



**Politechnika
Śląska**

PROJEKT INŻYNIERSKI

Modelowanie i analiza procesów związanych z działaniem systemu centralnego ogrzewania oraz przygotowania wody użytkowej w budynku mieszkalnym z wykorzystaniem uczenia maszynowego

Jakub KULA

Nr albumu: 296849

Kierunek: Automatyka i Robotyka

Specjalność: Technologie Informatyczne

PROWADZĄCY PRACĘ

dr inż. Szymon Ogonowski, prof. PŚ

KATEDRA Katedry Pomiarów i Systemów Sterowania

Wydział Automatyki, Elektroniki i Informatyki

Gliwice 2024

Tytuł pracy

Modelowanie i analiza procesów związanych z działaniem systemu centralnego ogrzewania oraz przygotowania wody użytkowej w budynku mieszkalnym z wykorzystaniem uczenia maszynowego

Streszczenie

(Streszczenie pracy – odpowiednie pole w systemie APD powinno zawierać kopię tego streszczenia.)

Słowa kluczowe

(2-5 słów (fraz) kluczowych, oddzielonych przecinkami)

Thesis title

Modeling and analysis of the heating and hot water preparation processes in residential building using machine learning

Abstract

(Thesis abstract – to be copied into an appropriate field during an electronic submission – in English.)

Key words

(2-5 keywords, separated by commas)

Spis treści

1	Wstęp	1
1.1	Cel i zakres pracy	1
1.2	Aktualny stan wiedzy	2
1.2.1	Przewidywania zużycia ciepłej wody użytkowej	2
1.2.2	Modelowanie zasobnika CWU	2
1.3	Charakterystyka rozdziałów	3
2	Narzędzia zastosowane w pracy	5
2.1	Język programowania	5
2.2	Biblioteka uczenia maszynowego	6
2.3	Narzędzia pomocnicze	7
2.3.1	Pandas i NumPy	7
2.3.2	Matplotlib	8
2.3.3	Scikit-learn, ELI5 i LIME	8
2.3.4	CUDA i Conda	8
3	Modelowanie sieci neuronowej	9
3.1	Metodologia projektowania modelu sieci neuronowej	9
3.2	Dane wejściowe i proces ich przetwarzania	11
3.3	Projektowanie i ocena modeli	14
3.4	Walidacja i próby dostrajania (?)	21
4	Wpływ warunków pogodowych	25
4.1	Permutacyjna ważność cech	27
4.2	Badanie wag wejściowych pierwszej warstwy	28
4.3	LIME	31
4.4	Ocena skuteczności metod	33
5	Modelowanie dynamiki temperatury wody użytkowej	37
5.1	Wyniki symulacji	39
5.2	Optymalizacja	41

6 Podsumowanie i wnioski	47
Bibliografia	50
Spis skrótów i symboli	53
Lista dodatkowych plików, uzupełniających tekst pracy	55
Spis rysunków	57
Spis tabel	59

Rozdział 1

Wstęp

1.1 Cel i zakres pracy

TODO NAPISAĆ wprowadzenie w problem/zagadnienie

1.2 Aktualny stan wiedzy

1.2.1 Przewidywania zużycia ciepłej wody użytkowej

Modele oparte na zebranych danych są szeroką kategorią modeli statystycznych i analitycznych, które wykorzystują dane zebrane z różnych źródeł do formułowania wniosków. Edwards et al.[1] stworzyli roczne profile poboru wody, na podstawie danych pomiarowych z 73 domostw w Québec (Kanada). Utworzone profile poboru ciepłej wody użytkowej mają trzy kluczowe elementy: Wysoka rozdzielczość czasowa. Cztery poziomy zużycia - odpowiadały one średniej, medianie, a także procentylo. Trzy wzorce czasowe - koncentowały się one na porannym, wieczornym lub równomiernym zużyciu wody przez cały dzień, aby uwzględnić różne style życia i zwyczaje mieszkańców.

Modele szeregów czasowych są specjalną kategorią modeli statystycznych przeznaczonych do analizowania i prognozowania danych, które są zebrane lub obserwowane w regularnych odstępach czasu. Gelazanskas i Gamage[2] zauważywszy silną dzienną i tygodniową korelację. Zastosowali kombinację metode dekompozycji sezonowej oraz model ARIMA. Po porównaniu zaproponowanego modelu wraz z wybranymi modelami referencyjnymi stwierdzono, że rozdzielenie szeregów czasowych na kilka składowych odgrywa znaczącą rolę w dokładności prognoz - uwzględnienie sezonowych wzorców zużycia może znacząco poprawić jakość przewidywań.

Modele behawioralne, są to modele oparte na zachowaniu ludzi. Biorą one pod uwagę różne czynniki takie jak motywację, preferencje czy wpływ społeczny. Są one zazwyczaj stosowane w dziedzinach naukowych takich jak psychologia czy ekonomia, ale również w inżynierii. Model behawioralny stworzony przez N.D. Pflugrad[3] symulował indywidualne zachowania w celu przewidywania zużycia energii i wody w gospodarstwie domowym. Model został stworzony w celu uniknięcia polegania na statystycznych rozkładach prawdopodobieństwa. Autor stwierdził, że przewidywania modelu są bardzo dobre dla gospodarstw zamieszkałych przez jedną osobę, gdyż największym problemem modelu jest jego skalowalność, ze względu na konieczność dokonania dużej ilości założeń i trudności w gromadzeniu danych o zachowaniach wielu mieszkańców.

1.2.2 Modelowanie zasobnika CWU

W kontekście modelowania zasobnika CWU, różne podejścia zostały przedstawione w literaturze. Jednym z efektywnych narzędzi w tym procesie jest modelowanie z linearyzacją. To podejście opiera się na przybliżeniu nieliniowego modelu do postaci liniowej, wykorzystując rozwinięcie w szereg Taylora wokół pewnego punktu pracy. Głównym celem

tego procesu jest uproszczenie złożonego modelu nieliniowego do postaci bardziej przystępnej - liniowej. Przykładem zastosowania linearyzacji w kontekście modelowania zasobnika CWU można odnaleźć w pracy autorstwa S. Claquina, A. Carriere'a i F. Rocaries'a.[4] W ich badaniach, skonstruowali bardzo precyzyjny model matematyczny, rozkładając go na trzy równania stanu. Te równania opisują zmienną temperatury wody w wymienniku, temperatury ramy wymiennika oraz temperatury komory spalania. Następnie, dzięki zastosowaniu linearyzacji, udało im się opracować kontroler adaptacyjny.

W literaturze można również znaleźć podejście modelowania, które polega na podziale zasobnika na warstwy o różnych temperaturach, umieszczone na różnych poziomach. Jest to podejście bardziej dokładne niż modelowanie zasobnika w modelu jednowarstwowym, gdyż bierze dokładniej pod uwagę takie zjawiska jak mieszanie się wody o różnej temperaturze, utratę ciepła do otoczenia czy stratę ciepła w transporcie. S. Alizadeh[5] prowadził badania nad zachowaniem ciepła w zbiornikach poziomych, które wykazały, że uwzględnienie rozkładu temperatury w zbiorniku ma istotne znaczenie dla oceny wydajności i efektywności. Podkreślał, jak utrata energii wewnętrznej zbiornika wynosiła 1.4%. Badania przeprowadzone przez M. Y. Haller, J. Streicher[6] wykazały, że utrata energii wewnętrznej mieściła się między 2.04% a 6.50%. Wnioski wysunięte przez nich wskazywały m.in. na różnice wynikające z kształtu zasobnika, pojemności oraz systemów wprowadzania i odbierania ciepła

1.3 Charakterystyka rozdziałów

W rozdziale drugim, zatytułowanym "Zastosowane Narzędzia w Pracy", zostanie szczegółowo omówione wymagania dotyczących języka programowania oraz głównej biblioteki używanej do tworzenia sieci neuronowych. Najpierw zostaną rozpatrzone dwa główne języki, które są popularne w dziedzinie uczenia maszynowego i zostanie dokonany wybór, który lepiej odpowiada wymaganiom. Następnie koncentracja zostanie położona na wybór biblioteki do tworzenia sieci neuronowe. Ponownie zostaną przedstawione dwie popularne biblioteki, które zostaną przeanalizowane pod kątem funkcjonalności, wydajności i wsparcia społeczności. Następnie zostanie wybrana jedna, która najbardziej sprzyja realizacji założeń. W dalszej części zostaną przedstawione biblioteki pomocnicze, które są niezbędne do przetwarzania i analizy danych. Zostaną opisane najważniejsze funkcje i zastosowania tych narzędzi. Zostanie także poruszony wątek kompatybilności tych bibliotek.

