

Predikce BSK₅ na odtoku z čistírny odpadních vod pomocí strojového učení

NASSIMA LEKOUAGHET, LAKHDAR BOURAGBI, BRAHIM MAHFOUD

Klíčová slova: biochemická spotřeba kyslíku (BSK₅) – čištění odpadních vod – strojové učení – soft senzor – XGBoost – predikce BSK₅ – optimalizace procesů

ABSTRAKT

Přesná a rychlá predikce biochemické spotřeby kyslíku (BSK₅) na odtoku je důležitá pro dodržování legislativních požadavků a zlepšení provozu čistíren odpadních vod (ČOV). Tradiční laboratorní stanovení BSK₅ trvá několik dní, což způsobuje prodlevy při hodnocení kvality vody a optimalizaci provozu čistírny. Aby vyřešila tento problém, vyvíjí tato studie model strojového učení pro odhad BSK₅ na odtoku s využitím snadno dostupných provozních dat. Byl použit soubor dat z velkokapacitní ČOV zahrnující 12 klíčových proměnných z přítokových i odtokových dat, jako jsou pH, chemická spotřeba kyslíku (CHSK), vodivost, celkové nerozpuštěné látky (NL), BSK₅ a teplota, přičemž cílovou proměnnou byla BSK₅ odtokové vody. Data byla vyčištěna a zpracována pomocí metody interkvartilového rozpětí (IQR) za účelem odstranění odlehlých hodnot a byly odvozeny nové proměnné, které lépe odrážejí chování procesu. Z několika testovaných modelů vykázal nejlepší výsledky model XGBoost, který dosáhl hodnoty $R^2 = 0,9695$ a střední kvadratické chyby (MSE) 0,05 při rozdělení dat na trénovací a testovací část v poměru 80 % ku 20 % na základě časové posloupnosti. Studie ukazuje, že strojové učení může sloužit jako účinný nástroj pro predikci BSK₅, jenž významně zkracuje pětidenní prodlevu laboratorního stanovení a umožňuje rychlejší rozhodování založené na datech při řízení ČOV.

ÚVOD

Správné nakládání s komunálními a průmyslovými odpadními vodami a jejich čištění jsou nezbytné pro ochranu vodních zdrojů a zdraví lidí i ekosystémů. K nejčastěji využívaným ukazatelům mezi různými parametry používanými k hodnocení kvality odtoku patří pětidenní biochemická spotřeba kyslíku (BSK₅). Vyjadřuje množství kyslíku potřebného pro biologický rozklad organických znečišťujících látek ve vodě mikroorganismy během pětidenního období [1]. BSK₅ poskytuje přímou informaci o biologicky rozložitelném organickém zatížení, a proto je klíčovým ukazatelem pro hodnocení účinnosti čištění a dopadů na životní prostředí. Z tohoto důvodu regulační orgány po celém světě stanovily přísné limity vypouštění s cílem udržet přijatelné hodnoty BSK₅ ve vyčištěných odpadních vodách [2].

Navzdory své důležitosti má konvenční stanovení BSK₅ významná omezení. Standardní metoda vyžaduje pětidenní inkubační dobu, což zpožďuje zpětnou vazbu o účinnosti čištění a omezuje možnost včasného provozního rozhodování [3]. Přesnost a reprodukovatelnost měření mohou navíc negativně ovlivnit faktory, jako jsou toxicita vzorku, vliv nitrifikace, variabilita mikrobiální inokulace a problémy spojené s uchováváním vzorků [4–7]. Tyto nedostatky poukazují na potřebu rychlých, spolehlivých a kontinuálních metod odhadu BSK₅.

Strojové učení (Machine Learning, ML) a modelování pomocí soft senzorů se jeví účinnými nástroji pro rychlý odhad BSK₅, protože umožňují predikci na základě rutinně měřených provozních parametrů, jako jsou CHSK, pH, vodivost, teplota a NL [8, 9]. Na rozdíl od empirických nebo deterministických modelů [7, 8], jež často selhávají při proměnlivých podmínkách v důsledku složitých biochemických a hydrodynamických procesů v ČOV [10, 11], dokážou přístupy založené na ML zachytit nelineární interakce a dynamiku systému, a tím podpořit proaktivní řízení procesu, včasnou detekci anomálií a udržitelnější provoz čistíren.

Nedávné studie ukazují, že moderní metody strojového učení mohou dosahovat velmi vysoké predikční úspěšnosti při odhadu BSK₅ a dalších parametrů kvality vody. Např. Dantas et al. [12] prokázali, že umělé neuronové sítě (ANN), tedy výpočetní modely inspirované způsobem zpracování informací biologickými neurony, dosahují dobrých výsledků při predikci BSK₅ a CHSK v ČOV v plném provozním měřítku. Zhang et al. [13] použili model Random Forest, tedy soubor mnoha rozhodovacích stromů společně zvyšujících přesnost predikce, pro odhad BSK₅ v říční vodě a dosáhli hodnoty R^2 přibližně 0,89. Jafar et al. [14] využili hluboké neuronové sítě (DNN), složitější formu architektury neuronových sítí, schopné zachytit vysoce nelineární vztahy, a zaznamenali hodnoty R^2 v rozmezí 0,89–0,93. Další studie prokázaly srovnatelné výsledky: modely XGBoost (Extreme Gradient Boosting), využívající gradientně posilované rozhodovací stromy, dosáhly v ekologických aplikacích hodnot R^2 kolem 0,92 [15], zatímco sítě RVFL (Random Vector Functional Link) vykázaly v podmínkách ČOV hodnotu $R^2 \approx 0,924$ [16]. Naproti tomu jednodušší modely často poskytují nižší přesnost. Např. Ali et al. [17] uvádějí hodnotu R^2 přibližně 0,88 při použití modelu CatBoost (Categorical Boosting), který umožňuje efektivní práci s kategoriálními i numerickými daty při minimálních nárocích na jejich předzpracování. Je třeba poznamenat, že modely lineární regrese nejsou při proměnlivých podmínkách čištění odpadních vod dostatečně robustní, protože často nedokážou zachytit nelineární interakce a složitou dynamiku systému [18].

Přesný a včasný odhad biochemické spotřeby kyslíku (BSK₅) zůstává v oblasti čištění odpadních vod trvalou výzvou, protože tradiční laboratorní analýzy vyžadují pětidenní inkubační dobu. Tato přirozená prodleva omezuje rychlost hodnocení procesu. K řešení tohoto problému studie navrhuje predikční rámec využívající algoritmus XGBoost, který byl zvolen díky své vysoké schopnosti pracovat s vysokodimenzionálními daty a nelineárními vztahy. Integrací robustního přístupu, který je založen na datech a tvorbě proměnných a vychází ze znalosti oboru, se model snaží přesně zachytit složitou dynamiku procesu čištění za reálných provozních podmínek. Hlavním cílem této práce je poskytnout spolehlivou alternativu pro rychlé hodnocení kvality odpadních vod, a tím umožnit rychlejší rozhodování a zlepšit reakci při řízení ČOV. Následující kapitoly podrobně popisují metodiku a proces vývoje modelu použité k dosažení těchto cílů.

METODIKA A VÝVOJ MODELU

Charakteristika lokality a sběr dat

Soubor dat použitý v této studii byl získán z komunální ČOV v provincii Skikda nacházející se v severovýchodním Alžírsku (36,89° s. š., 7,01° v. d.). Zařízení využívá aktivační proces se středním zatížením navržený pro zpracování proměnlivého organického zatížení z domácích a komunálních zdrojů. Projektovaná kapacita čistírny činí 46 000 m³/den a zařízení obsluhuje přibližně 279 700 ekvivalentních obyvatel.

Proces čištění zahrnuje mechanické předčištění, odstranění písku, primární sedimentaci, biologickou oxidaci v aktivačních nádržích, sekundární dosazování a konečnou dezinfekci (obr. 2). Systém je navržen pro průměrné přítokové zatížení BSK₅ přibližně 280 mg/l, což odpovídá 14 950 kg BSK₅ za den, 34 383 kg CHSK za den a 17 250 kg nerozpuštěných látek (NL) za den.

