Predykcja krótkoterminowa zużycia wody w budynkach wielorodzinnych z wykorzystaniem technik uczenia maszynowego

Short-term prediction of water consumption in multi-family buildings using machine learning techniques

SANDRA ŚMIGIEL, JUSTYNA STAŃCZYK, PAULINA DZIMIŃSKA, DAMIAN LEDZIŃSKI, TOMASZ ANDRYSIAK, PAWEŁ LICZNAR

DOI 10.36119/15.2023.12.11

W praktyce eksploatacyjnej systemów dystrybucji wody brak jest wdrożenia zaawansowanych narzędzi przetwarzania i analizowania danych monitoringowych. Dotyczy to wielu poziomów zarządzania wodociągami, w ramach których dochodzi do rejestracji pomiarów, najczęściej tworzących nieinterpretowalne zbiory danych. Wraz z pojawieniem się możliwości rejestracji danych, które można określić mianem wysokoczęstotliwościowych, istnieje potrzeba równoległego wdrażania odpowiednich technik z zakresu *data science*, stanowiących podstawę mądrych czy też inteligentnych sieci wodociągowych. By spełnić założenia dotyczące implementacji inteligencji na poziomie wodomierzy, koniecznym jest umożliwienie pomiaru zużycia wody z precyzyjnym interwałem pomiarowym oraz zaawansowanej analizy danych, które skutkować powinny efektywnym wnioskowaniem i zarządzaniem systemami dystrybucji wody. W niniejszym artykule zaprezentowano wyniki zastosowania modeli uczenia maszynowego w celu predykcji krótkoterminowej zużycia wody dla budynków wielorodzinnych. Do prognozowania zużycia wody wykorzystano modele liniowe, proste modele sieci neuronowych, algorytm najbliższych sąsiadów oraz drzewa decyzyjne. W ramach przeprowadzonych badań ocenie poddano cechy wyekstrahowane z przebiegu zużycia wody wraz z kombinacjami ich zestawów podawanymi na wejściu regresora. Zweryfikowano także, jak stopień agregacji danych oraz struktura budynku wpływają na błąd prognozowania.
Słowa kluczowe: sieci wodociągowe; uczenie maszynowe; wodomierze; zużycie wody

The operational practice of water distribution systems lacks the implementation of advanced tools for processing and analyzing monitored data. This is the case at many levels of water supply management, where measurements are recorded, most often creating uninterpretable data sets. With the arrival of data recording capabilities that can be described as high-frequency, there is a need for a simultaneous implementation of suitable data science techniques as the basis for smart water supply networks. To achieve the goals of implementing intelligence at the water meter level, it is necessary to allow measurement of water consumption with a precise measurement interval and advanced data analysis, which should result in effective inference and management of water distribution systems. This paper presents the results of the use of machine learning models to predict short-term water consumption for multifamily buildings. Linear models, simple neural network, nearest neighbour algorithm and decision trees were used to predict water consumption. The study evaluated features extracted from the water consumption waveforms and combinations of data sets given to the input of the regression model. It was also verified how the degree of data aggregation and the structure of the building influence the prediction error.
Keywords: water supply networks; machine learning; water meters; water consumption

Wprowadzenie

Rozwój technologii związanych z monitoringiem systemów dystrybucji wody, w tym dotyczących samych urządzeń pomiarowych jak i możliwości transmisji danych, spowodował znaczne zainteresowanie branży w aspekcie szeroko rozumianego zarządzania danymi, tzw. *big data*. W ślad za tym idą regulacje prawne, które obok konieczności zachowania odpowiednich zasad prawidłowej gospodarki wodomierzowej, mającej zapewnić jak najlepszą niezawodność wodomierzy i minimalizację błędu wskazań, wymuszają niejako stosowanie bardziej zaawansowanego opomiarowania infrastruktury technicznej. Przesłanką takiego stanu jest m.in. pojawienie się ustawy z dnia 20 kwietnia 2021 r. o zmianie ustawy o efektywności energetycznej oraz niektórych innych ustaw (Dz.U. 2021 poz. 868), będącej wynikiem implementacji do prawa krajowego dyrektywy Parlamentu Europejskiego i Rady (UE) ws. efektywności energetycznej

dr inż. Sandra Śmigiel – https://orcid.org/0000-0003-2459-5494 Politechnika Bydgoska im. Jana i Jędrzeja Śniadeckich, Wydział Inżynierii Mechanicznej, Bydgoszcz
dr inż. Justyna Stańczyk – https://orcid.org/0000-0002-5676-1787 Uniwersytet Przyrodniczy we Wrocławiu, Instytut Inżynierii Środowiska, Wrocław
mgr inż. Paulina Dzimińska – https://orcid.org/0000-0003-0427-5960 Miejskie Wodociągi i Kanalizacja w Bydgoszczy – sp. z o.o., Bydgoszcz
dr inż. Damian Ledziński – https://orcid.org/0000-0003-0796-4390, dr hab. inż. Tomasz Andrysiak, prof. PBŚ – https://orcid.org/0000-0001-8138-6619 Politechnika Bydgoska im. Jana i Jędrzeja Śniadeckich, Wydział Telekomunikacji, Informatyki i Elektrotechniki, Bydgoszcz
prof. dr hab. inż. Paweł Licznar – https://orcid.org/0000-0002-2559-5296 Wydział Instalacji Budowlanych, Hydrotechniki i Inżynierii Środowiska, Politechnika Warszawska, Warszawa.
Autor korespondencyjny/ corresponding author: sandra.smigiel@pbs.edu.pl

(Dyrektywa (UE) 2018/2002). W art. 16 niniejszej ustawy znajduje się zapis, że do dnia 1 stycznia 2027 r. właściciel lub zarządca budynku wielolokalowego zastąpi ciepłomierze, podzielniki kosztów ogrzewania lub wodomierze do pomiaru ciepłej wody urządzeniami umożliwiającymi zdalny odczyt ich stanu, co ma przyczynić się do maksymalizacji oszczędności energii. Pomimo, że rosnące ceny energii w aspekcie przygotowania ciepłej wody stanowią sensowny argument na poparcie tych działań, o tyle nadal trwają prace nad inteligencją mierników instalowanych w systemach dostarczających wodę do picia. Te z kolei, choć w dużej mierze już wchodzą w skład zdalnego, objezdnego systemu odczytu, uniemożliwiają realizację wysokoprecyzyjnych pomiarów, które docelowo mogą wraz z przetwornikami ciśnienia optymalizować pracę podsystemu pompowania wody, czyli jednego z najbardziej energochłonnych komponentów systemu zbiorowego zaopatrzenia w wodę. Jak wskazują wyniki badań, inteligentne planowanie pracy pomp na podstawie prognoz zużycia wody z wykorzystaniem technik ML (ang. Machine Learning) może zmniejszyć zużycie energii elektrycznej o ponad 15% [1]. Tego typu działania, mają tworzyć strukturę inteligentnych miast przy wykorzystaniu technologii IoT (ang. Internet of Things) [2], jako jednej z najprężniej rozwijających się w myśl idei Smart City, gałęzi technologii [3].