W kolejnym rozdziale pracy zostanie przedstawiona metodologia projektowania modelu sieci neuronowej. Szczegółowo omówione zostaną hiperparametry oraz zasady działania sieci neuronowych, aby zapewnić solidne podstawy teoretyczne dla dalszej części badań. Następnie zostaną omówione dane wejściowe, które są niezbędne do stworzenia modelu prognozującego dzienne zużycie wody. Szczególna uwaga zostanie poświęcona spo-

sobom przetwarzania tych danych, aby były one adekwatne do zastosowania w sieciach neuronowych. W dalszej części rozdziału zaprezentujemy cztery różne modele sieci neuronowych, które będą się różnić ilością wejść. Dokładnie zostaną omówione wyniki generowane przez każdy z modeli dla trzech różnych zestawów danych. Na zakończenie rozdziału zostanie zaproponowany sposób ulepszenia wybranego modelu sieci neuronowej. Ulepszenie to będzie polegało na wprowadzeniu dodatkowego parametru wykorzystującego dane historyczne.

W ramach rozdziału "Pogoda" niniejszej pracy dyplomowej dokonano prezentacji i analizy danych meteorologicznych, które posłużyły jako fundament dla procesu nauki dwóch modeli sieci neuronowych o zróżnicowanej złożoności architektonicznej. Następnie, w dalszej części rozdziału, skoncentrowano się na szczegółowym badaniu i ocenie ważności poszczególnych wejść dla obu modeli. Do tego celu wykorzystano trzy zaawansowane metody analityczne: Permutacyjną ważność cech, analizę wag wejściowych modelu oraz metodę LIME (Local Interpretable Model-agnostic Explanations). Każda z tych technik dostarcza istotnych informacji o znaczeniu poszczególnych zmiennych wejściowych w kontekście ich wpływu na wydajność i dokładność modeli. Następnie zostanie sprawdzenie poprawności implementacji metod poprzez zastosowanie dodatkowego wejścia do modelu - temperature zadaną.

Kolejnym etapem pracy będzie modelowanie zasobnika CWU. Zostranie przedstawiony wybrany model wraz z uwzględnieniem uproszczeń oraz szczegóły dotyczące sterowania piecem i modelu przewidującego zużycie wody. Następnie zostaną przedstawione dwie funkcje celu - funkcja komfortu i funkcja kosztów. Zostanie omówiona ich charakterystyka, oraz w jaki sposób reagują na zmiany współczynników symulacji.

Rozdział 2

Narzędzia zastosowane w pracy

2.1 Język programowania

Wybór głównego języka programowania zastosowanego w projekcie, wiązał się z postawieniem pewnych wymagań. Pierwszym z tych wymagań była dostępność dedykowanej biblioteki do uczenia maszynowego, która posiada narzędzia do efektywnej pracy nad modelami czy ich testowanie. Użycie biblioteki która jest dobrze utrzymana zapewnia ogromne wsparcie społeczności, które może okazać się nieocenione w procesie nauki czy rozwiązywania problemów.

Kolejnym wymaganiem jest aby wybrana technologia była ciągle wspierana i aktualizowana. Machine learning jest aktualnie jedną z najszybciej rozwijających się dziedzin programowania, co wiąże się z szybkimi zmianami. Po dokładnej analizie postawionych wymagań, zdecydowano o wyborze dwóch głównych języków programowania, które najlepiej odpowiadają potrzebom projektu.

- R
- Python

R jest językiem skoncentrowanym na analizie danych i statystyce. Posiada on bardzo bogaty ekosystem jednak może stanowić to przyczynę wielu konfliktów pomiędzy pakietami. Największą wadą tego języka jest problem ze skalowalnością. Praca z dużą ilością danych skutkuje zużyciem ogromnej ilości pamięci RAM.

Skorzystanie z Pythona będzie lepiej spełniać wymogi projektu. Jest on językiem bardziej wszechstronny oraz posiada obszerną bibliotekę standardową jak i bardzo liczne zewnętrzne biblioteki. Zgodnie z ankietą przeprowadzoną przez Stack Overflow w 2023 roku[7], Python zajmuje trzecie miejsce pod względem popularności wśród technologii oraz drugie miejsce pod względem preferencji na rynku pracy. Największą wadą Pythona jest jego niska wydajność która jest spowodowana faktem, że jest on językiem interpretowanym, więc nie jest on kompilowany do kodu maszynowego przed jego uruchomieniem.

2.2 Biblioteka uczenia maszynowego

Decyzja o wyborze odpowiedniej biblioteki do tworzenia sieci neuronowych jest kluczowym elementem każdego projektu związanego z uczeniem maszynowym i głębokim uczeniem. TensorFlow, będący jedną z dwóch głównych otwartych bibliotek do uczenia maszynowego w Pythonie, wyróżnia się swoją wszechstronnością i wsparciem od Google. Jego głównym konkurentem jest PyTorch, rozwijany przez Facebook, który zyskał popularność ze względu na swoją elastyczność i przyjazność dla użytkownika.

- **Rozbudowana Dokumentacja:** Jest niezbędne, aby biblioteka była wyposażona we wszechstronną i precyzyjnie zorganizowaną dokumentację, która wspiera proces edukacyjny oraz efektywne rozwiązywanie problemów technicznych. W kontekście tego wymogu TensorFlow wyróżnia się, dostarczając szczegółowe instrukcje i obszerną kolekcję przykładowych zastosowań, co stanowi bezcenny zasób zarówno dla osób rozpoczynających pracę z tą technologią, jak i dla programistów o zaawansowanym poziomie doświadczenia.
- **Integracja z Innymi Narzędziami i Bibliotekami:** Python oferuje szeroką gamę pomocniczych bibliotek, które dostarczają zróżnicowanych typów danych, co jest istotnym aspektem jego funkcjonalności, więc ważne aby biblioteka była z nimi kompatybilna. TensorFlow oferuje szerokie możliwości integracji, zwłaszcza z narzędziami Google Cloud[8], podczas gdy PyTorch jest bardziej ukierunkowany na modułowość i elastyczność w integracji z różnorodnymi ekosystemami.
- **Łatwość Użycia:** Intuicyjność i prostota użytkowania są niezbędne dla szybkiego prototypowania i eksperymentowania. PyTorch, ze swoim dynamicznym grafem obliczeń, jest często postrzegany jako bardziej dostępny dla nowych użytkowników oraz bardziej elastyczny w badaniach.
- **Wydajność i Optymalizacja:** Optymalna wydajność i możliwości skalowania są kluczowe w aplikacjach przemysłowych. TensorFlow, z lepszym wsparciem dla TPU[9] i rozbudowanymi opcjami treningu rozproszonego, często przewyższa PyTorch w dużych projektach i zastosowaniach o dużym obciążeniu.
- **Obsługa Wersjonowania Modeli:** Zdolność do zarządzania różnymi wersjami modeli jest istotna w długoterminowej pracy nad projektem. TensorFlow i PyTorch oferują różne podejścia do zarządzania i wersjonowania modeli, które należy rozważyć w kontekście specyficznych wymagań projektu.

Po dokładnej analizie i uwzględnieniu wszystkich kluczowych wymagań zdecydowano, że najlepszym wyborem dla niniejszego projektu jest TensorFlow. Jest on idealnym wyborem zapewniającym wysoką wydajność, ale również elastyczność i skalowalność.

2.3 Narzędzia pomocnicze

Python odgrywa kluczową rolę w dziedzinie uczenia maszynowego i analizy danych, oferując zróżnicowany zestaw narzędzi, które ułatwiają i usprawniają kluczowe procesy, takie jak przetwarzanie danych, ich testowanie oraz wizualizacja. Korzystając z nich, możliwe jest znaczne przyspieszenie procesów związanych z uczeniem maszynowym, od przygotowania danych aż po analizę wyników, co podkreśla ich niezastąpioną wartość w projektach naukowych i przemysłowych.

2.3.1 Pandas i NumPy

NumPy, znany również jako Numerical Python, jest fundamentalną biblioteką do obliczeń naukowych w Pythonie. Oferuje ona wsparcie dla wielowymiarowych, jednorodnych tablic, które w znaczący sposób przewyższają standardowe listy Pythona pod względem wydajności, szczególnie przy obróbce dużych zbiorów danych. NumPy znacząco upraszcza operacje na danych dzięki wektoryzacji, umożliwiając wykonanie operacji takich jak dodawanie, mnożenie czy transpozycja tablic bez konieczności stosowania pętli i iteracji.