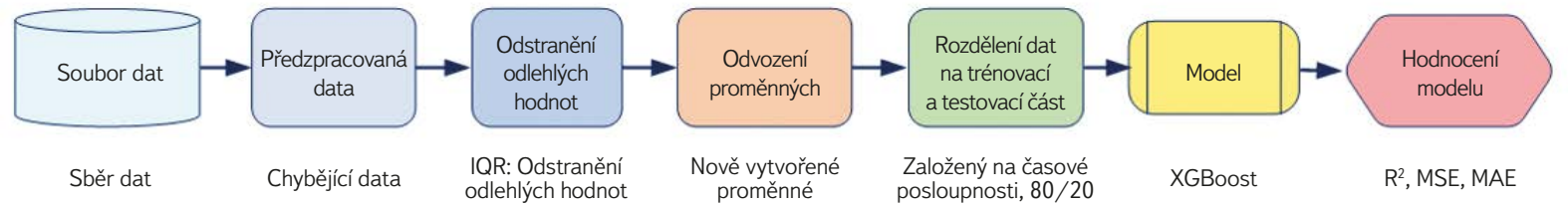
Vyčištěné odpadní vody jsou vypouštěny do řeky Oued Safsaf, citlivého vodního útvaru, který má v regionu významnou ekologickou funkci. Produkovaný přebytečný kal je odvodňován a následně využíván jako zemědělské hnojivo, což zdůrazňuje potřebu spolehlivé kontroly kvality odtoku. Nedostatečné odstranění organických látek zvyšuje množství zbytkové biologicky rozložitelné hmoty, jež vede ke zvýšení vlhkosti kalu a zhoršení jeho odvodnitelnosti, a tím ke snížení kvality výsledného biosolidu.

Informace o vzorcích použitých v této studii (2015–2024) byly digitalizovány z historických archivů ČOV. Soubor dat představuje průběžný monitoring přítoku a odtoku během všech 12 měsíců každého roku. Sledované parametry zahrnovaly biochemickou spotřebu kyslíku (BSK₅), chemickou spotřebu kyslíku (CHSK), celkové nerozpuštěné látky (NL), pH, teplotu, elektrickou vodivost a rozpuštěný kyslík (DO). Všechny analýzy byly prováděny v souladu se Standard Methods for the Examination of Water and Wastewater (APHA, 2017) [4].

Postup práce a strategie modelování

Tato kapitola popisuje vývoj vysoce přesného modelu pro predikci BSK₅ na odtoku z ČOV. Přístup využívá reálný soubor provozních dat, na který byl aplikován algoritmus XGBoost za účelem optimalizace výkonnosti modelu.

Pro zvýšení přesnosti modelu byla použita tvorba proměnných vycházející ze znalosti oboru; rozdělení dat na trénovací a testovací část podle časové posloupnosti (80 % / 20 %) pak umožnilo realistické hodnocení modelu. Postup zahrnuje předzpracování dat (zpracování chybějících hodnot a odstranění odlehlých hodnot), tvorbu proměnných, trénování modelu a jeho hodnocení pomocí ukazatelů R², MSE a MAE. Obr. 1 znázorňuje kompletní pracovní postup od surových dat až po finální predikci a prezentuje datově orientovaný rámec pro odhad BSK₅ na odtoku.



Obr. 1. Vývojový diagram modelovacího postupu pro predikci BSK₅ na odtoku
Fig. 1. Flowchart of the modeling pipeline for predicting effluent BOD₅

Sběr, čištění a předzpracování dat

Sběr dat a úvodní kontrola dat

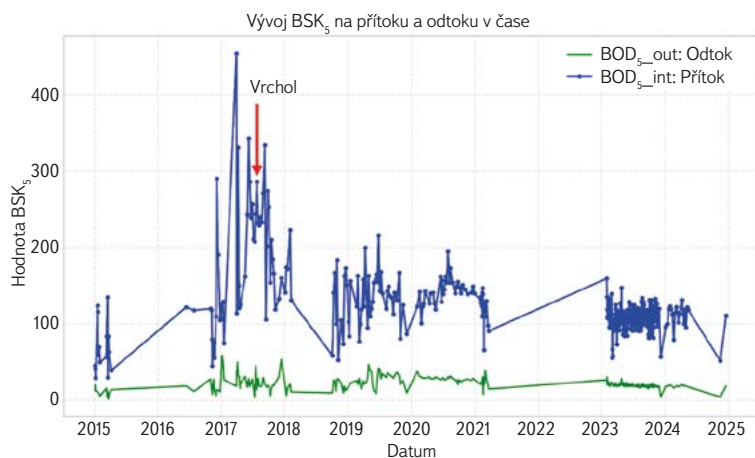
Soubor dat zahrnuje období devíti let (2015–2024) a obsahuje 410 platných záznamů, což odpovídá průměrně třem až čtyřem měřením za měsíc kvůli nepravidelné frekvenci odběru vzorků. Každý záznam obsahuje 15 provozních parametrů měřených jak na přítoku (vstupu), tak na odtoku (výstupu) z procesu čištění. Tyto proměnné byly vybrány na základě jejich významu pro účinnost biologického čištění a dostupnosti konzistentních dat.

Tab. 1. Data monitoringu procesu čištění odpadních vod
Tab. 1. Water treatment process monitoring data

Proměnné	Jednotka	Přítok	Odtok
Datum	dd/mm/rr	Datum	Datum
pH	-	pH_int	pH_out
Pětidenní biochemická spotřeba kyslíku (BSK ₅)	mg/l	BOD ₅ _int	BOD ₅ _out
Chemická spotřeba kyslíku (CHSK)	mg/l	COD_int	COD_out
Elektrická vodivost	μS/cm	Cond_int	Cond_out
Nerozpuštěné látky (NL)	mg/l	TSS_int	TSS_out
Teplota (T)	°C	Temp_int	Temp_out
Rozpuštěný kyslík (DO)	mg/l	O2_int	O2_out

Jako cílová proměnná byla zvolena BSK₅ na odtoku (BOD₅_out), jelikož představuje hlavní ukazatel kvality odpadní vody. Standardní pětidenní zkouška je pro operativní řízení považována za příliš pomalou; rychlý odhad však ve srovnání s konvenčními metodami měření umožňuje predikční modelování. Díky tomu lze problémy zachytit již v rané fázi a čistírnu provozovat efektivněji.

Obr. 2 znázorňuje časový vývoj koncentrací BSK₅ na přítoku a odtoku v období 2015–2025. Zatímco hodnoty na přítoku (BOD₅_int, modře) vykazují značnou proměnlivost s vrcholy v letech 2017–2018 v důsledku sezonních vlivů nebo hydraulických špiček, hodnoty na odtoku (BOD₅_out, zeleně) zůstávají od roku 2019 stabilní a nízké. Tato stabilita potvrzuje konzistentní účinnost čištění navzdory proměnlivému zatížení na přítoku. Výpadky dat v letech 2016, 2018 a 2022 způsobené problémy se senzory nebo záznamem dat byly řešeny během fáze předzpracování dat.