Inteligentne mierniki, z założenia mają poprawić wydajność energetyczną budynku i mogą wspierać podejmowanie odpowiednich działań w zakresie zrównoważonego rozwoju. Natomiast w kontekście wodociągów mogą umożliwić poznanie charakterystycznego wzorca zużycia wody, mogącego przyczynić się do optymalizacji dostaw wody. Choć przymiotniki „mądry" czy też „inteligentny" często są nadużywane w celach marketingowych, w branży wodociągowej posiadają jednoznacznie opisane przez Hamiltona (2021) [5] granice. I tak, w kontekście gospodarki wodomierzowej za system co najmniej mądry uznaje się taki, w ramach którego odbiorcy są opomiarowani z użyciem automatycznego systemu zdalnego odczytu AMR (ang. Automatic Meter Reading), realizowanego z wysoką dokładnością, w ramach rygorystycznie prowadzonej przez przedsiębiorstwo polityki wymiany wodomierzy. Z kolei inteligentny system posiada wdrożone kompletne algorytmy uczenia maszynowego (ML) wraz kontrolą wewnętrznego pomiaru przepływu, pomiaru i regulacji dokładności wskazań wodomierzy, z całkowitym auto-bilingiem i procedurą poboru opłat, bez jakiejkolwiek ingerencji człowieka. Można więc uznać, że bez wdrożenia technik uczenia maszynowego żadnego z mierników nie można określić mianem „intelligent". Inteligencja mierników objętości pobranej wody może być ukryta na poziomie nakładek wodomierzowych, których wymiana powinna przebiegać równolegle z koniecznością legalizacji wodomierzy [6]. Z analiz wykonanych przez Koral (2022) [7] wynika, że potencjał zastosowania inteligencji w gospodarce wodomierzowej uzależniony jest czasami od bardzo przyziemnych warunków, takich jak jakość łączności transmisji danych (głębokie, betonowe studnie tłumią sygnał) oraz okresowe braki w pomiarach. Inteligencja wbudowana w nakładkę powinna być odporna na tego typu niedogodności, a przede wszystkim powinna auto – adaptować się do zmieniających się w sposób dynamiczny warunków hydraulicznych panujących w przewodach. Inteligentny wodomierz powinien umożliwiać zarządzanie zużyciem wody, przeprowadzać prognozy poboru wody, kontrolować i diagnozować wycieki i sytuacje anomalne, ale także czuwać nad stanem technicznym samego urządzenia pomiarowego i sygnalizować jakiekolwiek problemy zarządcy.

Chociaż pojawiają się branżowe doniesienia dotyczące realizacji montażu coraz większej liczby zdalnych wodomierzy w warunkach polskich, brak jest wdrożonych procedur analizy zarejestrowanych danych. Zarządcy sieci i instalacji wodnych stają się coraz bardziej przekonani do korzyści płynących ze wsparcia, jakie oferują techniki uczenia maszynowego w codziennej eksploatacji. Także świadomość konsumentów nie jest tu bez znaczenia, bowiem jak udowodnili w swych badaniach Goulas i in. (2022) [4], ponad połowa uczestników przeprowadzonej ankiety zdecydowałaby się na instalację inteligentnego miernika, o ile wiązałoby się to z obniżeniem ponoszonych kosztów za zużytą wodę. Jak wskazali Siwoń i in. (2008) [8] pomiary w postaci szeregów czasowych parametrów pracy sieci wodociągowych są skarbnicą wiedzy o zjawiskach zachodzących w systemie, o ile posiada się umiejętność zgłębienia danych z zastosowaniem odpowiednich metod analitycznych, którymi docelowo stają się rozwiązania oferowane przez ML.

W prognozowaniu zużycia wody znajdują zastosowanie m.in. takie metody, jak: regresja liniowa LR (ang. Linear Regression), regresja wektorów wspierających SVR (ang. Support Vector Regression), drzewa losowe RF (ang. Random Forest) czy też perceptron wielowarstwowy MLP (ang. Multilayer Perceptron). Mogą one dawać porównywalne wyniki błędu dobowej prognozy [9] na poziomie analiz całych stref wodociągowych DMA (ang. District Metered Area), jednak w odniesieniu do predykcji zużycia wody w pojedynczych blokach mieszkalnych nie zostały one jeszcze dostatecznie zbadane. O ile jakość prognoz na poziomie większego obszaru zwodociągowania może być zadowalająca, w obrębie obiektów typu budynki wielorodzinne, ze względu na dużą losowość zużycia wody i skumulowany charakter danych [10], obarczone są znacznym błędem [11]. Nie bez znaczenia jest także interwał rejestrowanych pomiarów, na co zwrócono uwagę w branżowej monografii, powołując się na analogię pomiędzy układem krwionośnym człowieka, a systemem wodociągowym [8]. Pomimo tego, że system wodociągowy jest układem otwartym, a krwiobieg człowieka zamkniętym, poszukuje się przełożenia metod stosowanych do interpretacji napięcia mięśnia sercowego, rejestrowanego za pośrednictwem elektrokardiogramu EKG [12], [13], do detekcji stanu pracy wodociągów. Jednak by porównanie to było miarodajne, istnieje potrzeba wdrożenia wysokoczęstotliwościowych pomiarów [14], tak by krok czasowy był dostosowany do zachodzących w sieci zjawisk. Tymczasem, rejestrowanie danych w postaci skumulowanego zużycia w jednostce czasu, zwłaszcza w skali miesiąca, uniemożliwia tworzenie krótkoterminowych modeli predykcyjnych.

W artykule przedstawiono wyniki badań zastosowania modeli regresji, takich jak: kNN, drzewa decyzyjne, lasy losowe, perceptron wielowarstwowy, uogólniony model liniowy z rozkładem Poisson, modele LightGBM i XGBoost, uprzednio w większości wdrożonych do analiz sygnałów EKG [12], [13], [15], w kontekście predykcji 15-minutowej i 1-godzinowej zużycia wody dla budynków wielorodzinnych. Dla danych wejściowych dokonano przetwarzania wstępnego, opartego na agregacji i ekstrakcji cech z wykorzystaniem dyskretnej transformaty falkowej. Ponadto, zbadano skuteczność działania różnych modeli, kombinacji danych wejściowych (zarówno danych surowych, jak i wyekstrahowanych) oraz rozdzielczości danych wejściowych dla rozpatrywanych budynków. Wypracowano propozycje modeli oraz zestawu cech dla danych wejściowych istotnych w predykcji zużycia wody.

