastnosti (vstupní proměnné) k předpovědi.

SHAP má několik žádoucích vlastností:

- Aditivita: Součet Shapley hodnot pro všechny vlastnosti se rovná rozdílu mezi predikcí modelu a průměrnou predikcí pro datový soubor. Tato vlastnost zajišťuje úplnost vysvětlení a nevynechává žádné důležité vlastnosti.
- Konzistence: Pokud se model změní tak, že příspěvek vlastnosti se zvýší nebo zůstane stejný bez ohledu na ostatní vlastnosti, Shapley hodnota pro tuto vlastnost se nesníží. Tato vlastnost zajišťuje, že vysvětlení jsou v souladu s chováním modelu.
- Účinnost: SHAP poskytuje jednotný rámec pro interpretaci předpovědí jakéhokoli modelu strojového učení. Existují účinné metody aproximace pro některé třídy modelů ML, jako jsou stromové modely a hluboké neuronové sítě.

SHAP má však také některá omezení:

- Komplexní výpočet: Výpočet Shapley hodnot je výpočetně náročný, protože vyžaduje zvážení všech možných podskupin vlastností. U modelů s velkým počtem vlastností (proměnných) jsou nezbytné metody aproximace.
- Interpretační výzvy: Shapley hodnoty sice poskytují měřítko důležitosti proměnné (vlastnosti), ale interpretace těchto hodnot může být v některých případech náročná, zejména pokud vlastnosti jsou v interakci.

SHAP lze používat globálně i lokálně.

SHAP standardně nabízí řadu vizualizačních nástrojů pro zlepšení interpretovatelnosti modelu. Uvedeme dva z nich (upraveno podle DataCamp, www.datacamp.com): (1) graf důležitosti proměnné s vyznačením dopadu a (2) graf závislosti pro proměnnou (vlastnost).

1. Graf důležitosti proměnných s vyznačením dopadu proměnné na klasifikaci

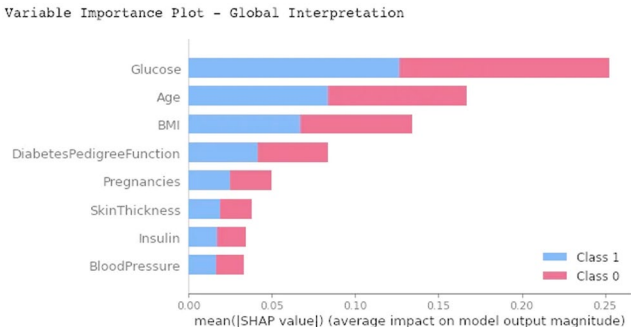
V tomto grafu jsou proměnné seřazeny podle jejich průměrných Shapley hodnot, které zobrazují nejdůležitější vlastnosti nahoře a ty nejméně důležité dole. To pomáhá pochopit vliv každé vlastnosti (proměnné) na předpověď modelu.

LIME (Local Interpretable Model-agnostic Explanations)

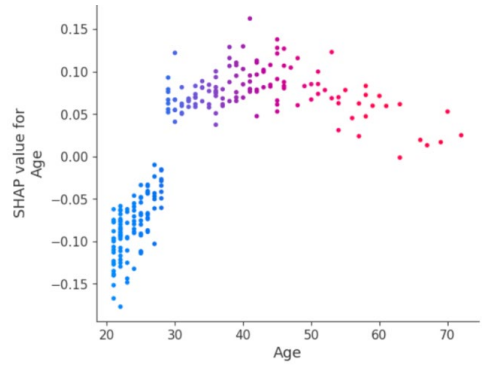
Technika LIME, kterou popsali Ribeiro a kol. [5], poskytuje vysvětlení jednotlivých predikcí vytvořených ML modely. Klíčovou myšlenkou LIME je přiblížení komplexního modelu jednodušším, interpretovatelným modelem v okolí konkrétní instance.