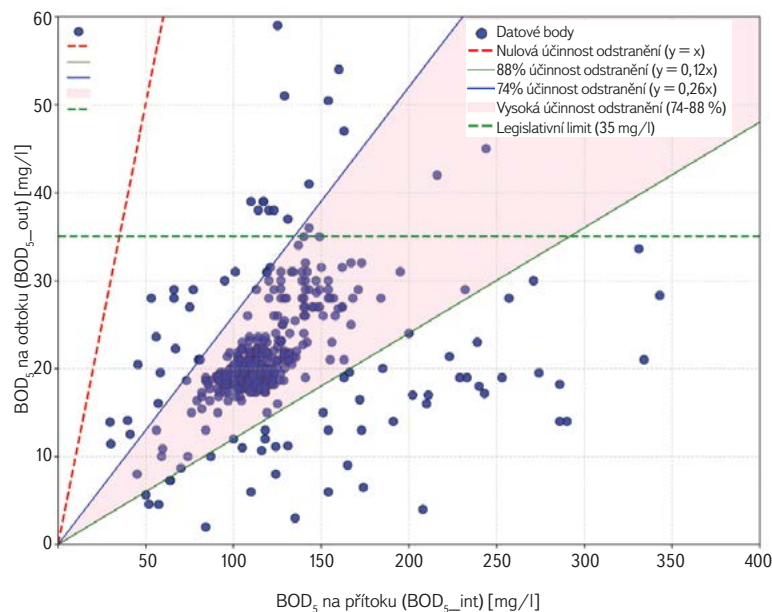


Obr. 2. Časová řada hodnot BSK_5 (= BOD_5) na přítoku a odtoku ukazuje stabilní účinnost čištění navzdory výrazným výkyvům zatížení na přítoku

Fig. 2. Time series of BOD_5 inflow and outflow showing consistent treatment performance despite significant fluctuations

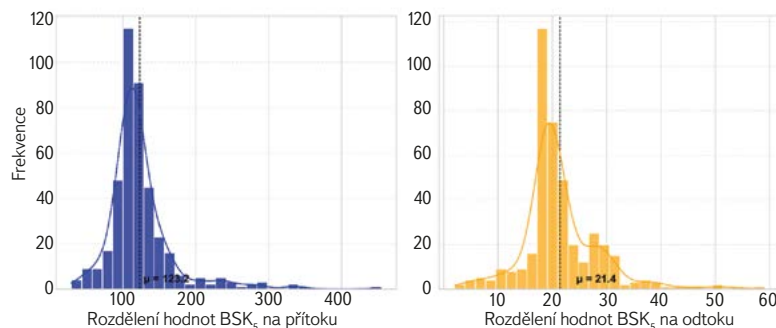
Obr. 3 ukazuje vztah mezi hodnotami BSK_5 na přítoku a odtoku s důrazem na celkovou účinnost odstraňování znečištění čistírenského systému v širokém rozsahu zatížení na přítoku. Hustý shluk datových bodů, kde se hodnoty BSK_5 na přítoku pohybují mezi 80 a 180 mg/l a koncentrace na odtoku zůstávají mezi 10 a 47 mg/l, ukazuje na stabilní a účinné snižování organického zatížení během provozních cyklů. To odpovídá odhadované účinnosti odstranění přibližně 74–88 %, jak je znázorněno na obr. 3 oblastí mezi přímkami ($y = 0,26x$) a ($y = 0,12x$), jež představují odstranění 74 %, resp. 88 %. Je pozoruhodné, že většina hodnot BSK_5 na odtoku zůstává pod legislativním limitem 35 mg/l pro vypouštění odpadních vod v souladu s alžírskými národními normami [19]. Pozorované nelineární rozložení odráží inherentní složitost biologických procesů čištění, a podporuje tak využití pokročilých metod strojového učení pro modelování účinnosti procesu. Jak je zjevné z obr. 3, několik izolovaných datových bodů se odchyluje od hlavního shluku a vykazuje relativně vysoké koncentrace BSK_5 na odtoku navzdory střednímu zatížení na přítoku. Tyto odchylky nejsou náhodné, ale pravděpodobně souvisejí s přechodnými provozními poruchami, jako jsou výkyvy teploty nebo hydraulické rázy (náhlé zvýšení průtoku nebo organického zatížení), které ovlivňují aktivitu mikroorganismů a účinnost čištění. Mohou také odrážet chyby měření nebo dočasné poruchy zařízení. Přesné určení příčiny těchto jevů je však obtížné vzhledem k absenci doplňkových dat, jako jsou podrobné údaje o průtocích, srážkové záznamy, teplota okolí nebo provozní záznamy z ČOV. Navzdory tomuto omezení přítomnost anomálií zdůrazňuje význam důkladného předzpracování dat, zejména v případě BSK_5 na odtoku. Současně poukazuje na potřebu adaptivních modelovacích přístupů schopných zachovat predikční přesnost i při proměnlivých a neideálních provozních podmínkách.

Histogramy BOD_{5_int} a BOD_{5_out} na obr. 4 vykazují pravoúhlé rozdělení, přičemž většina hodnot je soustředěna kolem centrálních maxim (např. kolem 200 mg/l pro BOD_{5_int}) a několik vysokých hodnot zasahuje do pravé části rozdělení (> 400 mg/l). Tyto extrémní hodnoty nemusí nutně představovat chyby; mohou odrážet skutečné události, jako jsou průmyslové vypouštění nebo přítok dešťových vod. Vyžadují však pečlivé zpracování ve fázi předzpracování dat, protože mohou zkreslit trénování modelu a ovlivnit jeho přesnost a věrohodnost. Byl použit vyvážený přístup: extrémní hodnoty spojené se skutečnými provozními poruchami byly ponechány, zatímco izolované odlehlé hodnoty byly za účelem zajištění kvality dat odstraněny.



Obr. 3. Bodový graf vztahu BSK_5 na odtoku a BSK_5 na přítoku

Fig. 3. Scatter plot of effluent versus influent BOD_5



Obr. 4. Rozdělení hodnot BSK_5 na přítoku a odtoku v období 2015–2024; je patrné shlukování kolem středních hodnot s pravoúhlými konci rozdělení, což ukazuje na výskyt vzácných událostí s vysokým zatížením

Fig. 4. Distribution of influent and effluent BOD_5 values between 2015 and 2024; there is a clustering around central values with right-skewed tails, which indicates the occurrence of rare, high-load events

Prvotní zpracování dat

Surová data vykazovala v důsledku problémů s měřením a záznamem dat chybějící hodnoty, nekonzistence a odlehlé hodnoty. Pro zlepšení kvality dat byl použit strukturovaný postup předzpracování zahrnující zpracování chybějících dat, detekci odlehlých hodnot a časovou synchronizaci dat.

Zpracování chybějících dat

Úvodní analýza souboru dat ukázala, že chybějící hodnoty byly označeny symbolem '/'; zejména u proměnných rozpuštěného kyslíku (O_{2_int} a O_{2_out}). Více než 62 % pozorování v těchto sloupcích chybělo, pravděpodobně v důsledku výpadků senzorů, údržby zařízení nebo přerušovaného záznamu dat, a proto nebyla pro spolehlivé modelování vhodná. Ponechání těchto proměnných by vedlo k významnému zkreslení nebo by vyžadovalo rozsáhlou imputaci, která by mohla narušit skutečný charakter procesu. Z tohoto důvodu byly v zájmu zachování integrity dat obě proměnné z analýzy odstraněny.

Po nahrazení všech položek '/' hodnotou NaN a převedení zbývajících měření na přítoku (int) a odtoku (out) do numerického formátu bylo zjištěno,

že ostatní proměnné (pH, BSK₅, CHSK, TSS, teplota a datum záznamu) jsou úplné a neobsahují chybějící hodnoty. Je zřejmé, že počet proměnných byl snížen z 15 (datum odběru, 7 vstupních a 7 výstupních proměnných uvedených v tab. 1) na 13 po vyřazení rozpuštěného kyslíku (DO) z dat měřených na přítoku i odtoku. Současně nebyly identifikovány žádné další chybějící hodnoty vyžadující imputaci. Tyto prediktory byly proto přímo použity v procesu modelování.

Je třeba zdůraznit, že DO představuje základní provozní parametr aktivních systémů, protože přímo ovlivňuje aktivitu mikroorganismů a účinnost nitrifikace [11]. Jeho vyloučení z této analýzy bylo způsobeno značnou ztrátou dat přesahující 60 %, nikoli sníženým významem DO pro samotné procesy čištění.

Detekce a odstranění odlehlých hodnot

Odlehlé hodnoty cílové proměnné, tedy BSK₅ na odtoku (BOD₅_out), byly identifikovány pomocí metody interkvartilového rozpětí (IQR), což je robustní statistický přístup běžně používaný při analýze environmentálních dat [20].