Biblioteka Pandas jest jedną z najbardziej popularnych i użytecznych narzędzi w języku Python przeznaczonych do analizy danych. Kluczowym elementem, który Pandas oferuje swoim użytkownikom, jest struktura danych znana jako DataFrame. Jest to dwuwymiarowa, etykietowana tablica umożliwiająca efektywną manipulację i analizę złożonych zbiorów danych. Dodatkowo biblioteka Pandas wyróżnia się również zoptymalizowaną wydajnością. Jest szczególnie skuteczna w obsłudze dużych zestawów danych, co stanowi kluczowe znaczenie w pracach wymagających szybkiego i efektywnego przetwarzania znacznych ilości informacji. Pandas wyposażona jest w bogaty zestaw narzędzi przeznaczonych do przetwarzania i analizowania danych, co znacznie ułatwia pracę z nimi. Istotną cechą tej biblioteki jest również jej zdolność do łatwego wczytywania i zapisywania danych z i do różnych formatów, takich jak CSV czy Excel.

NumPy i Pandas to dwie biblioteki, które idealnie współdziałają w ekosystemie Pythona, szczególnie w kontekście analizy danych i uczenia maszynowego. Pandas, będąc biblioteką wysokiego poziomu do manipulacji i analizy danych, jest zbudowany na fundamencie NumPy, większość funkcji w niej dostępnych, w tym te dotyczące operacji matematycznych i statystycznych, bazuje na mechanizmach i strukturach NumPy. Ponadto, zarówno NumPy, jak i Pandas wykazują doskonałą kompatybilność z bibliotekami używanymi w uczeniu maszynowym, takimi jak TensorFlow.

2.3.2 Matplotlib

Matplotlib jest obecnie uznawana za najpopularniejszą bibliotekę do tworzenia wykresów i wizualizacji danych w języku Python. Ta wszechstronna biblioteka umożliwia tworzenie różnorodnych rodzajów wykresów, w tym wykresów liniowych, słupkowych, kołowych, a także map ciepła, które będą odgrywały kluczową rolę w jednym z badań przedstawionych w tej pracy. Matplotlib wyróżnia się także szerokimi możliwościami personalizacji wykresów, umożliwiając użytkownikom pełną kontrolę nad aspektami wizualnymi takimi jak style linii, kolory, a także możliwość dodawania adnotacji i etykiet, co znacznie zwiększa czytelność i efektywność prezentacji danych. Jedną z najważniejszych cech Matplotlib jest jej integracja z Pandas i NumPy, które są dwiema głównymi bibliotekami do przetwarzania i analizy danych używanymi w tej pracy.

2.3.3 Scikit-learn, ELI5 i LIME

Scikit-learn jest biblioteka zawierająca bogaty zbiór algorytmów uczenia maszynowego, nie tylko oferuje szeroki zakres technik uczenia, ale także zapewnia dostęp do rozmaitych narzędzi przetwarzania danych. Dodatkowo scikit-learn oferuje różnorodne metryki oceny. ELI5 i Lime udostępniają nam narzędzia umożliwiające głębsze zrozumienie mechanizmów decyzyjnych modelu. Dzięki nim jesteśmy w łatwy sposób zbadać każdą predykcję modelu jak i wpływ każdej zmiennej z osobna.

2.3.4 CUDA i Conda

CUDA, będąca narzędziem opracowanym przez firmę NVIDIA, stanowi kluczową technologię umożliwiającą wykorzystanie mocy obliczeniowej procesorów graficznych do przyspieszenia działania aplikacji. CUDA umożliwia znaczne przyspieszenie obliczeń w porównaniu do tradycyjnego przetwarzania na CPU, dzięki tej technologii czas potrzebny do trenowania zaawansowanych modeli uczenia maszynowego zostanie znacznie skrócony.

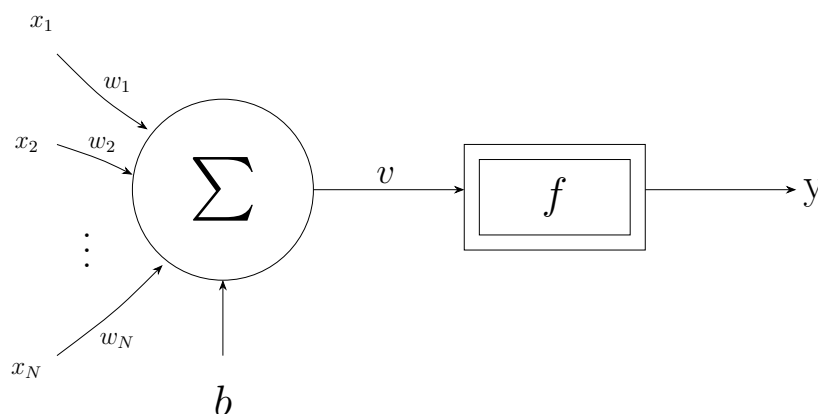
Conda jest środowiskiem wirtualnym i systemem do zarządzania pakietami. Pozwala ona na tworzenie odseparowanych środowisk dla różnych projektów, co jest bardzo ważne w momencie gdy różne projekty potrzebują różnych wersji tej samej biblioteki. Jest to szczególnie istotne w kontekście naszej pracy, gdzie narzędzia opisane wcześniej - mimo że idealnie współpracują ze sobą - wymagają specyficznych, niekoniecznie najnowszych wersji.

Rozdział 3

Modelowanie sieci neuronowej

3.1 Metodologia projektowania modelu sieci neuronowej

W początkowych etapach rozwoju sieci neuronowych, neuron, będący ich najmniejszym elementem był wzorowany na rzeczywistych neuronach ludzkiego układu nerwowego. Każdy taki neuron w sieci neuronowej funkcjonuje jako układ MISO, oznaczający wiele wejść i jedno wyjście. W praktyce oznacza to, że sygnał wyjściowy neuronu jest przekazywany do każdego neuronu w kolejnej warstwie sieci, jednak ponieważ jest to ta sama wartość, traktujemy to jako pojedyncze wyjście.



Rysunek 3.1: Model neuronu

Neuron w sieci neuronowej jest sumatorem. Sumuje on wartości wszystkich wejść, przy czym każde wejście jest mnożone przez określoną wagę oraz dodawana jest do nich wartość stała. Po zsumowaniu tych wartości sygnał wyjściowy neuronu jest przekształcany przez funkcję aktywacji, która decyduje o ostatecznej wartości wyjściowej tego neuronu.[10]

Wyjście pojedynczego neuronu można opisać za pomocą wzoru:

$$y = f \left(\sum_{j=1}^N w_j x_j + b \right), \quad (3.1)$$

gdzie:

- y - sygnał wyjściowy neuronu,
- x_j - sygnał wejściowy neuronu dla j -tego wejścia,
- w_j - waga przypisana do j -tego sygnału wejściowego neuronu,
- b - składnik stały,
- $f(\cdot)$ - funkcja aktywacji neuronu,
- N - liczba wejść neuronu.

Neurony organizują się w sieci, tworząc struktury znane jako sieci neuronowe. Każda sieć neuronowa składa się z wielu neuronów, które są wzajemnie połączone i mogą być zorganizowane w różne warstwy. Charakterystyczną cechą neuronów należących do tej samej warstwy jest to, że otrzymują one wspólny sygnał wejściowy.

Proces tworzenia architektury sieci neuronowej jest procesem wyboru hiperparametrów, czyli wartości konfiguracyjnych, które określają sposób działania i uczenia się modelu. Hiperparametry w sieciach neuronowych można podzielić na dwie kategorie:

1. Hiperparametry Architektury Sieci:

- (a) Liczba warstw.
- (b) Liczba neuronów.
- (c) Typy warstw.

2. Hiperparametry Procesu Uczenia:

- (a) Szybkość uczenia się: Wielkość kroku, o jaki wagi są aktualizowane podczas uczenia.
- (b) Liczba epok: Ilość jaką zestaw uczący jest przetwarzany przez sieć.
- (c) Rozmiar partii: Liczba próbek danych, które są przetwarzane jednocześnie w jednym kroku uczenia.
- (d) Optymalizator: Algorytm służący do aktualizacji wag w sieci.
- (e) Funkcja straty: Metryka oceniająca, jak dobrze model radzi sobie z zadaniem.