Proces generování vysvětlení LIME je tvořen následujícími kroky:

1. Pertubace: LIME generuje sadu modifikovaných instancí mírnou úpravou hodnot vlastností zkoumané instance,



1. Graf důležitosti – Můžeme vidět, že modrá a růžová zabírají polovinu vodorovných obdélníků pro každou třídu. To znamená, že každá vlastnost má stejný dopad na klasifikaci případů diabetu (label=1) i nediabetických (label=0). Glykemie, věk a BMI jsou proměnné s největší prediktivní schopností. Na druhou stranu těhotenství, tloušťka kůže, hladina inzulínu a krevní tlak nepřispívají tolik jako první tři vlastnosti.



2. Graf závislosti (barevně lze odlišit případy–instance hraniční, s diabetem a bez diabetu) – Ukazuje, že pacienti do 30 let mají nižší riziko diagnózy diabetu (body na levé straně obrázku). Naproti tomu jedinci starší 30 let musí počítat s vyšší pravděpodobností diagnózy diabetu (body ve středu a na pravé straně obrázku).

která má být vysvětlena. Tyto pertubace (poruchy) vznikají náhodným zapnutím či vypnutím vlastností.

2. Predikce: Model černé skříňky se používá k vytváření předpovědí pro modifikované instance.
3. Vážení: Modifikované instance se váží na základě jejich blízkosti k původní instanci pomocí metriky vzdálenosti (např. kosinová vzdálenost pro text, euklidovská vzdálenost pro obrázky).
4. Odhad jednoduchého modelu: Jednoduchý a přirozeně interpretovatelný model (např. lineární regrese, rozhodovací strom) je trénován na vážených modifikovaných instancích, aby se přiblížilo chování modelu černé skříňky lokálně k zkoumané instanci, která má být vysvětlena.
5. Vysvětlení: Jako vysvětlení původní předpovědi slouží koeficienty u vlastností (proměnných) interpretabilního modelu (u lineární regrese). Tyto koeficienty udávají podíl každé vlastnosti na předpovědi.

LIME má několik výhod, které z něj činí oblíbenou volbu pro vysvětlitelné AI:

- Model-agnostický: LIME lze aplikovat na jakýkoli model strojového učení bez ohledu na jeho architekturu nebo trénínkový algoritmus. Tato flexibilita umožňuje použití LIME s celou řadou modelů, včetně hlubokých neuronových sítí.
- Lokální vysvětlení: LIME poskytuje vysvětlení pro jednotlivé předpovědi, které mohou být v mnoha scénářích užitečnější než globální vysvětlení. Lokální vysvětlení pomáhají uživatelům pochopit, proč byla konkrétní instance klasifikována určitým způsobem.
- Interpretační vysvětlení: Pomocí jednoduchých, interpretačně zvládnutelných modelů, které přibližují místní chování modelu černé skříňky, LIME generuje vysvětlení, která jsou pro lidské uživatele snadno srozumitelná.

LIME má však také některá omezení:

- Nestabilita: Vysvětlení poskytnutá programem LIME mohou být citlivá na náhodné odchylky a volbu interpretačního modelu. Různé běhy LIME mohou pro stejnou instanci přinést mírně odlišná vysvětlení.
- Komplexnost výpočtů: Vytváření vysvětlení LIME může být výpočetně náročné, zejména u velkých datových souborů a složitých modelů. Potřeba dělat předpovědi pro více modifikovaných případů zvyšuje výpočetní náročnost.

LIME se aplikuje lokálně.

Ilustrační příklad (upraveno podle DataCamp, www.datacamp.com) zobrazuje standardní vysvětlení LIME pro 8. instanci v testovacích datech zařazenou pomocí klasifikátoru Náhodné lesy a prezentuje konečný příspěvek proměnných v tabulkovém formátu.

Popis i srovnání SHAP a LIME a problémy spojené s nasazením těchto postupů najdeme v mnoha pracích [př.: 7, 8, 9, 10].

Další vyvinuté a používané postupy i příslušnou literaturu podrobně rozebírá Molnár [10] ve své na internetu dostupné knížce.

Graf parciální závislosti (Partial Dependence Plot)

Graf parciální závislosti (zkrácené PDP nebo PD graf) ukazuje marginální vliv jedné nebo dvou vlastností na předpokládaný výsledek modelu strojového učení. Graf parciální závislosti může ukázat, zda je vztah mezi výstupem a vlastností (vstupní proměnnou) lineární, monotónní nebo složitější. U metody zvýšení interpretovatelnosti založené na pertubaci (modifikaci) je provedení relativně rychlé. PDP předpokládá nezávislost vlastností a může být zavádějící, pokud tento předpoklad není splněn.