Interkvartilové rozpětí (IQR) je definováno jako:

$$IQR = Q_3 - Q_1 \quad (1)$$

kde:

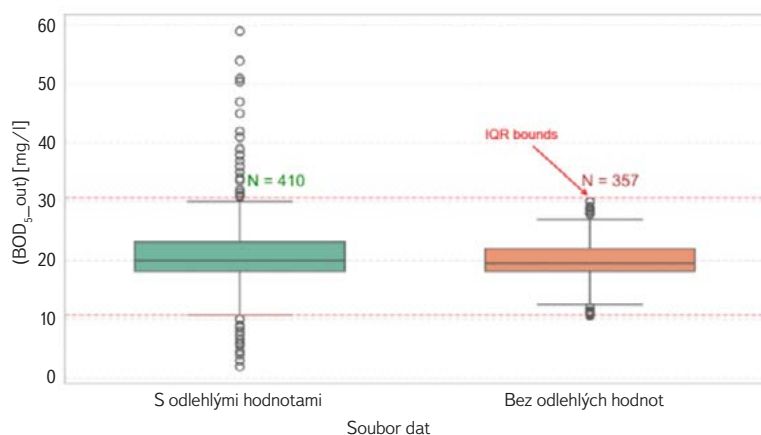
Q_1 a Q_3 představují 25. a 75. percentil.

Za odlehlé hodnoty byla považována pozorování nacházející se mimo následující meze:

$$\text{Lower Bound} = Q_1 - 1,5 \times IQR \quad (2)$$

$$\text{Upper Bound} = Q_3 + 1,5 \times IQR \quad (3)$$

Pro BOD₅_out činily hodnoty $Q_1 = 18,18$ mg/l a $Q_3 = 28,63$ mg/l, což odpovídá hodnotě $IQR = 10,45$ mg/l. Dolní a horní mez byly následně vypočítány na 2,50 mg/l, resp. 44,30 mg/l. Použití tohoto kritéria vedlo k odstranění 53 extrémních pozorování, včetně neobvykle vysokých hodnot (např. 59 mg/l), jež byly pravděpodobně způsobeny chybami měření nebo občasnými provozními poruchami. Tím se velikost souboru dat snížila ze 410 na 357 vzorků.



Obr. 5. Porovnání krabicových diagramů BOD₅_out před a po odstranění odlehlých hodnot

Fig. 5. Box plot comparison of BOD₅_out before and after outlier filtering

Následně byly odstraněny řádky obsahující chybějící hodnoty buď v cílové proměnné, nebo v klíčových prediktorech. Finální upravený soubor dat použitý pro modelování obsahoval N = 357 úplných vzorků. Jak ukazuje obr. 5, tyto kroky předzpracování zlepšily symetrii rozdělení a snížily jeho rozptýl, čímž posílily robustnost a zobecnitelnost predikčních modelů.

Časová synchronizace dat

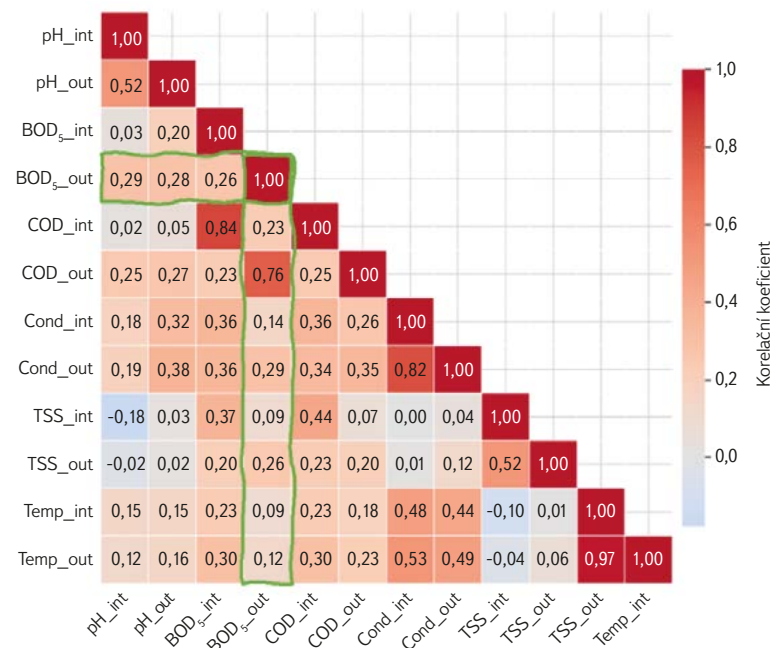
Sloupec s datem byl převeden do formátu datetime a použit k chronologickému seřazení souboru dat. Tento přístup byl zvolen za účelem zachování časových závislostí, což umožnilo realistické ověření modelu.

Rozšíření proměnných a výběr cílové proměnné pro zlepšení predikce

Při vývoji modelu strojového učení se ukázalo, že původní soubor dat obsahoval nedostatečný počet proměnných pro dosažení vysoké predikční přesnosti. K překonání tohoto omezení byl prostor proměnných rozšířen pomocí cílené tvorby nových proměnných a pečlivého výběru nejrelevantnějších vstupních a výstupních parametrů. Tato podkapitola popisuje proces vytváření nových proměnných a definování cílové proměnné za účelem zlepšení výkonnosti modelu.

Výběr proměnných a cílové proměnné

V prediktivním modelování lze za potenciální vstupy považovat všechny dostupné systémové parametry, avšak ne všechny přispívají stejnou měrou k výkonnosti modelu. Výběr a tvorba informativních proměnných jsou proto zásadní, neboť nerelevantní proměnné mohou snižovat přesnost modelu [21]. Za účelem zvýšení predikční schopnosti byly jako vstupy zahrnuty také parametry na odtoku, což je strategie, která v podobných aplikacích prokazatelně zvyšuje přesnost modelu.



Obr. 6. Matice Pearsonových korelačních koeficientů

Fig. 6. Pearson correlation matrix

Obr. 6 zobrazuje matici Pearsonových korelačních koeficientů znázorňující lineární vztahy mezi provozními proměnnými v souboru dat. Silné kladné korelace (červeně) označují proměnné, jež rostou společně (např. COD_int a BOD₅_int), zatímco záporné korelace (modře) vyjadřují opačné trendy. Podle očekávání

BSK₅ a CHSK na přítoku vykazují silnou kladnou korelaci ($r = 0,84$) a vysokou korelaci je možné pozorovat také mezi BSK₅ a CHSK na odtoku ($r = 0,76$). Teplota a TSS_{int} vykazují slabé korelace s ostatními proměnnými, což naznačuje jejich malý vliv na krátkodobé změny BSK₅. Teplota (Temp_{int}, Temp_{out}) a nerozpuštěné látky (TSS_{int}) vykazují slabé korelace se zbývajících proměnnými (obr. 6). Tyto proměnné však byly ponechány, aby bylo možné zachytit jemnější vlivy na BSK₅ a zvýšit přesnost modelu. Celkový počet proměnných použitých ve finální architektuře modelu tak činí 13.

Tvorba proměnných: vytváření nových prediktivních vstupních proměnných

Tvorba proměnných převádí surová data na informativní proměnné, které zvyšují interpretovatelnost modelu, jeho predikční výkonnost a schopnost zobecnění [20, 21]. V souladu s principy čištění odpadních vod byly odvozeny proměnné založené na znalostech oboru, jež lépe vystihují dynamiku procesů čištění. Bylo odvozeno šest dalších proměnných: ukazatele účinnosti odstranění, poměrové ukazatele a ukazatele rychlosti změny. Tím se původní sada 13 proměnných rozšířila na celkem 18 prediktorů, s výjimkou data odběru vzorků. Předběžné hodnocení potvrdilo, že tyto oborově specifické proměnné zvýšily přesnost modelu, zejména pro účely řízení procesů. Všechny odvozené proměnné jsou shrnuty v tab. 2, včetně jejich matematických definic a fyzikální interpretace. U každé proměnné byla posouzena její fyzikální smysluplnost a analyzována korelace s cílovou proměnnou BOD₅out.