Aby sieć neuronowa mogła efektywnie przewidywać wyniki, kluczowe jest jej odpowiednie nauczanie. Proces ten rozpoczyna się od inicjalizacji wag sieci, co stanowi punkt wyjściowy dla dalszego uczenia. Następnie, w trakcie procesu uczenia, wagi te są ciągle modyfikowane. Odbywa się to poprzez przesyłanie danych przez sieć i obliczanie błędu, który jest miarą różnicy między przewidywaniami sieci a rzeczywistymi wynikami. Kolejnym etapem jest obliczenie gradientu funkcji straty dla każdej wagi w sieci. Gradient ten wskazuje kierunek, w którym należy zmodyfikować wagi, aby zminimalizować błąd. Aktualizacja wag odbywa się przy pomocy wybranego algorytmu optymalizacyjnego zwanego optymalizatorem. Cały ten proces jest powtarzany przez ustaloną liczbę iteracji i epok, co pozwala sieci na stopniowe 'naukę' i poprawę swojej zdolności do przewidywania wyników.[11]

3.2 Dane wejściowe i proces ich przetwarzania

Projekt wykorzystuje dane zebrane przez M.J. Ritchie et al.[12] w 2018 roku. Informacje o zużyciu wody zostały pozyskane z 77 domostw znajdujących się na terenie Południowej Afryki, zamieszkiwanym przez dwoje osób dorosłych oraz dwójkę dzieci, oraz posiadających podstawowe urządzenia AGD takie jak pralka czy zmywarka. Dane były zbierane przez 16 tygodni, w cyklach trwających 4 tygodnie, aby równomiernie obejmować każdą porę roku. Próbkowanie danych odbywało się co minute, nieprzerwanie przez całą dobę. Całość danych zostało podzielone na 73 oddzielne pliki w formacie .csv. W każdym pliku znajdował się timestamp w formacie "RR-MM-DD HH:MM:SS", który określał dzień i godzinę pomiaru danych. Oprócz tego zawierał informacje o zużyciu wody, oraz temperaturę otoczenia. Dodatkowo dane były podzielone na poszczególne pory roku.

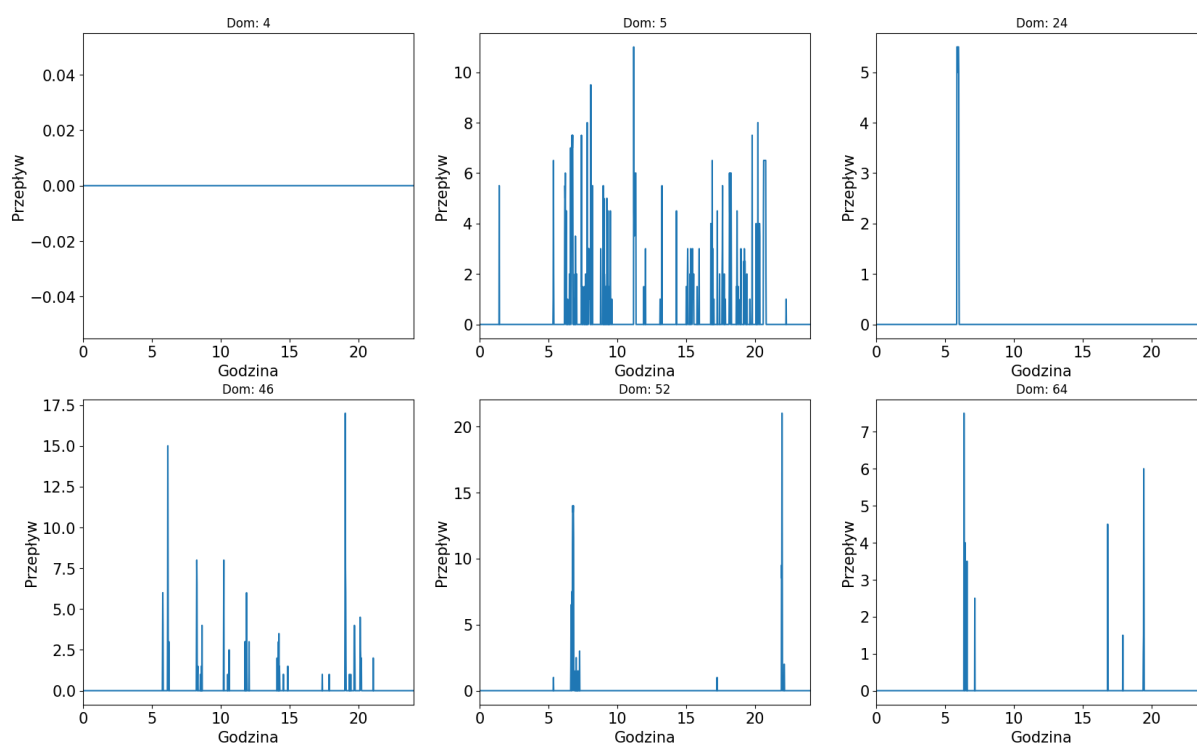


Rysunek 3.2: Porównanie wartości przepływów w różnych ramach czasowych.

Pierwszy wykres przedstawia dane zebrane z domu nr 13 zebrane w dniu 05/02/2018. Oś X oznaczająca godzinę, począwszy od północy. Oś Y reprezentuje przepływ wody w danym momencie dnia. Wykres przedstawia nieregularne piki o nierównomiernym rozkładzie.

Okresami o zwiększonym przepływie są godziny 7-13 oraz 14-15. Okresy mniejszej aktywności możemy zaobserwować w godzinach późno popołudniowych oraz nocnych. Drugi i trzeci wykres przedstawiają zbliżenie na godzinę 7 oraz 15. Dzięki zwięźeniu analizowanego zakresu czasu możliwe było dokładniejsze zbadanie struktury występujących pików. Ta metoda wizualizacji ujawniła, że poszczególne piki, które na ogólnym wykresie dobowym mogły sprawiać wrażenie pojedynczych punktów, w rzeczywistości są złożone z wielu pojedynczych zdarzeń. To odkrycie jest istotne, ponieważ wskazuje na bardziej złożoną dynamikę przepływu w określonych momentach doby, co na pierwszy rzut oka mogło umknąć uwadze

W celu lepszego wstępnego zrozumienia charakterystyki analizowanego zestawu danych, niezbędne jest także szczegółowe przyjrzenie się kilku losowo wybranym domostwom.



Rysunek 3.3: Porównanie przepływów dla przykładowych domów w dniu 05/02/2018

Analiza przedstawionych histogramów przepływów dla poszczególnych losowo wybranych domostw, wykonana na podstawie danych z dnia 05/02/2018, ukazuje wyraźne różnice w charakterystyce przepływów dla każdego z nich. Zgodnie z Rysunkiem 3.2, można stwierdzić, że każdy dom prezentuje unikalny wzór aktywności, co może odzwierciedlać różnorodność nawyków, planów dnia lub specyficznych potrzeb mieszkańców. Na przykład, dla domu nr 5 zużycie wody jest rozłożone przez większą część dnia, z obserwowaną aktywnością w rozmaitych godzinach. Jednakże zarejestrowano również pojedyncze zużycie w nocy. Z kolei Dom 52 charakteryzuje się wyraźnym szczytem wieczornym, co stanowi

kontrast w stosunku do pozostałych domów.

W przypadku Domu nr 4, nie odnotowano żadnego przepływu w analizowanym dniu. Brak danych może wynikać z co najmniej dwóch potencjalnych przyczyn. Pierwszą z nich jest zastosowany czas próbkowania, który wynosił jedną minutę. Taki interwał może nie być wystarczająco krótki, aby zarejestrować sporadyczne lub krótkotrwałe zdarzenia przepływu. Drugą możliwością, która może wyjaśniać brak zarejestrowanej aktywności, jest potencjalna nieobecność mieszkańców w domu w danym dniu.

Pomimo iż zgromadzone dane charakteryzowały się wysoką jakością, ich specyficzny format wymagał przygotowania skryptu celem ich przetwarzania i ekstrakcję istotnych informacji.

```

1 for f in csv_files:
2     dataset = pd.read_csv(f)
3     for i in range(len(dataset)):
4         dt = dataset.loc[i, "Summer_Timestamps"]
5         data, time = dt.split('□')
6         year, month, day = (int(x) for x in data.split('-'))
7         ans = datetime.date(year, month, day)
8         dzien_tygodnia = dni_tygodnia_mapa[ans.strftime("%A")]
9
10        hours, minutes, null = time.split(":")
11        time = (int(hours)*60+int(minutes))/(60*24)
12
13        przeplyw = dataset.loc[i, "Summer_Water_Consumption"]
14
15        dane.loc[len(dane)] = {'Pora_roku': 2,
16                               'Dzien_tygodnia': dzien_tygodnia,
17                               'Czas_dnia': time,
18                               'Przeplyw': przeplyw}

```

Rysunek 3.4: Fragment skryptu przetwarzającego dane.

W ramach procesu dostawania formatu aby przystosować go do wymagań tensorflow dzień tygodnia został zamieniona na etykietę liczbową, która przyjmuje wartość od 1 do 7, co odpowiada kolejnym dniom tygodnia. Podobny proces został zastosowany do etykietowania pór roku. Każda została zakodowana jako etykieta w zakresie od 1 do 4 co prezentuje kolejno wiosnę, lato, jesień i zimę. Dodatkowo czas dnia został zmieniony na procent dnia w skali od 0 do 1.