Lze aplikovat pouze globálně.

Akumulované lokální efekty (ALE, Accumulated Local Effects)

Accumulated Local Effects (ALE) je metoda pro výpočet efektů vlastností. Algoritmus poskytuje modelově agnostická globální vysvětlení klasifikačních a regresních modelů pro tabulková data. ALE řeší některé klíčové nedostatky grafů parciálních závislostí (PDP).

Tuto metodu lze použít pouze globálně.

Kotvy (Anchors)

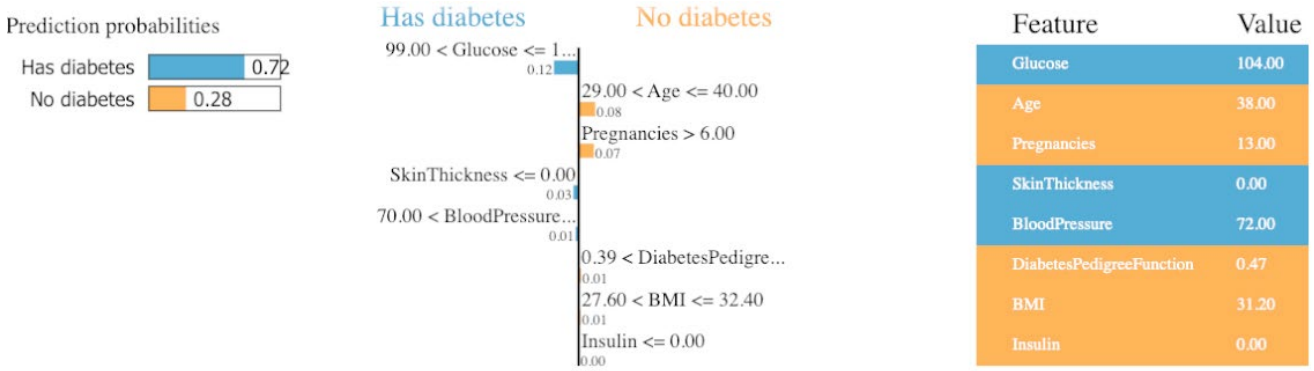
Za představou kotev je nápad vysvětlit chování složitých modelů pomocí vysoce přesných pravidel nazývaných kotvy. Tyto kotvy jsou lokálně postačující podmínky pro zajištění určité predikce s vysokou mírou spolehlivosti.

Z toho plyne, že metodu lze aplikovat pouze lokálně.

Metoda kontrastního vysvětlení (CEM, Contrastive Explanation Method)

CEM generuje lokální vysvětlení pro daný případ v klasifikačních modelech z hlediska relevantních pozitiv (PP, Pertinent Positives) a relevantních negativ (PN, Pertinent Negatives). Zdůrazňuje nejen to, co by mělo být minimálně a dostatečně přítomno, aby ospravedlnilo klasifikaci vstupního příkladu neuronovou sítí (příslušná pozitiva), ale také to, co by mělo minimálně a nutně chybět (příslušná negativa), aby se navrhlo úplnější a kvalitnější vysvětlení.

CEM se hodí pro lokální použití.



LIME Výsledek obsahuje tři hlavní informace (číslované zleva doprava): (1) předpovědi modelu, (2) příspěvky parametrů (proměnných) a (3) skutečná hodnota každého parametru (proměnné). Můžeme nahlédnout, že u pacienta se předpokládá diabetes se 72% spolehlivostí. Důvody, které vedly model k tomuto rozhodnutí, jsou následující: Hladina glukózy u pacienta je vyšší než 99, krevní tlak je vyšší než 70. Tyto hodnoty lze ověřit z tabulky vpravo.