Metodyka badawcza

Przygotowanie zbioru danych do badań

W początkowym etapie badań przeprowadzono przetwarzanie wstępne danych, w ramach którego z analizowanego zbioru metodą okna kroczącego zostały wycięte szeregi czasowe trwające 7 dni dla docelowo 1 godzinnej prognozy w przód. Przyjęto, że przesuwanie okna będzie następowało co

1 godzinę. Parametrami wejściowymi modelu były wielkości (cechy) uzyskane z 7 dni, a wyjściami modelu było zużycie wody zagregowane do 15 minut oraz godziny. Uzyskane po wstępnym przetworzeniu dane poddano walidacji. Niepełne dane, tj. takie, w przypadku których brakowało wyników pomiarów, odrzucono.

W kolejnym etapie, dane podzielono na zbiór danych treningowych i testowych. Granicą podziału był czas. Dane treningowe stanowiły sumarycznie 18440 rekordów (pomiary z okresu styczeń-kwiecień) oraz odpowiednio dane testowe – 14196 rekordów (pomiary z okresu maj-czerwiec). Dokładną liczbę danych, wykorzystanych do procesu uczenia i testowania w odniesieniu do poszczególnych budynków przedstawiono w tabeli 1.

Tab. 1 Liczność zbiorów treningowego i testowego dla poszczególnych budynków

Lp.	ID budynku	Liczba rekordów treningowych	Liczba rekordów testowych	Suma rekordów zbioru
1	04u	830	798	1628
2	06u	429	514	943
3	08u	753	585	1338
4	09u	651	656	1307
5	12u	1274	1016	2290
6	13u	743	873	1616
7	14u	572	636	1208
8	15u	691	627	1318
9	16u	2263	1297	3560
10	17u	785	178	963
11	18u	599	683	1282
12	19u	1229	991	2220
13	20u	699	597	1296
14	22u	687	692	1379
15	24u	876	679	1555
16	25u	487	426	913
17	26u	728	424	1152
18	27u	1385	439	1824
19	28u	2076	1466	3542
20	30u	683	619	1302

W kolejnym etapie dokonano ekstrakcji cech. Jest to proces, który polega na przetwarzaniu danych prowadzący do uzyskania reprezentacji parametrów odpowiednio sprofilowanych do analizowanego problemu. Procedura ta ma za zadanie poprawić efektywność przetwarzania informacji i jednocześnie zoptymalizować modelowanie [16]. Ostatecznie, w wyniku wstępnego przetwarzania danych uzyskano następujące parametry:

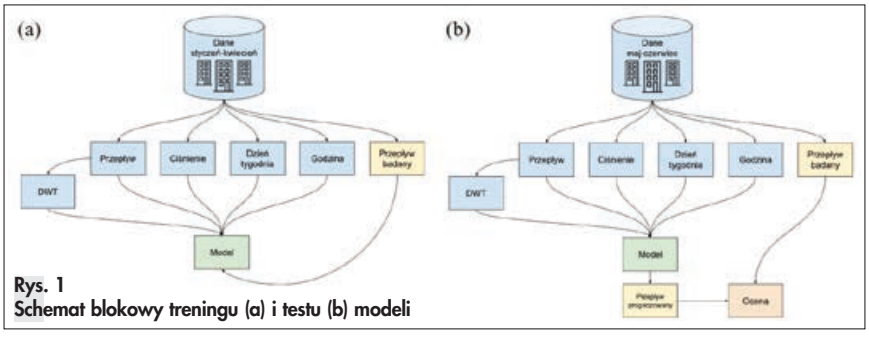
- zużycie wody o rozdzielczości 1 min. (surowe dane bezpośrednio z bazy),
- natężenia przepływu o rozdzielczościach 15, 30 i 60 min. (będących sumą zużycia 1-minutowego),
- ciśnienie o rozdzielczości 1 min. (surowe dane bezpośrednio z bazy),
- ciśnienia o rozdzielczościach 15, 30 i 60 min. (stanowiące średnią wartość ciśnienia 1-minutowego),
- dyskretna transformata falkowa DWT (ang. Discrete Wavelet Transform) zrealizowana dla szeregów czasowych natężenia przepływu o rozdzielczości 1 min; podobnie jak Stańczyk (2019) [17] wybrano falkę daubechies 4 (db4), 3 poziomy dekompozycji,
- dyskretna transformata falkowa DWT zrealizowana dla szeregów czasowych natężenia przepływu o rozdzielczościach 15, 30, 60 min. (jak wyżej, funkcja falkowa daubechies db4, 3 poziomy),
- dzień tygodnia,
- konkretna godzina doby.

Modele uczenia maszynowego

Do prognozowania zużycia wody wykorzystano modele uczenia maszynowego. Przebadano szereg modeli (algorytmów) regresji.

Pierwszy z nich, algorytm k-najbliższych sąsiadów (ang. kNN – k Nearest Neighbors), jest prostym modelem przypisującym nowe dane do klasy na podstawie k najbliższych sąsiadów, czyli zestawów danych najbardziej przybliżonych do siebie, czego miarą jest odległość euklidesowa. Kolejny rozpatrywany model to drzewa decyzyjne (ang. Decision Trees), czyli niezłożony model budujący drzewo decyzyjne, w którym każdy węzeł reprezentuje regułę decyzyjną. Następnie wdrożono lasy losowe RF (ang. Random Forest) składające się z zestawu drzew decyzyjnych, które są trenowane niezależnie od siebie, a wyniki z tych drzew są agregowane, co pomaga w redukcji wariancji i poprawie stabilności predykcji. Kolejny rozpatrywany model, perceptron wielowarstwowy (ang. MLP – Multilayer Perceptron), to rodzaj sieci neuronowej, która składa się z wielu warstw połączonych neuronów. Jest to model zdolny do rozwiązywania złożonych problemów, ale wymaga bardziej zaawansowanego przetwarzania danych. Bardziej złożone modele, takie jak uogólniony model liniowy GLM z rozkładem Poissona (ang. Generalized Linear Model with a Poisson distribution) stanowi liniowy model regresji, który wykorzystuje rozkład Poissona do modelowania zmiennej objaśnianej. Kolejny model, LightGBM LGBM (ang. Light Gradient-Boosting Machine) jest to model, który oparto na drzewach decyzyjnych i wykorzystuje algorytm GBDT (ang. Gradient Boosting Decision Trees). Znany jest także ze swojej wydajności i efektywności w obszarze gradient boosting. Ostatnim modelem wykorzystanym w badaniach był XGBoost (ang. eXtreme Gradient Boosting), który również oparty jest na algorytmie GBDT. Obydwa te modele w odróżnieniu od lasów losowych tworzą drzewa po kolei, przy czym kolejne są wynikiem wniosków wyciągniętych na podstawie poprzednich drzew. Różnica między nimi polega natomiast na sposobie tworzenia samego drzewa.