Tab. 2. Odvozené proměnné a jejich matematický popis
Tab. 2. Engineered features and their mathematical descriptions

Odvozená proměnná	Vzorec	Poznámky
COD_removal_eff	$\frac{COD_{int} - COD_{out}}{COD_{int}} \quad (4)$	Účinnost odstranění CHSK; vyšší hodnoty indikují účinnější odstranění znečištění
pH_change	$pH_{out} - pH_{int} \quad (5)$	Rozdíl pH mezi přítokem a odtokem; ukazatel neutralizačních procesů
COD_BOD ₅ _ratio_int	$\frac{COD_{int}}{BOD_{5int}} \quad (6)$	Poměr CHSK/BSK ₅ na přítoku; vysoké hodnoty indikují nízkou biologickou rozložitelnost
COD_out_BOD ₅ _int_ratio	$\frac{COD_{out}}{BOD_{5int}} \quad (7)$	Poměr CHSK/BSK ₅ na odtoku a přítoku; ukazatel změny biologické rozložitelnosti a zbytkového znečištění během procesu čištění
BOD ₅ _int_pH_int	$BOD_{5int} \times pH_{int} \quad (8)$	Vliv pH na biologický rozklad organických látek (BSK ₅)
Cond_ratio	$\frac{Cond_{out} - Cond_{int}}{Cond_{int}} \quad (9)$	Změna elektrické vodivosti (Cond _{int} , Cond _{out}) mezi přítokem a odtokem

Strategie rozdělení na trénovací a testovací sadu

Jelikož datový soubor zahrnuje téměř desetileté období (2015–2024) a obsahuje časově označená sekvenční měření, byla data rozdělena na trénovací a testovací sadu podle časové posloupnosti, aby bylo zachováno chronologické uspořádání a zabránilo se úniku informací. Na rozdíl od náhodného rozdělení, které může vést k časovému zkreslení tím, že umožňuje budoucím datům ovlivnit proces trénování, tento přístup chronologicky přiřadil prvních 80 % vzorků (286 pozorování, leden 2015 až prosinec 2022) do trénovací/validační sady a zbývajících 20 % (71 pozorování, leden 2023 až květen 2024) do testovací sady. Tím byla simulována situace z reálného provozu, kdy jsou modely trénovány na historických datech a následně vyhodnocovány na datech z pozdějšího období [22, 23].

Tento přístup umožňuje hodnotit model za realistických podmínek predikce, jež zohledňují jak sezonní variabilitu, tak dlouhodobé provozní trendy. Výkonnostní metriky použité v této studii, včetně koeficientu determinace R² (který udává, jakou část variability BSK₅ na odtoku je model schopen vysvětlit), střední kvadratické chyby (MSE, jež zdůrazňuje přítomnost velkých chyb predikce) a střední absolutní chyby (MAE, která vyjadřuje průměrnou odchylku mezi predikovanými a naměřenými hodnotami), tak poskytují smysluplný obraz o skutečné predikční schopnosti modelu. Přísným oddělením testovacích dat od fází trénování a ladění modelu zůstává hodnocení nezájaté a reprezentativní pro reálné využití.

Zvolený model: architektura a hyperparametry XGBoostu

XGBoost (Extreme Gradient Boosting) byl zvolen pro svou vysokou výpočetní efektivitu a vynikající schopnost pracovat s nelineárními biologickými daty. Jde o vysoce optimalizovanou implementaci gradientního boostingu, která aditivním způsobem kombinuje rozhodovací stromy a využívá gradientní sestup k minimalizaci chyby predikce. Efektivitu trénování a výkonnost modelu zvyšuje prostřednictvím:

- gradientů druhého řádu (optimalizace založené na Hessově matici), jež umožňují rychlejší konvergenci,
- vestavěné regularizace L1 (Lasso) a L2 (Ridge), které regulují složitost modelu a omezují přeučení,
- podvzorkování sloupců a řádků, jež zlepšuje schopnost zobecnění modelu.

Díky své robustnosti, rychlosti a vysoké predikční přesnosti patří XGBoost mezi nejpoužívanější algoritmy v soutěžích strojového učení (např. Kaggle) i v reálných aplikacích napříč technickými obory, financemi a environmentálními vědami [24, 25].

Pro trénování modelu XGBoost byly v této studii použity následující hyperparametry:

- `n_estimators = 200`: počet boostingových iterací (stromů),
- `learning_rate = 0,05`: míra učení (learning rate) pro omezení přeučení,
- `max_depth = 5`: omezuje hloubku stromů s cílem vyvážit flexibilitu modelu a jeho interpretovatelnost,
- `subsample = 0,8`: pro každý strom využívá 80 % vzorků, čímž zavádí prvek náhodnosti a zvyšuje robustnost modelu,
- `colsample_bytree = 0,8`: při každém dělení náhodně vybírá 80 % proměnných, čímž zvyšuje rozmanitost modelu,
- `reg_alpha = 0,1`, `reg_lambda = 0,1`: používá mírnou regularizaci L1 a L2 k omezení velikosti koeficientů.

Tyto parametry byly vybrány na základě předchozích experimentů a odpovídají osvědčeným postupům používaným při řešení regresních úloh. Model byl implementován pomocí knihovny XGBoost pro Python (verze 1.7.6).

Metriky hodnocení výkonnosti modelu

Predikční výkonnost modelu byla hodnocena pomocí tří standardních regresních metrik: střední absolutní chyby (MAE), střední kvadratické chyby (MSE) a koeficientu determinace (R^2). Jejich matematické definice jsou uvedeny níže:

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (10)$$

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (11)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (12)$$

kde:

y_i	je	pozorovaná hodnota BOD_{5_out}
$\hat{y}_i$		predikovaná hodnota
$\bar{y}$		aritmetický průměr pozorovaných hodnot
n		počet pozorování v testovací sadě

MAE vyjadřuje průměrnou velikost chyb predikce bez ohledu na jejich směr, přičemž všechny odchylky hodnotí stejně, a poskytuje tak srozumitelný ukazatel celkové přesnosti modelu. Díky své snadné interpretovatelnosti je široce využívána, zejména v environmentálním a geoprostorovém modelování. MSE naproti tomu před výpočtem průměru umocňuje jednotlivé chyby na druhou, čímž přikládá větší váhu velkým odchylkám, a je proto účinná při identifikaci modelů náchylných k vytváření výrazných odlehlých hodnot [26]. Koeficient determinace R^2 tyto metriky doplňuje tím, že vyjadřuje podíl variability cílové proměnné vysvětlený modelem. Hodnota blízká 1 svědčí o velmi dobré shodě modelu s daty, zatímco nižší hodnoty ukazují na slabší schopnost modelu vysvětlit variabilitu cílové proměnné. Všechny metriky výkonnosti byly vypočítány na vyčleněné testovací sadě (leden 2023 až květen 2024), což zajistilo nezkreslené hodnocení schopnosti modelu zobecňovat na nová data.

VÝSLEDKY A DISKUZE

Tato kapitola prezentuje výkonnost jednotlivých základních modelů a stohovaného ansámblu při predikci BSK_5 na odtoku (BOD_{5_out}). Analýza zahrnuje kvantitativní výsledky, vizualizace a jejich interpretaci v kontextu řešeného problému s cílem zdůraznit přesnost, robustnost a praktickou využitelnost modelů.

Analýza výkonnosti modelů

Predikční výkonnost modelu XGBoost byla hodnocena po důkladném čištění dat a odvození proměnných. Z původního datového souboru zůstalo pro analýzu 357 vysoce kvalitních vzorků, které byly rozděleny na trénovací sadu ($n = 286$) a testovací sadu ($n = 71$).

Model byl sestaven s využitím 18 proměnných, jež zahrnovaly jak surová data ze senzorů, tak i odvozené ukazatele, jako jsou účinnost odstraňování znečištění a poměry $CHSK/BSK_5$. Z nich 17 parametrů sloužilo jako vstupní proměnné, včetně 11 původních parametrů přítoku a odtoku a šesti odvozených proměnných. Zbývající proměnná ($BSK_{5_výstup}$) byla určena jako cílový výstup. Metriky výkonnosti pro testovací datovou sadu jsou shrnuty v tab. 3.