Kontrafaktuální případ (Counterfactual Instance)

Kontrafaktuální vysvětlení „dotazují" model, aby ukázaly, jak moc by se musely změnit hodnoty jednotlivých vlastností, aby se změnila predikce. Kontrafaktuální vysvětlení výsledku nebo situace má formu „Kdyby nenastalo X, pak by nenastalo Y".

„Counterfactual Instance" metoda je navržena pro lokální aplikaci.

Integrované gradienty (Integrated Gradients)

Integrované gradienty mají za cíl přiřadit hodnotu důležitosti každé vstupní vlastnosti modelu strojového učení na základě gradientů výstupu modelu s ohledem na vstup. Má mnoho příkladů použití, včetně snah o pochopení významu vlastností, identifikace zkreslení dat a ladění práce modelu.

Integrované gradienty jsou navrženy pro lokální použití.

Globální interpretace prostřednictvím rekurzivního dělení (Global Interpretation via Recursive Partitioning, GIRP)

Kompaktní binární strom, který globálně interpretuje modely ML tím, že představuje nejdůležitější rozhodovací pravidla implicitně obsažená v modelu pomocí matice příspěvků vstupních proměnných.

GIRP lze aplikovat pouze globálně.

Protodash

Novější přístup k hledání „prototypů" pro existující program strojového učení. Prototypem se myslí podmnožina dat, která mají silný vliv na prediktivní schopnost modelu. Smyslem prototypu je říci něco v tom smyslu, že pokud bychom odstranili tyto datové body, model by tak dobře nefungoval, takže člověk může pochopit, co nejvíce ovlivnilo predikci.

Protodash lze aplikovat pouze lokálně.

Náhradní stromy (Tree Surrogates)

Tree Surrogates je interpretovatelný model, který je trénován k aproximaci předpovědí složitých modelů. Můžeme vyvodit závěry o původním složitém modelu tím, že interpretujeme náhradní model. Stromy jsou snadno interpretovatelné a poskytují kvantitativní předpovědi budoucího chování.

Tree Surrogates lze použít globálně i lokálně.

Dodatečné využití metod XAI

Postupy pro zvýšení vysvětlení a interpretovatelnosti jsou vhodné i v případech, které nesouvisí přímo s prací daného AI modelu a systému AI. Zmíníme některé možnosti.

Ladění a vylepšování modelů. Prediktivní modely dělají chyby. Pochopení základních příčin, které vedou k těmto nepřesnostem, je nezbytné pro zvýšení výkonu modelu. Techniky XAI nám pomáhají pochopit důvody předpovědí a pomáhají nám charakterizovat situace, kdy model dělá chyby. Díky tomu je XAI cenným nástrojem pro ladění modelů a přispívá ke zpřesňování modelu.

Detekce a zmírnění předsudků. Prediktivní modely mohou být zkreslené, což vede k sociálním nerovnostem a poškozuje menšinové skupiny. Algoritmické zkreslení obvykle pochází z používání zkreslených nebo neúplných dat nebo z předpokladů zabudovaných v modelech.

Podobné případy algoritmického zkreslení byly zdokumentovány v procesu soudního systému a procesu žádostí o zaměstnání. Tyto příklady zdůrazňují potřebu odpovědnosti v systémech umělé inteligence, zejména při nasazení v oblastech s významnými společenskými dopady.

Techniky XAI mohou pomoci identifikovat takové zkreslení tím, že kontextualizují, proč modely vytvářejí určité předpovědi.

Scénáře s vysokým rizikem. Chybné nebo zaujaté předpovědi mohou mít hluboký dopad na životy a pohodu jednotlivců, zejména při použití ve vysoce rizikových scénářích. Například při použití prediktivních algoritmů pro systémy sociálního bodování, individuální kategorizaci, technologie rozpoznávání obličejů, nástroje pro vymáhání práva, procesy prověřování zaměstnání a poskytování základních služeb, jako je zdravotní péče. V praktických aplikacích může být vysvětlení stejně důležité jako samotná předpověď.