3.3 Projektowanie i ocena modeli

W ramach realizacji badań nad optymalizacją architektury sieci neuronowej oraz doбором hiperparametrów zdecydowano się na podział danych uczących na trzy zbiory. Pierwszy z nich to zestaw który zawiera dane pochodzące z 12 losowo wybranych domostw, co ma na celu zapewnienie reprezentatywności i różnorodności w ramach próby badawczej. Drugi zestaw stanowi podzbiór zawierający dane z pojedynczego gospodarstwa domowego, co pozwala na szczegółową analizę wydajności modelu w warunkach bardziej jednorodnych danych. Dodatkowo, utworzony został trzeci zestaw danych, który obejmował informacje z wszystkich 77 domów biorących udział w badaniu.

Podział zbioru danych na trzy zestawy okazał się kluczowy dla efektywnego doboru hiperparametrów modelu, szczególnie biorąc pod uwagę, że cały zbiór danych zawierał aż 12,5 miliona wierszy. Zestaw wybranych domostw, zawierający blisko 2 miliony wierszy, oraz pojedyncze domostwo z 161 tysiącami wierszy, umożliwiły przeprowadzenie dokładniejszych i bardziej zróżnicowanych testów. Czas uczenia sieci był znacząco różny dla poszczególnych zestawów danych. Przykładowo, dla całego zbioru danych proces uczenia trwający 10 epok przy rozmiarze partii równym 64 zajmował około 46 minut. Tymczasem dla wybranych domostw czas ten skracał się do 9 minut, a dla pojedynczego domostwa uczenie trwało zaledwie 40 sekund. W ramach badań podjęto próby wykorzystania Google Colab, będącego popularnym narzędziem służącym do programowania i przetwarzania danych w chmurze. Po odpowiednim skonfigurowaniu środowiska napotkano na pierwszy znaczący problem - czas trwania uploadu pliku. Zaskakująco, przesyłanie pliku o rozmiarze 300 MB, zawierającego 12,5 miliona wierszy, zajęło znacznie więcej czasu, niż można było przewidywać. Kolejnym krokiem było przeprowadzenie procesu uczenia maszynowego na danych, zaplanowanego na 10 epok. Niestety, cały proces trwał ponad 100 minut, co wskazuje na ograniczenia wersji darmowej Google Colab. W związku z tym stwierdzono, że bez inwestycji w wersję płatną, Google Colab nie zapewnia oczekiwanej redukcji czasu niezbędnego do nauki modelu.

Początkowo hiperparametry były testowane na najmniejszym zbiorze, co pozwalało na szybką i efektywną ocenę różnych konfiguracji. Po uzyskaniu zadowalających wyników na zbiorze pojedynczego domostwa testy były rozszerzane kolejno na zbiór wybranych domostw, a następnie na pełny zbiór danych. Taka strategia pozwoliła na stopniowe i metodyczne dostosowywanie hiperparametrów, minimalizując przy tym czas i zasoby potrzebne do przeprowadzenia eksperymentów, a jednocześnie maksymalizując ogólną skuteczność modelu. Na podstawie serii przeprowadzonych testów dotyczących dostrajania modelu wybrano następujące hiperparametry:

- Optymalizator: Adam
- Funkcja strat: mse
- Rozmiar partii: 64
- Początkowy współczynnik uczenia: 0.0001

Po przeprowadzeniu serii eksperymentów, w procesie selekcji optymalnej architektury sieci neuronowej, najbardziej efektywną konfiguracją okazała się struktura składająca się z sześciu warstw, z których cztery pełniły funkcję warstw ukrytych. W procesie iteracyjnego dostosowywania i ewaluacji różnych architektur sieci, model o takiej budowie wykazał najlepsze wyniki w zakresie dokładności i generalizacji na testowanych zbiorach danych. Architektura ta charakteryzowała się kolejno malejącą liczbą neuronów w poszczególnych warstwach: pierwsza warstwa zawierała 512 neuronów, druga 256, trzecia 128, czwarta 64, piąta 32, a szósta, będąca warstwą wyjściową, miała 1 neuron. Wszystkie warstwy, z wyjątkiem ostatniej, wykorzystywały funkcję aktywacji ReLU. Natomiast ostatnia warstwa, pełniąca rolę warstwy wyjściowej, zastosowała funkcję aktywacji typu 'linear'.

W ramach opracowanego modelu sieci neuronowej zastosowano dynamicznie zmieniający się współczynnik uczenia, oparty na metodzie wykładniczego spadku, opisanego wzorem:

$$\text{Współczynnik uczenia}(epoka) = \begin{cases} \text{Początkowy współczynnik uczenia} & \text{jeżeli } epoka < 5 \\ \text{Współczynnik uczenia}(epoka - 1) \times e^{-0.1} & \text{jeżeli } epoka \geq 5 \end{cases} \quad (3.2)$$

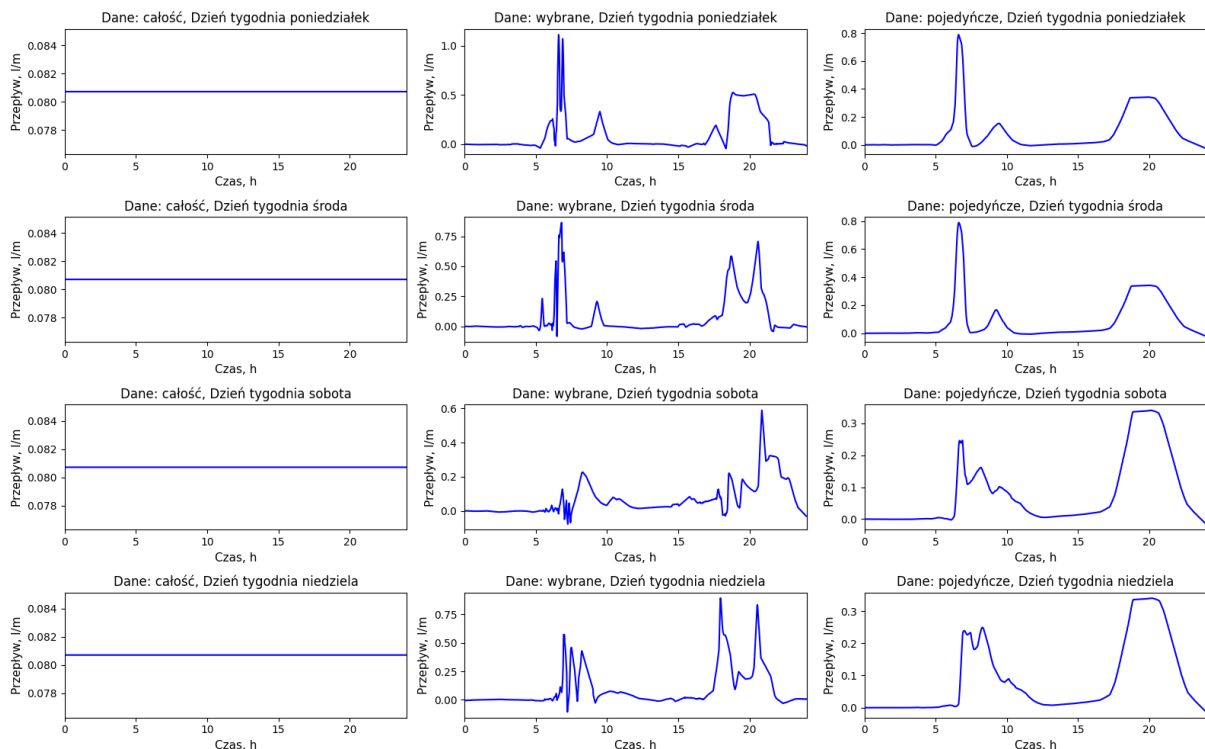
Użycie tej metody pozwoliło na zmniejszanie wartości współczynnika uczenia w trakcie procesu trenowania, co zwiększyło zdolności adaptacyjne sieci. Został on zastosowany, gdyż częstym zjawiskiem było generowanie przez sieć stałej wartości wyjściowej, niezależnie od różnych danych wejściowych.