Modele były badane dla ustawień domyślnych, to znaczy nie był dokonywany tzw. tuning modeli, polegający na celowej ich optymalizacji. Rozpatrywano różne kombinacje cech wyekstrahowanych z danych wejściowych. Łącznie podejmowano przy siedmiu przyjętych modelach 3360 kombinacji. Każda z kombinacji zawsze zawierała dzień tygodnia i godzinę. Schemat blokowy trenowania (a) i testowania (b) modeli przedstawiono na rysunku 1.



Rys. 1
Schemat blokowy treningu (a) i testu (b) modeli

Zastosowane miary trafności prognoz

Zastosowane miary trafności prognoz stanowią kluczowy element oceny jakości modeli predykcyjnych. Uzyskane wyniki poddano ocenie przy wykorzystaniu metryk, takich jak: współczynnik determinacji R^2, średni błąd absolutny MAE (ang. Mean Absolute Error) oraz błąd średniokwadratowy MSE (ang. Mean Square Error). Poniżej przedstawiono równania (1)-(3), za pomocą których opisuje się wymienione miary trafności prognoz.

$$R^2 = 1 - \frac{\sum_{t=1}^{n}\left(Y_t - \hat{Y}_t\right)^2}{\sum_{t=1}^{n}\left(Y_t - \bar{Y}_t\right)^2}. \qquad (1)$$

$$MAE = \frac{1}{n}\sum_{t=1}^{n}\left|Y_t - \hat{Y}_t\right| \qquad (2)$$

$$MSE = \frac{1}{n}\sum_{t=1}^{n}\left(Y_t - \hat{Y}_t\right)^2 \qquad (3)$$

gdzie:

$\hat{Y}_t$ – wektor predykcji zużycia wody wygenerowany przez model, tj.

prognozowana wartość zużycia wody,

$\overline{Y}_t$ – średnia arytmetyczna empirycznych wartości zmiennej podlegającej prognozie,

Y_t – wektor obserwacji zużycia wody (zmienna podlegająca prognozie),

t – 1, 2, 3, …, n – wskaźnik czasu dla prognozowania zbioru n – elementowego tworzącego szereg czasowy.

Korzystając z tych miar trafności prognoz, możliwe jest dokładne zrozumienie, jak dobrze model radzi sobie z przewidywaniem rzeczywistych danych. Współczynnik determinacji R^2 mierzy procent zmienności objaśnianej przez model, co oznacza, że im wyższa wartość R^2, tym lepiej model wyjaśnia zmienność danych. Natomiast MAE i MSE pomagają określić, jak dokładne są prognozy i jak duże są błędy modelu. W przypadku metryki MAE, która mierzy średnią wartość bezwzględną różnicy między prognozami modelu a rzeczywistymi danymi, im mniejsza jest jej wartość, tym modele predykcyjne są bliżej rzeczywistych danych.

Wyniki badań

Predykcję krótkoterminową zużycia wody w budynkach wielorodzinnych z wykorzystaniem technik uczenia maszynowego weryfikowano etapowo. Oceny uzyskanych wyników dokonano z wykorzystaniem wyżej wymienionych miar. Na potrzeby omówienia wyników w tej części zdecydowano się operować metryką R^2, tj. współczynnikiem determinacji. R^2 jest najbardziej uniwersalną metryką, najczęściej wybieraną w publikacjach przez innych autorów, choć nie zawsze potrafi trafnie odzwierciedlić jakość prognoz. Dla porównania, MAE i MSE polaryzują wynik badania, ukierunkowując go na konkretny rodzaj błędu. Wyniki badań przedstawiono z uwzględnieniem wpływu ekstrakcji cech na modele wytrenowane przy wykorzystaniu metod uczenia maszynowego.

Analiza znaczenia cech w 15-minutowych prognozach zużycia wody

Początkowy etap badań objął swoim zakresem analizę dwudziestu budynków o charakterze mieszkalnym wielorodzinnym. Rozpatrywano wszystkie możliwe kombinacje uwzględniające cechy wyekstrahowane z danych wejściowych (tj. natężenie przepływu (Q), ciśnienie (p), transformaty falkowej dla natężenia przepływu DWT (Q) – wymienionych w podrozdziale dotyczącym przygotowania danych) dla siedmiu przyjętych modeli uczenia maszynowego. Analiz dokonano dla 15-minutowych prognoz zużycia wody. W tabeli 2 przedstawiono zestawienie wyników obliczeń pierwszych pięciu najlepszych kombinacji, ułożonych według rosnących wartości metryki oceny MAE. Odpowiednio w tabeli 3 zestawiono wyniki, ułożone według malejących wartości metryki oceny R^2.

Tab. 2 Analiza wyników pierwszych pięciu najlepszych kombinacji, pod względem miary MAE w 15-minutowych prognozach zużycia wody

ID budynku	Wyekstrahowane kombinacje cech	Model ML	R^2	MAE	MSE
22u	Q – p	MLP	0,402	18,936	691,985
17u	Q – DWT (Q)	LGBM	0,423	19,662	662,180
27u	Q	MLP	0,371	23,045	1057,257
26u	Q – DWT (Q) – p	LGBM	0,589	23,464	1078,275
18u	DWT (Q)	MLP	0,647	26,133	1177,511

gdzie: Q – natężenie przepływu, p – ciśnienie, DWT (Q) – dyskretna transformata falkowa dla natężenia przepływu

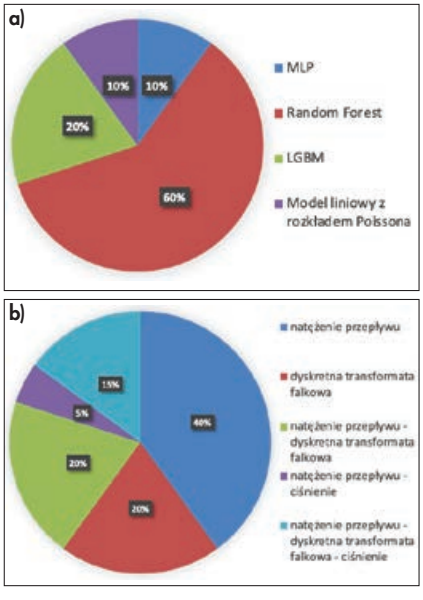
Tab. 3 Analiza wyników pierwszych pięciu najlepszych kombinacji pod względem miary R^2 w 15-minutowych prognozach zużycia wody

ID budynku	Wyekstrahowane kombinacje cech	Model ML	R^2	MAE	MSE
18u	DWT (Q)	MLP	0,647	26,133	1177,511
25u	Q – DWT(Q)	Random Forest	0,636	30,772	1746,068
14u	Q	Random Forest	0,617	33,714	1913,833
24u	Q – DWT (Q)	Random Forest	0,617	32,615	1942,967
19u	Q	LGBM	0,611	30,470	1663,699

gdzie: Q – natężenie przepływu, p – ciśnienie, DWT (Q) – dyskretna transformata falkowa dla natężenia przepływu