Výsledky prokazují vysokou míru predikční přesnosti. Hodnota R^2 ve výši 96,95 % naznačuje, že model účinně zachycuje nelineární dynamiku procesu

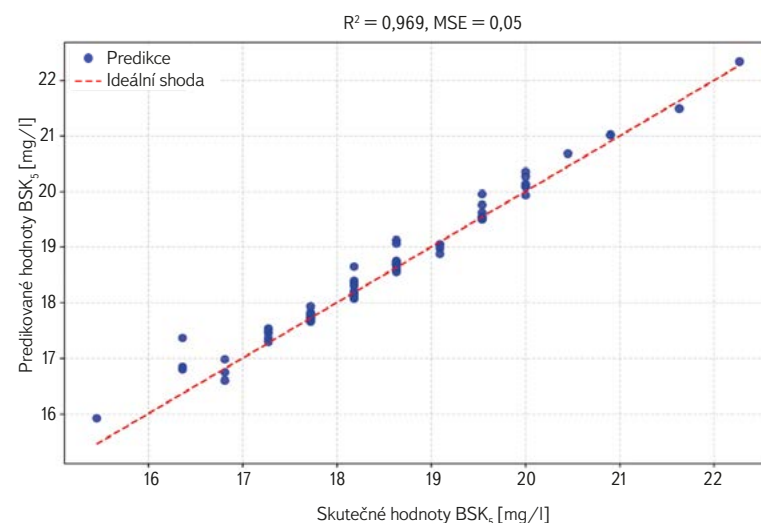
čištění. Nízká hodnota MAE (0,146) je obzvláště významná, protože ukazuje, že průměrná chyba predikce modelu je zanedbatelná ve vztahu k legislativním limitům pro BSK_5 .

Tab. 3. Výsledky modelu XGBoost při predikci BSK_5

Tab. 3. XGBoost performance metrics for BOD_5 prediction

Metrika	Hodnota
Střední kvadratická chyba (MSE)	0,05
Střední absolutní chyba (MAE)	0,146
Koeficient determinace (R^2)	96,95 %

Obr. 7 ukazuje těsnou shodu mezi skutečnými a predikovanými hodnotami BSK_5 . Model přesně sleduje jak stabilní období, tak i kolísání hodnot a zachovává vysokou přesnost i během období se zvýšeným zatížením. Tato výkonnost překonává výsledky dosažené v předchozích studiích [13–16]. Použití validace založené na časovém rozdělení dat navíc zajišťuje, že model je spolehlivý pro reálný provoz a není zatížen únikem informací mezi datovými sadami.



Obr. 7. Skutečné a predikované hodnoty BOD_{5_out} ; červená přerušovaná čára představuje ideální predikci ($y = x$)

Fig. 7. Actual vs. predicted BOD_{5_out} values; the red dashed line represents perfect prediction ($y = x$)

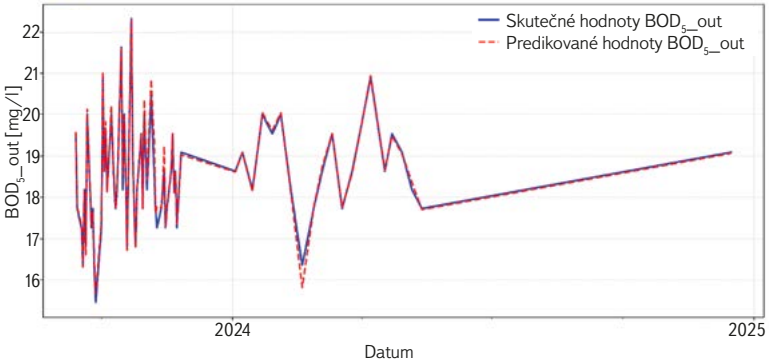
Význam proměnných a poznatky o procesu

Zařazení celkem 17 proměnných bylo pro úspěšnost modelu klíčové. Zejména začlenění proměnných založených na znalostech oboru, jako jsou $COD_removal_eff$, $COD_out_BOD_{5_int_ratio}$, $BOD_{5_int_x_pH}$ a změna vodivosti vyjádřená ukazatelem $Cond_ratio$ (rovnice 4, 7, 8 a 9 v tab. 2), poskytlo modelu hlubší kontext o biologickém stavu systému. Nízká hodnota MSE (0,05) dále potvrzuje stabilitu modelu a ukazuje, že model nevytváří velké, nahodilé chyby (odlehlé hodnoty).

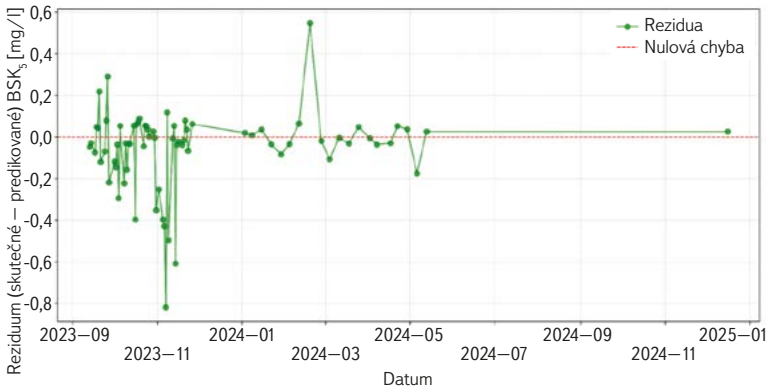
Tato robustnost je zásadní pro spolehlivé monitorování za provozních podmínek, protože umožňuje obsluze ČOV důvěřovat predikovaným hodnotám na odtoku a využívat je k proaktivnímu řízení a optimalizaci procesu.

Časová přesnost predikce

Obr. 8 znázorňuje časový vývoj skutečných a predikovaných hodnot BSK₅ během testovacího období. Model přesně sleduje dynamické chování procesu čištění, včetně krátkodobých výkyvů i postupných trendů. Navzdory drobným odchylkám model rychle reaguje na změny a nadále poskytuje přesné sledování vývoje hodnot. Analýza reziduí dále potvrzuje stabilní výkonnost modelu. Jak ukazuje obr. 9, rezidua jsou náhodně rozložena kolem nulové hodnoty, a to bez zjevného trendu nebo nárůstu rozptylu v čase. Svědčí to o homoskedasticitě chyb a absenci systematického zkreslení, což jsou klíčové vlastnosti pro spolehlivé využití modelu v reálných monitorovacích systémech.



Obr. 8. Časové porovnání skutečných a predikovaných hodnot BSK₅ na odtoku během testovacího období
Fig. 8. Temporal comparison of actual and predicted effluent BOD₅ over the test period



Obr. 9. Rezidua během testovacího období; chyby jsou soustředěny kolem nulové hodnoty bez systematického posunu
Fig. 9. Residuals over the test period; errors are centered around zero, with no systematic drift

Porovnání s odbornou literaturou

Tab. 4 porovnává výsledky této studie s nedávno publikovanými pracemi zaměřenými na predikci BSK₅/CHSK na ČOV. Náš model překonává několik dosavadních přístupů z hlediska hodnoty R². Je třeba poznamenat, že řada předchozích studií používala náhodné rozdělení dat na trénovací a testovací sadu, což může vést k nadhodnocení výkonnosti modelu v důsledku časového úniku informací. Naproti tomu tato studie využívá striktní chronologické rozdělení dat v poměru 80 %/20 %, jež zachovává časovou strukturu dat a zajišťuje schopnost modelu zobecňovat výsledky v podmínkách reálného využití. Náš model tak dosahuje vysoké přesnosti při současném použití přísných validačních postupů a zajištění plné reprodukovatelnosti výsledků.

Tab. 4. Srovnání výkonnosti modelů pro predikci BSK₅ podle publikovaných studií
Tab. 4. Comparative performance of BOD₅ prediction models in recent literature

Studie	Metoda	R ² [%]
Zhang et al. (2025) [13]	Random Forest	~89 %
Jafar et al. (2022) [14]	Deep Neural Network	89–93 %
Sethy et al. (2025) [15]	XGBoost	~93 %
Ali et al. (2025) [17]	CatBoost	~88 %
Tato studie	XGBoost (time-split)	96,95 %

Zobecnitelnost a přenositelnost modelu

Navzdory vysoké výkonnosti dosažené na studované lokalitě (R² = 0,9695) zůstává použitelnost modelu mimo tento kontext omezena jeho závislostí na datech specifických pro danou lokalitu, použitých při trénování. Predikční výkonnost je významně ovlivněna charakteristikami přítoku, provozními podmínkami a environmentálními faktory, které jsou obsaženy v trénovacím datovém souboru. Lze proto očekávat, že model bude spolehlivě fungovat především na čistírnách s obdobnou technologickou konfigurací a podobným provozním režimem.