W ramach procesu testowania różnych konfiguracji sieci neuronowej zaproponowano eksplorację wydajności modeli przy różnorodnych kombinacjach wejść. Wszystkie zaproponowane modele miały jedno wyjście - temperaturę w mieszkaniu. Celem tego podejścia było zbadanie, jak zmiana danych wejściowych wpłynie na zdolność modelu do nauki i generalizacji przewidywania przepływu. Poniżej przedstawiono zestawienie wejść modeli, które zostały uwzględnione w analizie:

1. Wejścia modelu A - dzień tygodnia, pora dnia
2. Wejścia modelu B - pora dnia
3. Wejścia modelu C - pora roku, dzień tygodnia, pora dnia
4. Wejścia modelu D - pora roku, pora dnia

Modele A

Modele typu A zostały opracowane w celu badania korelacji tygodniowej, z uwzględnieniem dwóch kluczowych zmiennych wejściowych: czasu dnia oraz dnia tygodnia. Równocześnie zauważono, że zużycie energii cieplnej w ciągu dni roboczych znacznie różni się od tego w dniach wolnych od pracy, kiedy to ludzie spędzają więcej czasu w swoich domach.



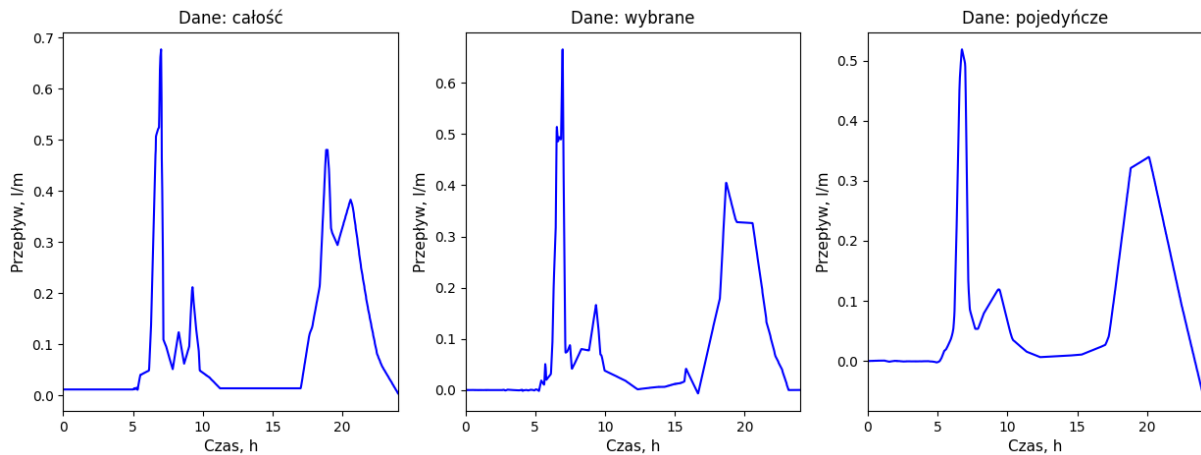
Rysunek 3.5: Porównanie symulacji dla modeli typu A

Predykcje modeli uczących się na całym zbiorze danych cechują się brakiem dynamiki, niezależnie od godziny czy dnia tygodnia. Model utrzymuje stałą wartość w okolicach 0.081 l/min przez cały czas. Pozostałe modele radzą sobie znacznie lepiej. Model uczący się na podstawie danych z 12 domostw symuluje zwiększoną aktywność w godzinach porannych, około 8-9 oraz wieczornych, w okolicach godziny 20. Podobnie zachowuje się

model uczony na pojedynczym domostwie, choć zaproponowane przez niego zużycie wody jest mniej gwałtowne. Ponadto, zauważalne są różnice między symulacjami w dni powszednie a dniami wolnymi. Dni wolne charakteryzują się znacznie większym zużyciem wody w godzinach wieczornych, a także dłuższym, choć mniej intensywnym, korzystaniem z wody w godzinach porannych.

Modele B

Modele typu B zostały zaprojektowane z zastosowaniem jednego wejścia, tj. pory dnia. Celowe zastosowanie tego podejścia wynika z intencji zbadania, w jaki sposób wybrana architektura sieci oraz hiperparametry radzą sobie z zadaniem predykcji, pomijając dodatkowe informacje takie jak dzień tygodnia czy pora roku. Decyzja o ograniczeniu modelu do jednego wejścia wynika z potrzeby zbadania skuteczności predykcyjnej bez wpływu korelacji tygodniowej i sezonowej.

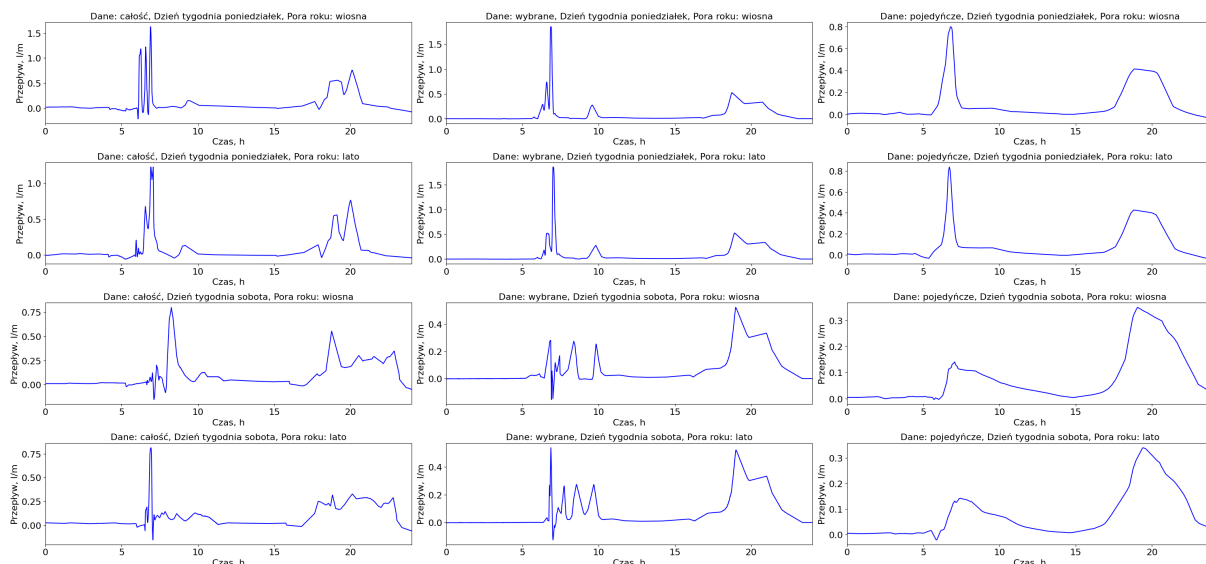


Rysunek 3.6: Porównanie symulacji dla modeli typu B

Niezależnie od ilości dostępnych danych, charakteryzują się bardzo zbliżonymi prognozami dotyczącymi zużycia wody. Zaobserwowano znaczne zwiększenie zużycia wody w godzinach od 6:00 do 9:00 oraz ponownie od 18:00 do 23:00. To idealnie ilustruje, że bez uwzględnienia kontekstu tygodniowego i sezonowego modele generalizują predykcje w bardzo podobny sposób. Zauważalne jest, że prognozy generowane przez model uczący się na pojedynczym domostwie odznaczają się największą płynnością.

Modele C

Modele typu C zostały opracowane w celu oceny skuteczności proponowanej architektury sieci neuronowej oraz dobranych hiperparametrów w kontekście jednoczesnego uwzględniania dwóch rodzajów korelacji: korelacji tygodniowej i sezonowej. W tym celu, do struktury tych modeli wprowadzono trzy kluczowe wejścia: czas dnia, dzień tygodnia oraz porę roku.