Soulad s předpisy a zákony. Jak se zavádění umělé inteligence rozšiřuje, zákony a předpisy začínají vyžadovat, aby rozhodnutí umělé inteligence byla transparentní a vysvětlitelná. Cílem je zaručit, že algoritmy budou fungovat v rámci etických hranic. Například obecné nařízení Evropské unie o ochraně osobních údajů (GDPR) vyžaduje značnou transparentnost ohledně logiky automatizovaných rozhodnutí, zejména u rozhodnutí, která významně ovlivňují jednotlivce, jako jsou rozhodnutí týkající se zaměstnání, bonity a právních záležitostí.

Objevování znalostí. Použití XAI pro generování nových znalostí je pravděpodobně nejvíce přehlíženou aplikací. Tato schopnost je zvláště zajímavá vzhledem k tomu, že modely strojového učení jsou vhodnější pro identifikaci složitých vzorců a trendů v datech ve srovnání s tradičními statistickými testy, jako jsou t-testy nebo testy chí-kvadrát.

Vyhodnocení vysvětlení a interpretace

Tento odstavec se opírá o závěry autorů [16, s. 95]. XAI vysvětlení a interpretace jsou určeny pro člověka a k jejich správnému vyhodnocení jsou zapotřebí koncepty behaviorálních věd. Pro

hodnocení systémů XAI je k dispozici řada metod a ne všechny systémy XAI musí být nutně hodnoceny výzkumem jejich uživatelů (tedy lidí). Podle autorů je však hodnocení výkonu uživatele při skutečném používání systému rozhodujícím činitelem úspěchu XAI. Výzkum s využitím lidských účastníků je náročný na zdroje a čas. Jsou s ním také spojeny značné metodologické problémy nebo problémy s měřením:

- Vysvětlení generovaná XAI lze hodnotit z hlediska kritérií správnosti a výsledky korelovat s výkonem uživatelů.
- Lze hodnotit porozumění vysvětlení ze strany uživatelů a výsledky korelovat s kvalitami jejich mentálních modelů a s jejich výkonností.
- Znalosti nebo mentální modely uživatelů lze měřit nebo nějak reprezentovat a výsledky korelovat se změnou výkonu, kterou lze připsat procesu vysvětlování.

Je možné si představit celou řadu měr výkonnosti a mnoho z nich bylo také ve výzkumu využito. Představují výzvu pro tvorbu operacionalizovaných definic a výzvu pro experimentální design.

- Úspěšná vysvětlení umožňují uživateli účinně a efektivně využívat AI v jeho kognitivní činnosti pro účely, kterým má AI sloužit.
- Úspěšná vysvětlení umožňují uživateli vybrat nejlepší ze souboru alternativních formálních modelů (např. algoritmů).
- Úspěšná vysvětlení umožňují uživateli správně předvídat, co systém AI v daných případech udělá. To může zahrnovat případy, které AI vyřeší správně, a také případy, které vyřeší špatně (např. selhání, anomálie).
- Úspěšná vysvětlení umožňují uživateli vysvětlit případy, ve kterých se AI mýlí.
- Úspěšná vysvětlení umožňují uživatelům správně posoudit, zda určení systému je správné, a tím překalibrovat jeho důvěru.
- Úspěšná vysvětlení umožňují uživateli vysvětlit ostatním lidem, jak AI funguje.
- Úspěšné vysvětlení umožňují uživateli posoudit kdy a do jaké míry se lze spolehnout na AI navzdory uznání určitých omezení AI.

Závěr

„Umělá inteligence není budoucnost; je to současnost." – podobná tvrzení se objevují v amerických odborných přehledech problematiky [např. 15]. Autoři z projektu DARPA se vyjádřili takto (originál je na konci článku): „Dramatický úspěch ve strojovém učení způsobil explozi nových schopností umělé inteligence. Pokračující pokrok slibuje produkci autonomních systémů, které vnímají, učí se, rozhodují a jednají samy. Tyto systémy nabízejí ohromné výhody, ale jejich účinnost bude omezena neschopností AI vysvětlit svá rozhodnutí a jednání lidským uživatelům. Tato otázka je zvláště důležitá pro Ministerstvo obrany Spojených států, které čelí výzvám vyžadujícím vývoj inteligentnějších, více autonomních a spolehlivějších systémů." [11]