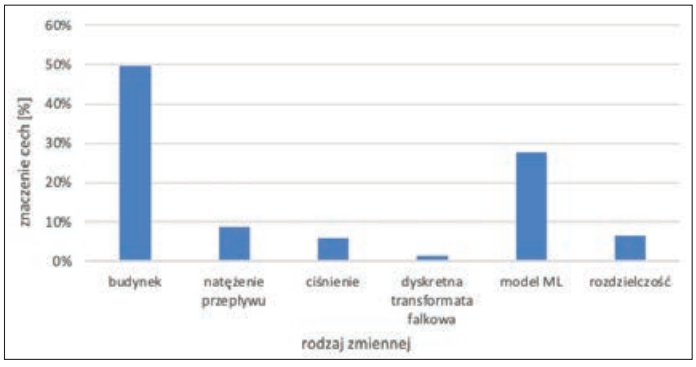
Wyniki zawarte w tabelach 2-3 potwierdzają stochastyczność zużycia wody. Losowość jej poboru na poziomie pojedynczego budynku, praktycznie uniemożliwia uzyskanie lepszej jakości prognoz, co potwierdzono także w badaniach Silva i in. (2021) [11]. Cykl okołodobowy rytmu życia mieszkańców pozwala na wyodrębnienie pewnej okresowości, zwłaszcza 24, 12 i 8 godzinnej [18], jednak w niewielkiej skali czasu zmienność związana przede wszystkim ze zwyczajami konsumentów jest kluczowa. Przeprowadzone dotychczas badania nad determinizmem zużycia wody w gospodarstwach domowych ograniczają się jedynie do konkretnych studiów przypadku, w związku z czym interpretacja tego zjawiska w mikroskali jest skomplikowana [19]. Jak udowodnili w badaniach Stańczyk i in. (2023) [20] determinizm zużycia wody uwidacznia się dopiero podczas analiz minimum jedenastu średnich wielkościowo budynków typu wielorodzinnego.

Szczegółowa analiza wybranych modeli uczenia maszynowego i cech wyekstrahowanych z przebiegów zużycia wody została przedstawiona na rysunku 2. Uwzględniono wyniki posortowane po najwyższej wartości miary R^2. Miara R^2, jak opisano wcześniej, jest najczęściej używaną miarą trafności prognoz. Sprowadza ona bowiem wynik do jednej liczby, która w pewnym stopniu określa całkowitą sprawność predyktora. Najczęściej (w 60%), najlepsze wyniki uzyskiwały modele lasów losowych (Random Forest) w połączeniu z parametrem natężenia przepływu. Najgorzej wypadły modele, takie jak MLP oraz model liniowy z rozkładem Poissona (po 10%) z uwzględnieniem danych wejściowych jakimi były natężenie przepływu i ciśnienie. Brak znaczącej przydatności do tego typu analiz parametru ciśnienia związany jest z dynamiką hydraulicznego przepływu w przewodach. Choć wielkość ta była próbkowana w interwale 1-min., stanowiła wartość chwilową, która mogła w danym momencie być odniesiona do warunków statycznych panujących w instalacji.



Rys. 2
Analiza częstości wystąpień modeli ML (a) i wyekstrahowanych cech (b) w 15-minutowych prognozach zużycia wody

Uzupełnieniem badań są wyniki graficzne ukazujące znaczenie wpływu cech wyekstrahowanych z przebiegu zużycia wody.



Rys. 3
Analiza znaczenia cech w 15-minutowych prognozach zużycia wody

Znaczenie cech (ang. feature importances) w tym podejściu wskazuje jak ważna była każda z przyjętych cech w modelowaniu. Zasadniczo, zabieg ten ma na celu podkreślenie znaczenia konkretnej zmiennej wchodzącej na wejściu do modelu. Procentowy rozkład znaczenia cech dla przyjętych modeli uczenia maszynowego przedstawiono na rysunku 3.

Przeprowadzona analiza potwierdziła, że w dużej mierze (prawie 50%) istotne są zwyczaje mieszkańców co do poboru wody w obrębie każdego z budynków, co jednocześnie świadczy o tym, że nawet w budynkach o podobnym typie, główną rolę odgrywają czynniki behawioralne i te związane ze strukturą budynku, liczbą mieszkańców oraz ich wiekiem, standardem wyposażenia w urządzenia sanitarne, rokiem budowy [21]. Kolejnym, ważnym czynnikiem są same modele ML, ponieważ posiadają one różne zdolności prognostyczne, i jak wykazano wyżej nie wszystkie pozwalają na uzyskanie zadowalającej jakości predykcji.

Analiza 1-godzinowych prognozach zużycia wody

Na podstawie wcześniej uzyskanych wyników badań zdecydowano się na analizę 1-godzinowych prognoz zużycia wody dla wspomnianych wcześniej dwudziestu budynków o charakterze mieszkalnym wielorodzinnym. Podobnie jak poprzednio rozpatrywano wszystkie kombinacje uwzględniające wyekstrahowane cechy i przyjęte modele.

W tabeli 4 przedstawiono zestawienie wyników obliczeń pierwszych pięciu najlepszych kombinacji, ułożone według rosnących wartości metryki oceny MAE. Odpowiednio została na rysunku 4. Uwzględniono wyniki posortowane po najwyższej wartości miary R^2, zgodnie z uzasadnieniem jak poprzednio. Zrealizowane badania wskazują, że najwyższe wyniki miary R^2 zapewniał najczęściej algorytm lasów losowych, co pokrywa się z analizą dla 15-minutowych prognoz zużycia wody. Odpowiednio, najlepiej w tym ujęciu radziły sobie modele w połączeniu z cechami natężenia przepływu i dyskretnej transformaty falkowej dla natężenia przepływu.

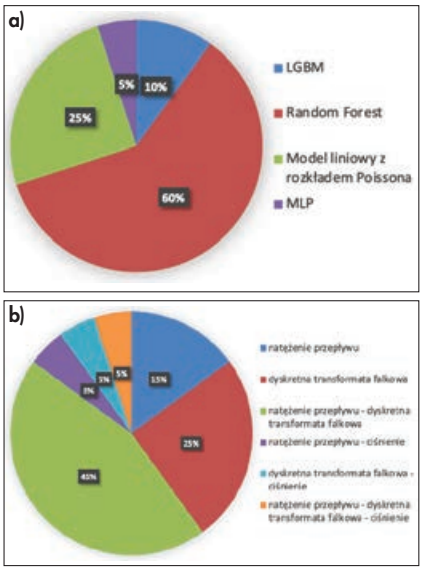
Tab. 5 Analiza wyników pierwszych pięciu najlepszych kombinacji pod względem miary R^2 w 1-godzinowych prognozach zużycia wody

ID budynku	Wyekstrahowane kombinacje cech	Model ML	R^2	MAE	MSE
25u	Q – DWT (Q)	LGBM	0,820	76,143	10704,547
24u	Q – DWT (Q)	Random Forest	0,815	79,990	11258,287
14u	Q	Random Forest	0,812	82,462	11345,437
19u	DWT (Q)	Model liniowy z rozkładem Poissona	0,803	76,007	10325,496
16u	Q – DWT (Q)	Random Forest	0,803	77,572	10920,205

gdzie: Q – natężenie przepływu, p – ciśnienie, DWT (Q) – dyskretna transformata falkowa dla natężenia przepływu

w tabeli 5 zestawiono wyniki, ułożone według malejących wartości metryki oceny R^2.