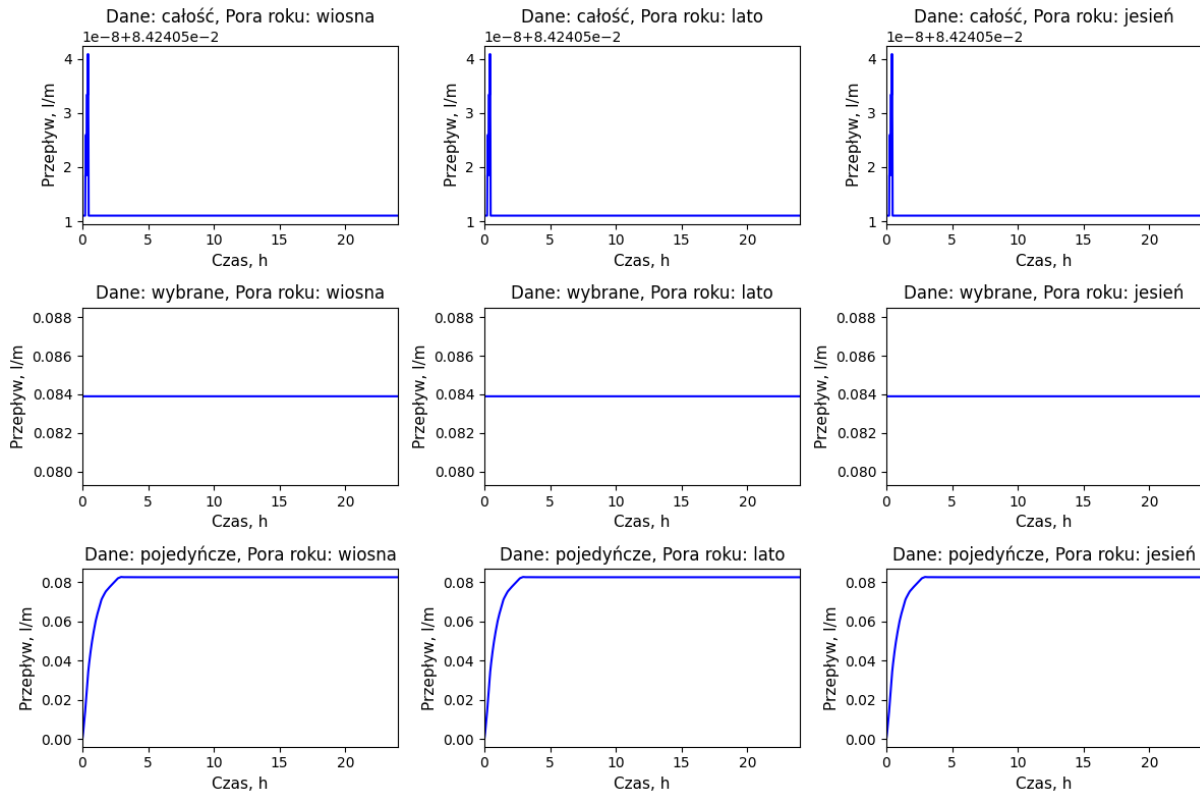
Rysunek 3.7: Porównanie symulacji dla modeli typu C

Analizując rysunek 3.7, można zauważyć, że wszystkie modele prezentują bardzo zbliżoną charakterystykę zużycia wody. Dwa największe szczyty zużycia wody występują w godzinach 6-8 oraz 17-20:30. Wykresy modeli, które zostały nauczone na większej ilości danych, cechują się wyższymi wartościami przepływu, osiągając nawet 1.5 l/min przy uczeniu na pełnym zbiorze danych. Równocześnie przebiegi czasowe przepływu ciepłej wody użytkowej dla tych modeli są bardziej zmiennorodne, charakteryzujące się znacznie większą dynamiką. W przypadku modelu nauczonego tylko na jednym zbiorze danych, dynamika ta jest stosunkowo bardziej płynna.

Należy zauważyć, że przewidywania dla dni roboczych różnią się od tych dla dni weekendowych. Na przykład, dla modelu nauczonego na wybranych domostwach, pojedynczy szczyt na wykresie, obrazujący przepływ w poniedziałek lub środę, został zastąpiony przez trzy pojedyncze szczyty o mniejszej wartości. Dla modelu nauczonego na pojedynczym domostwie pojedynczy szczyt zmienia się w długotrwałe zużycie, zaczynające się o podobnej godzinie, ale trwające nawet do godziny 13.

Modele D

modele typu D zostały skonstruowane w celu badania korelacji sezonowej, wydzielając ją od wpływu korelacji tygodniowej. Modele tego typu zostały wyposażone w dwa kluczowe wejścia: porę roku oraz czas dnia. Zaproponowany model i zestawienie hiper-



Rysunek 3.8: Porównanie symulacji dla modeli typu D

parametrów nie są adekwatne do dokładnej predykcji zużycia ciepłej wody użytkowej. Modele uczące się na pełnym zbiorze danych, niezależnie od pory roku, wykazywały jedynie jedno zdarzenie, symulowane w okolicach godziny 00:30, a następnie prognozowały stałą wartość w granicach 1.1 l/min. Natomiast modele uczące się na podstawie danych związanych z 12 wybranymi domostwami, bez względu na porę roku czy godzinę, prezentowały stałą wartość na poziomie 0.084 l/min. W przypadku modeli uczących się na pojedynczych przypadkach domów, wykazywały one charakterystykę interccji pierwszego rzędu. Rozpoczywały się od zera, stopniowo wzrastając do wartości około 0.08 m/l.

W celu oceny skuteczności zaproponowanych modeli przeprowadzono proces walidacji krzyżowej, z wykorzystaniem metody holdout. Dane zostały podzielone na dwa podzbiory: zbiór uczący i zbiór testowy, stosując proporcję 80-20. Oznacza to, że 80% dostępnych danych zostało przeznaczone do trenowania modelu, podczas gdy pozostałe 20% stanowiło niezależny zbiór testowy, służący do oceny skuteczności modeli w warunkach nieznanych.

Tabela 3.1: Porównanie wartości MSE dla różnych modeli przy zróżnicowanej ilości danych

Liczba domostw	Model A	Model B	Model C	Model D
Całość	0.5841	0.5891	0.5883	0.5842
Wybrane	0.5649	0.5648	0.5650	0.5572
Pojedyncze	0.7276	0.7277	0.7271	0.7271

Analiza wyników wykazuje, że modele uczące się na pojedynczych domostwach wykazują tendencję do generowania najwyższych błędów predykcji, niezależnie od rodzaju modelu. Efekt może być związany z charakterystyką metody holdout, która jest obciążona w sposób pesymistyczny. Dzielenie zbioru danych na dwa mniejsze, może wystąpić pewne zniekształcenie w ocenie skuteczności modelu. Gdyż próbki wybierane losowo, mogą dać niereprezentatywny obraz działania modelu. Z tego powodu rozszerzenie zbioru danych na 12 wybranych domostw lub całość zbioru, przyczynia się do zmniejszenia błędu na zbiorze testowym.

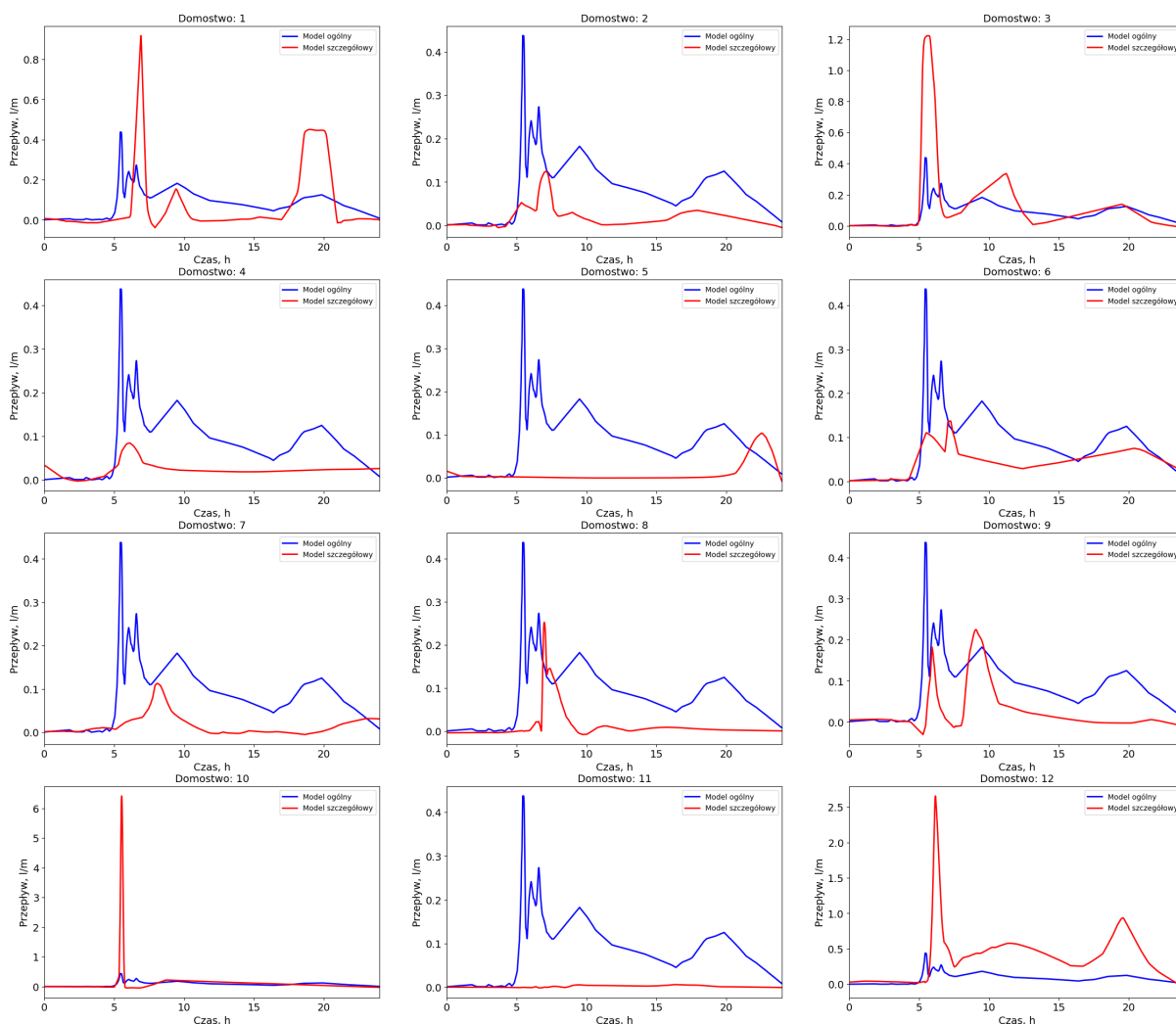
Niezależnie od wybranego typu modelu, obserwuje się, że modele uczące się na wybranych 12 domostwach osiągają niższe wartości błędu średniokwadratowego w porównaniu do modeli uczonych na pełnym zbiorze danych. Jednym z możliwych wyjaśnień tego zjawiska jest bogactwo różnych profili zużycia wody obecnych w całym zbiorze danych.