Faktory omezující přenositelnost modelu zahrnují rozdíly ve složení přítoku, hydraulickém zatížení a technologickém uspořádání procesu (např. strategii provzdušňování, stáří aktivovaného kalu a stupně odstraňování živin), stejně jako regionální podmínky, jako jsou klimatická variabilita a charakter průmyslových odpadních vod. Využití modelu na ČOV s výrazně odlišnými charakteristikami bez jeho předchozí recalibrace může vést ke snížení přesnosti predikce nebo ke zvýšení nejistoty výsledků. Model je navíc primárně určen pro provozy se středním zatížením a nemusí být přímo použitelný na čistírnách s nízkým zatížením.

Budoucí výzkum by se měl zaměřit na validaci modelu na více lokalitách a na integraci metod doménové adaptace nebo transfer learningu, které by zvýšily robustnost modelu a umožnily jeho širší využití v různorodých systémech čištění odpadních vod.

ZÁVĚR

Tato studie ukazuje, že samostatný model XGBoost umí docílit vysoce přesné predikce BSK₅ na odtoku (BOD_{5-out}) na základě téměř desetileté řady provozních dat (2015–2024) z ČOV v plném provozním měřítku. Při hodnotách R² = 96,95 %, MSE = 0,05 a MAE = 0,146 mg/l dosažených na chronologicky vyčleněné testovací sadě vykazuje model mimořádnou výkonnost, aniž by bylo nutné využívat složité strategie ansámblového slučování modelů.

Klíčovým zjištěním je, že této úrovni přesnosti bylo dosaženo kombinací tvorby proměnných založené na znalostech oboru, zahrnující např. účinnost odstranění CHSK, změny pH a poměry CHSK/BSK₅, spolu s důslednou validací založenou na časovém rozdělení dat, která zajišťuje realistické hodnocení modelu v podmínkách predikce.

Přestože datový soubor neobsahoval údaje o rozpuštěném kyslíku (DO), model si zachoval vysokou predikční výkonnost, což naznačuje, že běžně monitorované proměnné, jako jsou CHSK na odtoku, vodivost a pH na přítoku,

poskytovaly dostatečné informace o provozních podmínkách a biologické aktivitě systému. Ačkoli je zásadní význam rozpuštěného kyslíku pro kinetiku procesů čištění nesporný, výsledky této studie naznačují, že vysoce korelované provozní parametry mohou při prediktivním modelování do určité míry kompenzovat absenci přímých měření DO.

Ještě důležitější je, že model funguje jako praktický virtuální senzor a poskytuje téměř okamžité odhady BSK₅ na základě přímých měření. Nahrzení pětidenní prodlevy laboratorního stanovení okamžitou predikcí vytváří podmínky pro využití BSK₅ nejen ke kontrole plnění emisních limitů, ale také k operativnímu řízení procesu. To zajišťuje:

- Včasnou detekci provozních poruch: Průběžná predikce BSK₅ umožňuje včas odhalit zhoršení biologické účinnosti způsobené změnami zatížení nebo hydraulickými poruchami, což přispívá k přijetí provozních opatření ještě před zhoršením kvality odtoku.
 - Optimalizaci spotřeby energie na provzdušňování: Predikované hodnoty BSK₅ umožňují přesnější řízení provzdušňování a zajišťují, že přívod kyslíku odpovídá aktuálnímu organickému zatížení. Tím se omezuje plýtvání energií způsobené nadměrným provzdušňováním při zachování vysoké účinnosti procesu.
 - Proaktivní úpravy provozu pro zajištění souladu s legislativními požadavky: Predikce BSK₅ umožňuje přijímat preventivní provozní opatření v situacích, kdy se očekávané hodnoty blíží limitům pro vypouštění odpadních vod. Tím je zaručeno trvalé plnění legislativních požadavků a současně se předchází vysokým pokutám i environmentálním a ekonomickým škodám.
- Tato studie ukazuje, že i relativně jednoduchý model může docílit vynikající přesnosti. Při správném zpracování dat a využití znalostí oboru může vhodně nastavený model XGBoost dosahovat srovnatelných nebo lepších výsledků než složitější metody, a zároveň je rychlejší, snáze interpretovatelný i využitelný. Díky tomu představuje perspektivní řešení pro škálovatelné aplikace v oblasti řízení a provozu systémů čištění odpadních vod.

Literatura

- [1] METCALF & EDDY, Inc., TCHOOBANOGLOUS, G., STENSEL, H. D., TSUCHIHASHI, R., BURTON, F. L. *Wastewater Engineering: Treatment and Resource Recovery*. 5. vyd. New York, NY: McGraw-Hill Education, 2014. ISBN 978-0-07-340118-8. Dostupné z: <https://www.mheducation.com/highered/product/wastewater-engineering-treatment-and-resource-recovery-metcalf-and-eddy.html?viewOption=student>
- [2] YOUNG, J. C., CLARK, J. W. Second-Order Analysis of BOD. *Journal of the Sanitary Engineering Division*. 1965, 91(1). Dostupné z: <https://doi.org/10.1061/JSEDAI.0000532>
- [3] PAINTER, H. A. A Review of Literature on Inorganic Nitrogen Metabolism in Microorganisms. *Water Research*. 1970, 4(6), s. 393–450. ISSN 0043-1354. Dostupné z: [https://doi.org/10.1016/0043-1354\(70\)90051-5](https://doi.org/10.1016/0043-1354(70)90051-5)
- [4] APHA. *Standard Methods for the Examination of Water and Wastewater*. 23rd ed. Washington, DC: American Public Health Association, 2017. ISBN 9780875532875.
- [5] JOUANNEAU, S., RECOULES, L., DURAND, M. J., BOUKABACHE, A., PICOT, V., PRIMAULT, Y., LAKEL, A., SENGELIN, M., BARILLON, B., THOUAND, G. Methods for Assessing Biochemical Oxygen Demand (BOD): A Review. *Water Research*. 2014, 49, s. 62–82. ISSN 0043-1354. Dostupné z: <https://doi.org/10.1016/j.watres.2013.10.066>
- [6] OLSSON, G., NIELSEN, M., YUAN, Z., LYNNGAARD-JENSEN, A., STEYER, J.-P. *Instrumentation, Control and Automation in Wastewater Systems*. IWA Publishing. 2005, 4. ISBN electronic 9781780402680. Dostupné z: <https://doi.org/10.2166/9781780402680>
- [7] CHING, P. M. L., SO, R. H. Y., MORCK, T. Advances in Soft Sensors for Wastewater Treatment Plants: A Systematic Review. *Journal of Water Process Engineering*. 2021, 44, 102367. ISSN 2214-7144. Dostupné z: <https://doi.org/10.1016/j.jwpe.2021.102367>
- [8] SHYU, H. Y., CASTRO, C. J., BAIR, R. A., LU, Q., YEH, D. H. Development of a Soft Sensor Using Machine Learning Algorithms for Predicting the Water Quality of an Onsite Wastewater Treatment System. *ACS Environmental Au*. 2023, 3(5), s. 308–318. Dostupné z: <https://doi.org/10.1021/acsenvironau.2c00072>
- [9] PATTNAIK, B. S., PATTANAYAK, A. S., UDGATA, S. K., PANDA, A. K. Machine Learning Based Soft Sensor Model for BOD Estimation Using Intelligence at Edge. *Complex & Intelligent Systems*. 2021, 7, s. 961–976. ISSN 2198-6053. Dostupné z: <https://doi.org/10.1007/s40747-020-00259-9>
- [10] JU, F., GUO, F., YE, L., XIA, Y., ZHANG, T. Metagenomic Analysis on Seasonal Microbial Variations of Activated Sludge from a Full-Scale Wastewater Treatment Plant over 4 Years. *Environmental Microbiology Reports*. 2014, 6(1), s. 80–89. ISSN 1758-2229. Dostupné z: <https://doi.org/10.1111/1758-2229.12110>