Analiza poszczególnych modeli uczenia maszynowego i cech wyekstrahowanych z przebiegów zużycia wody przedstawiona



Rys. 4
Analiza częstości wystąpień (a) modeli ML i (b) wyekstrahowanych cech w 1-godzinowych prognozach zużycia wody

Podobnie jak poprzednio, dokonano analiz znaczenia wpływu cech wyekstrahowanych z 1-godzinowych prognoz zużycia wody (rysunek 5). Niezmiennie największe znaczenie ma cecha taka jak swoisty charakter budynku, głównie w kontekście zwyczajów mieszkańców i typu wykorzystanego modelu uczenia maszynowego. Jednocześnie należy zauważyć, że wzrosło znaczenie rozdzielczości wykonanych pomiarów z około 7% do ponad 11%, co także świadczy o istnieniu minimalnego progu agregacji danych, przy którym uwypuklony zaczyna być deterministyczny charakter zużycia wody.

Prognoza zużycia wody

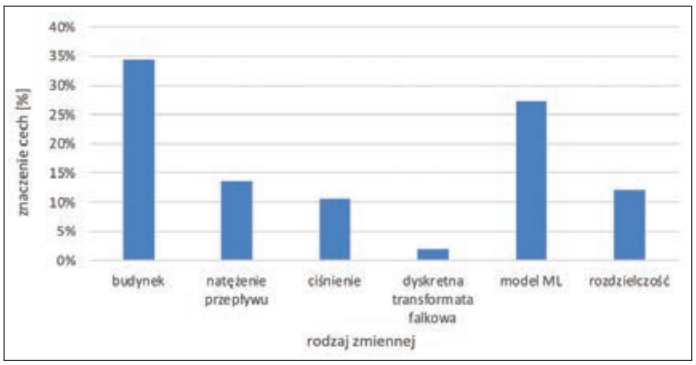
W kolejnym etapie badań, z uwagi na mnogość wyników, skupiono się na budynku, w ramach którego udało się uzyskać najbardziej trafną prognozę zużycia wody. W przypadku predykcji 15-minutowej najlepszy wynik miary R^2 uzyskał budynek o ID 18u. Budynek ten posiada 83 mieszkań, ze 151 mieszkańcami i stanowi jeden z największych spośród analizowanych obiektów.

Dla budynku o ID 18u najlepszy model uczenia maszynowego to perceptron wielowarstwowy (MLP), wytrenowany na danych wejściowych, stanowiących dyskretną transformatę falkową dla natężenia przepływu o rozdzielczości 15 minut. W tym wypadku współczynnik R^2 wynosił 0,647. Przykładowy przebieg zużycia wody dla tego budynku przedstawiono na rysunku 6. Jak można zauważyć największa różnica pomiędzy wartościami obserwowanymi, a prognozowanymi istnieje w godzinach około 9:00 – 11:00 oraz pomiędzy 19:00 a 21:00. Tłumaczyć to można nieco zachwianą zmiennością zużycia wody

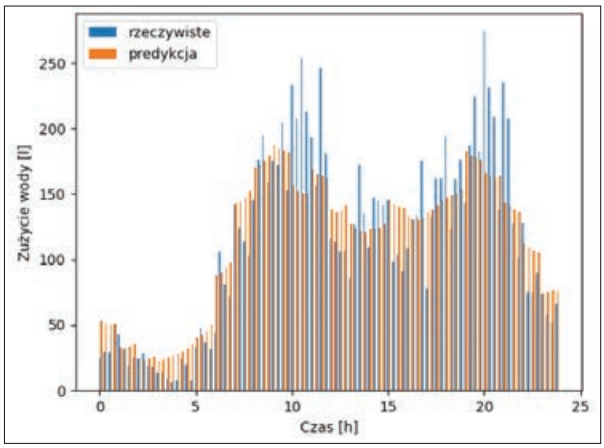
Tab. 4 Analiza wyników dla pierwszych pięciu najlepszych kombinacji pod względem miary MAE w 1-godzinowych prognozach zużycia wody

ID budynku	Wyekstrahowane kombinacje cech	Model ML	R^2	MAE	MSE
17u	Q – DWT (Q)	MLP	0,708	43,841	3322,753
22u	Q – DWT (Q) – p	MLP	0,647	49,825	4386,106
26u	Q – DWT (Q)	MLP	0,794	59,899	6377,836
27u	Q – DWT (Q) – p	LGBM	0,596	61,053	7334,793
15u	Q – DWT (Q) – p	MLP	0,763	62,134	7196,860

gdzie: Q – natężenie przepływu, p – ciśnienie, DWT (Q) – dyskretna transformata falkowa dla natężenia przepływu



Rys. 5
Analiza znaczenia cech w 1-godzinowych prognozach zużycia wody



Rys. 6
Wyniki prognoz 15-minutowego zużycia wody dla budynku ID 18u

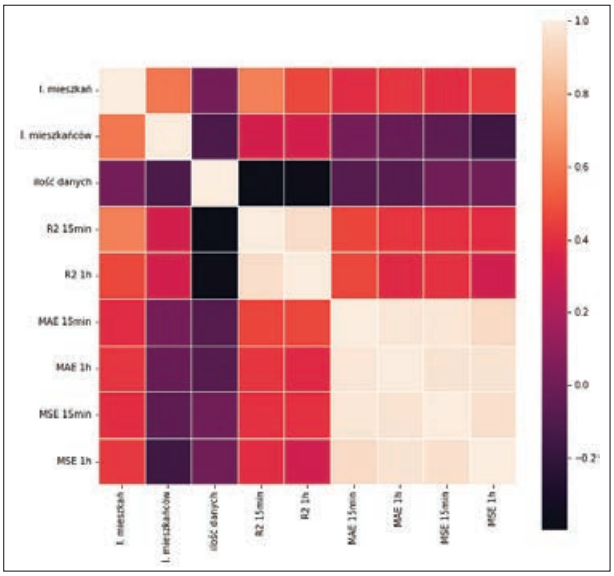
spowodowaną pandemią COVID-19, ponieważ badania były prowadzone w okresie *lockdownu*, co wymusiło na użytkownikach sieci wodociągowych niejako zmianę swoich przyzwyczajeń, widoczną zwłaszcza w "pikach" porannego i wieczornego zużycia wody, co szczegółowo opisała Dzimińska i in. (2021) [22].