Umělá inteligence (AI) a konkrétně modely strojového učení (ML) určitě prokázaly svůj potenciál tím, že vyrovnávají, někdy dokonce překonávají lidskou úroveň přesnosti diagnostiky a rozhodování u celé řady problémů reálného světa [12]. Nejvyšší přesnosti však často dosahují složité modely, s jejichž interpretací mají potíže i odborníci, což vytváří velký problém pro uživatele. Tyto modely jsou často „černou skříňkou" a neprůhledné, což je obzvláště problematické v odvětvích, jako je zdravotnictví. Pochopení důvodů, které stojí za predikcemi, je klíčové pro vytváření důvěry uživatele. Z tohoto důvodu přichází na scénu vysvětlitelná umělá inteligence (XAI), která pomáhá lidem pochopit a důvěřovat výsledkům a výstupům navržených modely AI [3, 10, 12, 13, 14, 15, 16, 17]. Novější přehledy popisují metody a oblasti aplikace XAI [18, 19, 20, 21, 22, 23, 24]. V práci [25] se zabývají autoři návrhem zásad tvorby XAI zaměřené na veřejnost.

XAI není však bez problémů a omezení. Pro efektivní implementaci technik XAI je zásadní pochopit, jaká je cílová skupina pro XAI vysvětlení. Existuje velký rozdíl mezi vysvětleními vhodnými pro odborníky na umělou inteligenci a vysvětleními pro netechnické uživatele. Technické detaily mohou zahltit netechnické uživatele, což vede k nejistotě nebo dokonce k chybné interpretaci. Při návrhu XAI je proto důležité uvažovat znalosti, cíle, dovednosti a schopnosti koncového uživatele. Post-hoc vysvětlení složitých modelů je dosaženo lokální aproximací jejich chování kolem konkrétního datového bodu instance pomocí jednoduššího, transparentního modelu. I když tyto aproximace nabízejí vhled, nemohou zcela zachytit chování původního modelu. Navíc, jak se složitost modelů s dobou **zvyšuje**, schopnost tyto modely přesně vysvětlit se **snižuje**.

XAI hraje klíčovou roli při řešení neprůhlednosti složitých AI ML modelů a umožňuje získat důvěru v rozhodnutí založená na AI. XAI usnadňuje identifikaci zkreslení, ladění modelů, dodržování předpisů a generování nových znalostí vysvětlením chování modelu. Pro snadnou a efektivní implementaci XAI je k dispozici řada „open source" zdrojů.

Větší transparentnost a vysvětlitelnost užitého AI modelu jsou považovány za jednu ze součástí úsilí o větší přijetí systémů AI. Navíc se uvažuje řešení problémů napříč právními, etickými, technickými a ekonomickými dimenzemi. Vysvětlitelnost a interpretovatelnost sama o sobě nezaručuje účinnost AI aplikace.

Poznámka: Anglické položky ze seznamu literatury (ne kniha Molnara) byly před využitím přeloženy pomocí českého překladače LINDAT MFF (non advance modus).

Originál výňatku DARPA (2021): Dramatic success in machine learning has created an explosion of new AI capabilities. Continued advances promise to produce autonomous systems that perceive, learn, decide, and act on their own. These systems offer tremendous benefits, but their effectiveness will be limited by the machine's inability to explain its decisions and actions to human users.

This issue is especially important for the United States Department of Defense (DoD), which faces challenges that require the development of more intelligent, autonomous, and reliable systems.

EXPLAINABILITY AND INTERPRETABILITY OF AI SYSTEMS

Abstract

The paper gives an overview of explainable and interpretaible artificial intelligence and describes some points that contribute to establishing trust and can provide developers with principles for creating trusted AI systems.

We briefly look at model-agnostic methods with a special focus on the most famous of them, namely LIME and SHAP.