- [11] HENZE, M., GUJER, W., MINO, T., VAN LOOSDRECHT, M. *Activated Sludge Models ASM1, ASM2, ASM2d and ASM3*. IWA Publishing. 2000, 5. ISBN 1900222248. Dostupné z: <https://doi.org/10.2166/9781780402369>
- [12] DANTAS, M. S., CHRISTOFARO, C., OLIVEIRA, S. C. Artificial Neural Networks for Performance Prediction of Full-Scale Wastewater Treatment Plants: A Systematic Review. *Water Science and Technology*. 2023, 88(6), s. 1 447–1 470. ISSN 0273-1223. Dostupné z: <https://doi.org/10.2166/wst.2023.276>
- [13] ZHANG, X., ZHANG, Y., YANG, X., WANG, Z., LIU, X. Biochemical Oxygen Demand Prediction Based on Three-Dimensional Fluorescence Spectroscopy and Machine Learning. *Sensors*. 2025, 25(3), 711. ISSN 1424-8220. Dostupné z: <https://doi.org/10.3390/s25030711>
- [14] JAFAR, R., AWAD, A., JAFAR, K., SHAHROUR, I. Predicting Effluent Quality in Full-Scale Wastewater Treatment Plants Using Shallow and Deep Artificial Neural Networks. *Sustainability*. 2022, 14(23), 15598. ISSN 2071-1050. Dostupné z: <https://doi.org/10.3390/su142315598>
- [15] SETHY, K., MUDHOLKAR, M., MUDHOLKAR, P. K., BEHERA, P. K., KARTHIK, B., BALAMURALITHARAN, S. Predicting Water Quality Using Ensemble Machine Learning Models and Remote Sensing Data. *International Journal of Environmental Sciences*. 2025, 11(16s), s. 166–175. ISSN 2229-7359. Dostupné z: <https://doi.org/10.64252/cw5pq178>
- [16] ELMAADAWY, K., ABD ELAZIZ, M., ELSHEIKH, A. H., MOAWAD, A., LIU, B., LU, S. Utilization of Random Vector Functional Link Integrated with Manta Ray Foraging Optimization for Effluent Prediction of Wastewater Treatment Plant. *Journal of Environmental Management*. 2021, 298, 113520. ISSN 0301-4797. Dostupné z: <https://doi.org/10.1016/j.jenvman.2021.113520>
- [17] ALI, A., ZHOU, G., TAN, Y., ANTEZANA LOPEZ, F. P. Biochemical Oxygen Demand Estimation Using Explainable Ensemble Learning Methods. *Remote Sensing Applications: Society and Environment*. 2025, 41, 101835. ISSN 2352-9385. Dostupné z: <https://doi.org/10.1016/j.rsase.2025.101835>
- [18] GULSHIN, I., KUZINA, O. Machine Learning Methods for the Prediction of Wastewater Treatment Efficiency and Anomaly Classification with Lack of Historical Data. *Applied Sciences*. 2024, 14(22), 10689. ISSN 2076-3417. Dostupné z: <https://doi.org/10.3390/app142210689>
- [19] Algerian Executive Decree No. 06-141 of 19 April 2006 establishing the limit values for liquid industrial effluent discharges (official French text: *Décret exécutif n° 06-141 du 19 avril 2006 définissant les valeurs limites des rejets d'effluents liquides industriels*). *Journal Officiel de la République Algérienne Démocratique et Populaire (JORA)*, No. 26, 23 April 2006, pp. 4–6 (in French). Dostupné z: <https://www.joradp.dz/FTP/jo-francais/2006/F2006026.pdf>
- [20] KOUKARAS, P., TJORTJIS, C. Data Preprocessing and Feature Engineering for Data Mining: Techniques, Tools, and Best Practices. *AI*. 2025, 6(10), 257. ISSN 2673-2688. Dostupné z: <https://doi.org/10.3390/ai6100257>
- [21] ZHENG, A., CASARI, A. *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*. Sebastopol, CA: O'Reilly Media, 2018. ISBN 978-1491953242. Dostupné z: <https://www.oreilly.com/library/view/feature-engineering-for/9781491953235/>
- [22] BÜYÜKKEÇECİ, M., OKUR, M. C. A Comprehensive Review of Feature Selection and Feature Selection Stability in Machine Learning. *Gazi University Journal of Science and Technology*. 2023, 36(4), s. 1 506–1 520. Dostupné z: <https://doi.org/10.35378/gujs.993763>
- [23] HYNDEMAN, R. J., ATHANASOPOULOS, G. *Forecasting: Principles and Practice*. 3. vyd. Melbourne, Australia: OTexts, 2021. ISBN 9780987507136. Dostupné z: <https://otexts.com/fpp3/fpp2/>
- [24] GÉRON, A. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. 2. vyd. Sebastopol, CA: O'Reilly Media, 2019. ISBN 978-1492032649. Dostupné z: <https://www.amazon.com/Hands-Machine-Learning-Scikit-Learn-TensorFlow/dp/1492032646>
- [25] CHEN, T., GUESTRIN, C. XGBoost: A Scalable Tree Boosting System. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY: ACM, 2016, s. 785–794. ISBN 978-1450342322. Dostupné z: <https://doi.org/10.1145/2939672.2939785>
- [26] ROBESON, S. M., WILLMOTT, C. J. Decomposition of the Mean Absolute Error (MAE) into Systematic and Unsystematic Components. *PLOS ONE*. 2023, 18(2), e0279774. ISSN 1932-6203. Dostupné z: <https://doi.org/10.1371/journal.pone.0279774>

Autoři

Dr. Nassima Lekouaghet¹
ORCID: 0000-0003-1694-1430

Dr. Lakhdar Bouragbi^{2, 3}
ORCID: 0009-0002-3170-1275

Dr. Brahim Mahfoud³
✉ b.mahfoud@univ-bouira.dz
ORCID: 0000-0002-0709-8958

¹Laboratoř stavebního a hydraulického inženýrství,
Univerzita 8. května 1945, Guelma (Alžírsko)

²Katedra procesního inženýrství, Fakulta technologií,
Univerzita 20. srpna 1955, Skikda (Alžírsko)

³Laboratoř energetiky a mechaniky (LEM), Fakulta aplikovaných věd,
Univerzita Bouira (Alžírsko)

Originál příspěvku v angličtině prošel recenzním řízením.

DOI: 10.46555/VTEI.2026.05.003

ISSN 0322-8916/© 2026 Autoři. Tuto práci je kdokoli oprávněn šířit a využívat
za podmínek licence CC BY-NC 4.0

PREDICTING EFFLUENT BOD₅ IN WASTEWATER TREATMENT USING MACHINE LEARNING

LEKOUAGHET, N.¹; BOURAGBI, L.^{2, 3}; MAHFOUD, B.³

¹Laboratory of Civil Engineering and Hydraulics,
University of 8th May 1945, Guelma (Algeria)

²Department of Process Engineering, Faculty of Technology,
University 20 août 1955, Skikda (Algeria)

³Energy and Mechanics Laboratory (LEM), Faculty of Applied Science,
University of Bouira (Algeria)

Keywords: Biochemical Oxygen Demand (BOD₅) –
wastewater treatment – machine learning – soft sensor –
XGBoost – BOD₅ prediction – process optimization

Accurate and quick prediction of effluent Biochemical Oxygen Demand (BOD₅) is important for maintaining regulatory standards and improving the operation of wastewater treatment (WWT) plants. Traditional laboratory tests for BOD₅ take several days, causing delays in assessing water quality and adjusting plant performance. To overcome this issue, this study develops a machine learning model to estimate effluent BOD₅ using easily available plant data. A dataset from a large-scale wastewater treatment plant was used, comprising 12 variables from both influent and effluent data, such as pH, Chemical Oxygen Demand (COD), conductivity, Total Suspended Solids (TSS), BOD₅, and temperature, with effluent BOD₅ as the target variable. The data were cleaned and processed using the Interquartile Range (IQR) method to remove outliers, and new features were created to better reflect process behavior. Among several tested models, XGBoost showed the best performance, achieving an R² of 0.9695 and a Mean Squared Error (MSE) of 0.05, based on a time-based train/test split (80 %/20 %). The study demonstrates that machine learning can serve as a powerful predictive tool for BOD₅ estimation, significantly reducing the five-day analytical lag of laboratory testing and supporting faster, data-driven decision-making in WWT management.