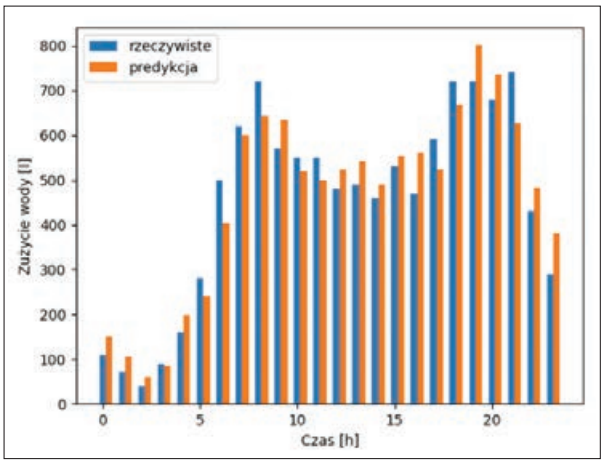
Dla predykcji 1-godzinnej, najlepszy wynik miary R^2 uzyskano dla budynku ID 25u. Budynek ten strukturalnie liczy 82 mieszkań, z 151 mieszkańcami. Przykładową prognozę godzinową dla tego budynku przedstawiono na rysunku 7.

Dla tego przypadku najlepszy model uczenia maszynowego to LightGBM (LGBM), trenujący z danymi wejściowymi w postaci natężenia przepływu i dyskretnej transformaty falkowej wykonanej na danych o przepływie o rozdzielczości 60 minut. W tym wypadku współczynnik R^2 wynosił 0,819. Jak wskazuje wcześniej przedstawiony wykres znaczenia cech (rysunek 4), na możliwość zmniejszenia błędu predykcji, duże znaczenie mógł mieć fakt, że sygnał zużycia wody został poddany DWT. Pozwoliło to na wyeliminowanie wpływu obarczenia szeregów czasowych rejestrowanych z dużą częstotliwością, białym szumem pomiarowym o płaskim widmie częstotliwościowym [17].

W obliczu widocznej różnicy w jakości prognoz zużycia wody dla poszczególnych budynków, na koniec postanowiono jednoznacznie zweryfikować, czy liczba mieszkań, mieszkańców, długość zarejestrowanych szeregów czasowych może wpływać na wyniki modelowania. Szczegółowo studium analizowanego przypadku zostało opisane w publikacji Dzimińska i in. (2022) [14]. Na rysunku 8 przedstawiono macierz korelacji pomiędzy liczbą mieszkań, liczbą mieszkańców, liczbą danych treningowych, a uzyskanymi błędami modelowania. Pomijając zależności pomiędzy poszczególnymi miarami błędu predykcji, na rysunku widać, że współczynnik R^2 dla prognoz 15 minut w dużym stopniu zależy od liczby mieszkań w budynku, a w ogóle nie zależy od liczby danych jaką dysponujemy. W mniejszym stopniu od liczby mieszkań zależy R^2 dla predykcji 1-godzinowej, jednak kwestia liczby mieszkańców wpływa na porównywalnym poziomie jak w przypadku predykcji 15-minutowych. Silna korelacja pomiędzy liczbą mieszkań, a mniejsza związana z liczbą mieszkańców, świadczyć może o tym, że w obrębie mieszkań może być różna liczebność mieszkańców, co wynika z czysto demograficznych przesłanek. Liczba mieszkańców ma wpływ na poprawę R^2, jednak wpływ ten jest znikomy dla błędów MAE i MSE. Długość zastosowanych do

Rys. 8
Macierz korelacji struktury budynków i osiąganych błędów predykcji





Rys. 7
Wyniki prognoz godzinowego zużycia wody dla budynku ID 25u

prognozowania szeregów czasowych nie ma znaczenia lub wpływ ten jest znikomy, co pozwala wnioskować, że modele ML nie wymagają zbyt dużych zbiorów danych o zużyciu wody, by dokonywać trafnych predykcji.

Podsumowanie

Wdrażanie do codziennej praktyki eksploatacyjnej systemów dystrybucji wody wysokoczęstotliwościowych pomiarów wymaga równoległej implementacji modeli uczenia maszynowego (ML). Obecność ML w sferze wodociągowej, będąca kluczowym elementem inteligentnych sieci, ma na celu usprawnienie analizy oraz interpretacji zjawisk zachodzących w sieciach wodociągowych, zarówno w kontekście predykcji panujących parametrów hydraulicznych jak i detekcji anomalii. Analogia struktury systemu wodociągowego do układu krwionośnego i pracy serca człowieka, pozwoliła na weryfikację czy metody wcześniej wykorzystane w medycynie mają przełożenie na czysto techniczne aspekty. Przeprowadzono badania w celu oceny sensowności ekstrakcji cech oraz możliwości trafnej prognozy zużycia wody w różnych przedziałach czasowych za pomocą starannie wyselekcjonowanych zestawów cech.

Wyniki badań jednoznacznie wskazują, że lepszą jakość predykcji, wyrażoną największym współczynnikiem R^2, można uzyskać dla prognoz godzinowych (R^2=0,820), a nieco gorszą dla 15-minutowych (R^2=0,647). Choć wydaje się to oczywiste, mając na uwadze losowość zużycia wody, jak wskazuje Li i in. (2020) [23] o mądrości systemów wodociągowych świadczy także jak najmniejsze opóźnienie czasowe pomiędzy danymi wejściowymi, a tymi diagnostycznie użytecznymi, dlatego też autorzy wdrożyli do badań dane o małej rozdzielczości czasowej. Tymczasem, jak sugerują wyniki przeprowadzonych badań szeregi czasowe są obarczone szumem pomiarowym, a dekompozycja sygnału z wykorzystaniem dyskretnej transformaty falkowej (DWT) w stosunku do natężenia przepływu w parze z natężeniem przepływu, pozwala na osiągnięcie nieco lepszych wyników. Ciśnienie jako jeden z badanych parametrów, ze względu na zachowanie statyczne w odniesieniu do przyłącza wodociągowego, wpływało na predykcję jedynie w 10%. Kluczowe natomiast okazały się zachowania mieszkańców w poborze wody, potwierdzając, że na poziomie pojedynczego budynku nie jest możliwe osiągnięcie jakiegokolwiek determinizmu w krótkim przedziale czasowym, a charakter poborów wody różni się pomiędzy budynkami, nawet tego samego typu.

Nie bez znaczenia jest również struktura budynku, która ma wpływ na błąd predykcji zużycia wody. Wyniki badań wskazują, że w największej mierze istotna jest liczba mieszkań, a następnie dopiero liczba mieszkańców zamieszkujących dany obiekt. Najmniejszy wpływ na jakość prognoz wykazała długość zarejestrowanych szeregów czasowych, sugerując tym samym, że zarówno w procesie uczenia jak i testowania wykorzystano odpowiednią liczbę danych, stanowiących ośmiomiesięczny okres pomiarów.