Keywords

artificial intelligence, LIME, SHAP, XAI

Literatura

[1.] Hendl J Jazykové modely a umělá inteligence. Praha LF1 UK, Sborník konference Medsoft 1-9 2023.
[2.] Abgrall G et al. Should AI models be explainable? Critical Care28:301 2024.
[3.] Sajid Ali et al. Explainable Artificial Intelligence (XAI): What we know and what is left to Attain Trustworthy Artificial Intelligence. Information Fusion 2023. (www.elsevier.com/locate/inffus)
[4.] Hendl J Big data. Praha: Grada 2021.
[5.] Ribeiro M T, Singh S a Guestrin C „why should i trust you?" explaining the Predictions of any classifier, in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144, 2016.
[6.] Lundberg S M, Lee S A unified approach to interpreting model predictions, Advances in neural informatics processing systems, vol. 30, 2017.
[7.] Ahmed S et al. A Comparative Analysis of LIME and SHAP Interpreters with Explainable ML-Based Diabetes Predictions. IEEE Acces2016. DOI 10.1109/ACCESS.2024.3422319
[8.] Roberts C V On the Bias-Variance Characteristics of LIME and SHAP in High Sparsity Movie Recommendation Explanation Tasks. Conference'17, July 2017, Washington, DC, USA 2017.
[9.] Juscafresa A N An introduction to explainable artificial intelligence with LIME and SHAP. University of Barcelona 2022.
[10.] Molnar C Interpretable machine learning. E-Book. 2022. (https://christophm.github.io/interpretable-ml-book)
[11.] Editorial DARPA's explainable AI (XAI) program: A retrospective. Defense Advanced Research Projects Agency (DARPA), Applied AI Letters. https://doi.org/10.1002/ail2.61 2021.
[12.] Kolektiv: Jednoduše. Umělá inteligence. Praha: DK Universum 2023.
[13.] Nguyen T A Survey on Explainable Artificial Intelligence: Techniques, XAI-based Model Improvement Methods, Applications Research ·Research Gate April 2024. DOI: 10.13140/RG.2.2.20608.65289 2024.
[14.] Weber L et al. Beyond explaining: Opportunities and challenges of XAI-based model improvement. Information Fusion, 2022.
[15.] Ortiogossa E S et al. EXplainable Artificial Intelligence (XAI)—From Theory to Methods and Applications. IEEE Access, June, 2024. DOI: 10.1109/ACCESS.2024.3409843
[16.] Mueller S T et al. Explanation in Human-AI Systems: A Literature Meta-Review Synopsis of Key Ideas and Publications and Bibliography for Explainable AI. DARPA XAI Program February 2019.
[17.] Dwivedi R et al. Explainable AI (XAI): Core Ideas, Techniques, and Solutions. ACM Comput. Surv. 55 2023.
[18.] Tjoa E, Guan V Survey on Explainable Artificial Intelligence (XAI): towards Medical XAI.IEEE Transactions on Neural Networks and Learning Systems, Vol. 32, 11, 2021.
[19.] Sheu R K et al.Survey on Medical Explainable AI (XAI): Recent Progress, Explainability Approach, Human Interaction and Scoring System. Sensors 2022, 22 https://doi.org/10.3390/s22208068
[20.] Chaddad A et al. Survey of Explainable AI Techniques in Healthcare. Sensors 2023, 23 DOI: 10.3390/s23020634
[21.] Hea X et al.What Are the Users' Needs? Design of a User-Centered Explainable Artificial Intelligence Diagnostic System. International Journal of Human–Computer Interaction Vol. 39, 7 2023.
[22.] Frasca M et al. Explainable and interpretable artificial intelligence in medicine: a systematic bibliometric review. Discover Artificial Intelligence, 4, 2024 DOI: 10.1007/s44163-024-00114-7
[23.] Bharati S et al. A Review on Explainable Artificial Intelligence for Healthcare: Why, How, and When? IEEE Transactions on Artificial Intelligence, pp. 1429-1442, vol. 5 2024 DOI: 10.1109/TAI.2023.3266418
[24.] Kinger S, Kulkarni V A review of explainable AI in medical imaging: implications and applications. International Journal of Computers and Applications, Vol. 46, 11, 2024 DOI: 10.1080/1206212X.2024.2404082
[25.] He X What Are the Users' Needs? Design of a User-Centered Explainable Artificial Intelligence Diagnostic System. International Journal of Human–Computer Interaction, Vol. 39, 7, 2023. DOI: 10.1080/10447318.2022.2095093#

Kontakt

Prof. Jan Hendl

1. LF UK
jhendl1111@gmail.com