Na podstawie wyników badań stwierdzono użyteczność ekstrakcji cech i zastosowania metod ML do prognozy krótkoterminowej zużycia wody, nawet do tak losowych danych jak zużycie wody w mikroskali pojedynczych budynków. W przyszłych badaniach planuje się dopracowanie pozyskanych modeli ML nie tylko pod względem skuteczności ich działania, ale także pod względem optymalizacji w zakresie czasu potrzebnego na trening oraz, co ważniejsze, czasu i zasobów potrzebnych na inferencję (predykcję), tak aby możliwe było zastosowanie ich w urządzeniach IoT.

Podziękowanie

Autorzy pracy składają podziękowania Miejskim Wodociągom i Kanalizacji w Bydgoszczy Sp. z o.o. za udostępnienie poligonu badawczego i współpracę.

LITERATURA

[1] E. Sarmas, E. Spiliotis, V. Marinakis, G. Tzanes, J.K. Kaldellis, H. Doukas, „ML-based energy management of water pumping systems for the application of peak shaving in small-scale islands", Sustainable Cities and Society, 82, 2022. DOI: 10.1016/j.scs.2022.103873.

[2] Andrić, A. Vrsalović, T. Perković, M.A. Čuvić, P. Šolić, „IoT approach towards smart water usage", Journal of Cleaner Production, 367, 2022. DOI: 10.1016/j.jclepro.2022.133065.

[3] T. Cichoń, J. Królikowska, „Remote Reading of Water Meters as an Element of a Smart City Concept", Rocznik Ochrona Środowiska, 23, 2021. DOI: 10.54740/ros.2021.060.

[4] A. Goulas, D. Goodwin, C. Shannon, P. Jeffrey, H.M. Smith, „Public perceptions of household IoT smart water "event" meters in the UK—implications for urban water governance", Frontiers in Sustainable Cities, 4, 2022. DOI: 10.3389/frsc.2022.758078.

[5] S. Hamilton, B. Charalambous. G.Wyeth, „Improving Water Supply Networks: Fit for Purpose Strategies and Technologies", 2021. DOI: 10.2166/9781780409207.

[6] T. Cichoń, J. Królikowska, „Problemy eksploatacji wodomierzy w aspekcie ich niezawodności", Instal, pp. 41-44, 2022. DOI: 10.36119/15.2022.12.7.

[7] W. Koral, „Analiza danych z systemu stacjonarnego odczytu wodomierzy dla średniej wielkości miasta", Instal, pp. 46-52, 2022. DOI: 10.36119/15.2022.12.8.

[8] Z. Siwoń, J. Łomotowski, W. Cieżak, P. Licznar, J. Cieżak, „Analiza i prognozowanie rozbiorów wody w systemach wodociągowych", Studia z Zakresu Inżynierii, ISSN 0137-5393, 2008, Warszawa.

[9] J. Stańczyk, J. Kajewska-Szkudlarek, P. Lipiński, P. Rychlikowski, „Improving short-term water demand forecasting using evolutionary algorithms", Scientific Reports, 12(1), 2022. DOI: 10.1038/s41598-022-17177-0.

[10] J. Wawrzosek, S. Ignaciuk, J. Stańczyk i J. Kajewska-Szkudlarek, „Water Consumption Variability Based on Cumulative Data From Non-simultaneous and Long-term Measurements", Water Resources Management, pp. 2799–2812, 35, 2021. DOI: 10.1007/s11269-021-02868-6.

[11] K.P.T.D Silva, A. Kalbusch, E. Henning, G.A.L. Menezes, „Modeling water consumption in multifamily buildings: a case study in Southern Brazil", Urban Water Journal, 18(10), pp. 783-795, 2021. DOI: 10.1080/1573062X.2021.1934040.

[12] K. Pałczyński, S. Śmigiel, D. Ledziński, S. Bujnowski, „Study of the few-shot learning for ECG classification based on the PTB-XL dataset", Sensors, 22(3), 2022. DOI: 10.3390/s22030904.

[13] S. Śmigiel, K. Pałczyński, D. Ledziński, „ECG signal classification using deep learning techniques based on the PTB-XL dataset", Entropy, 23(9), 2021. DOI: 10.3390/e23091121.

[14] P. Dzimińska, J. Stańczyk, S. Drzewiecki, P. Licznar, „Wykorzystanie systemu rejestracji danych z dużą częstotliwością do analizy nierównomierności zużycia wody", Instal, pp. 24-30, 2022. DOI: 10.36119/15.2022.1.3.

[15] S. Śmigiel, „ECG Classification Using Orthogonal Matching Pursuit and Machine Learning", Sensors, 22(13), 2022. DOI: 10.3390/s22134960.

[16] M. Grzegorowski, „Governance of the redundancy in the feature selection based on rough sets' reducts", In International Joint Conference on Rough Sets, pp. 548-557, 2016. DOI: 10.1007/978-3-319-47160-0_50.

[17] J. Stańczyk, „Detekcja stanów pracy sieci wodociągowej z wykorzystaniem cyfrowego przetwarzania sygnałów", Praca doktorska, Politechnika Wrocławska, Wrocław, 2019.

[18] K. Ramm, P. Licznar, „Analiza baz danych zapisów wodomierzy do dynamicznego modelowania rozbiorów wody", Wydawnictwo Seidel-Przywecki, 2021.

[19] Cominola, L. Preiss, M. Thyer, H.R. Maier, P. Prevos, R.A. Stewart, A. Castelletti, „The determinants of household water consumption: A review and assessment framework for research and practice", npj Clean Water, 6(1), 11, 2023. DOI: 10.1038/s41545-022-00208-8.

[20] J. Stańczyk, K. Pałczyński, P. Dzimińska, D. Ledziński, T. Andrysiak, P. Licznar, „The impact of the number of high temporal resolution water meters on the determinism of water consumption in a district metered area", Scientific Reports, 2023. DOI: 10.1038/s41598-023-46086-z (w trakcie publikacji).

[21] T.F. Dias, A. Kalbusch, E. Henning, „Factors influencing water consumption in buildings in southern Brazil", Journal of Cleaner Production, 184, pp. 160-167, 2018. DOI: 10.1016/j.jclepro.2018.02.093.

[22] P. Dzimińska, S. Drzewiecki, M. Ruman, K. Kosek, K. Mikołajewski, P. Licznar, „The Use of Cluster Analysis to Evaluate the Impact of COVID-19 Pandemic on Daily Water Demand Patterns", Sustainability, 13(11), 2021. DOI: 10.3390/su13115772.

[23] J. Li, X. Yang, R. Sitzenfrei, „Rethinking the framework of smart water system: A review", Water, 12(2), 2020. DOI: 10.3390/w12020412.

■