Pomimo ograniczonej dynamiki predykcji obserwowanej w modelach typu D, ich wartość błędu średniokwadratowego utrzymuje się na poziomie porównywalnym z modelem o bardziej zróżnicowanej charakterystyce. To zjawisko wynika z faktu, że modele typu D symulują wartości przepływu, które są zbliżone do zera (w okolicach 0.08 l/s). W rzeczywistych warunkach pomiarowych, przepływ w badanych domach utrzymuje się na poziomie zerowym przez większość czasu dnia, co wpływa na porównywalność tych dwóch przypadków.

Interesującym aspektem wynikającym z analizy modeli typu A, B i C jest obserwacja, że ich wartości błędu średniokwadratowego są znacznie zbliżone w przypadku uczenia na próbie 12 wybranych domostw. Jednakże, w kontekście modeli uczonych na pełnym zbiorze danych obejmującym 77 domów, ujawnia się zauważalna różnica między modelami A, B i C. Co pokazuje, że dostrzegalny jest istotny wpływ korelacji tygodniowej i sezonowej na różnice w ich osiągnięciach predykcyjnych.

3.4 Walidacja i próby dostrajania (?)

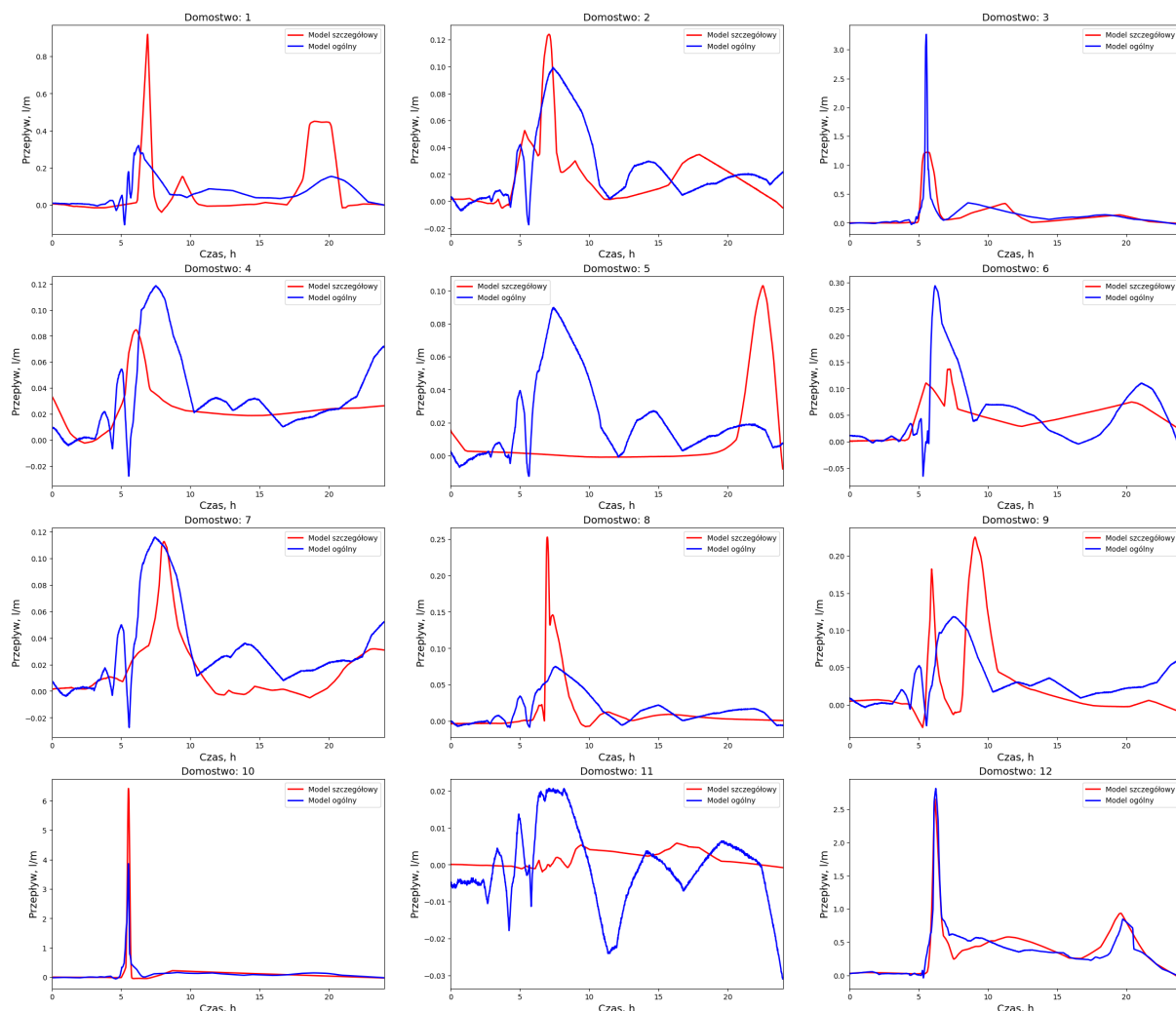
W celu weryfikacji poprawności i efektywności opracowanego modelu sieci neuronowej przeprowadzono porównanie modelu nauczonych na danych ze wszystkich 12 domostw z modelami utworzonymi dla każdego z tych domów osobno. Taki eksperyment miał na celu ocenę zdolności generalizacji modelu nauczonych na zbiorze 12 domostw w porównaniu z modelami specyficznymi dla poszczególnych domów.



Rysunek 3.9: Porównanie modelu ogólnego z modelami szczególnymi

W celu dalszego zwiększenia skuteczności modelu sieci neuronowej zaproponowano wprowadzenie dodatkowego wejścia do systemu – tygodniowego zużycia. Implementacja tego rozwiązania została przeprowadzona w specyficzny sposób, mający na celu uniknięcie przekształcenia tego parametru w niezamierzony label identyfikujący poszczególne domy. W fazie uczenia modelu, do każdego tygodnia przypisywano sumę zużycia zarejestrowanego w tym okresie. Natomiast w fazie testowania, model otrzymywał średnią wartość tygodniowego zużycia. Celem tej strategii było umożliwienie modelowi korzystania z danych historycznych zużycia w sposób, który poprawiałby jego zdolność do przewidywania,

jednocześnie zachowując elastyczność i możliwość generalizacji wyników na różne domostwa.



Rysunek 3.10: Porównanie modelu po dodaniu kolejnego wejścia do sieci

Analizując charakterystykę modelu ogólnego i szczegółowego dla każdego z domostw, przedstawioną na Rysunku 3.5, można zauważyć, że dla domostwa nr 10 i 12 model ogólny wykazał się wysoką zgodnością z modelem szczegółowym, co świadczy o jego zdolności do precyzyjnego odwzorowania charakterystyki przewidywania przepływu w ciągu dnia. W przypadku tych dwóch domów przebieg przewidywań dla obu modeli jest podobny, co wskazuje na to, że model ogólny efektywnie uchwycił dynamikę zużycia charakterystyczną dla tych konkretnych domostw. Natomiast w kontekście domu nr 5, wyniki ukazują, że model ogólny miał znaczne trudności z dopasowaniem się do wzorców przepływu.

W celu dokładniejszej oceny i porównania efektywności modelu ogólnego, nauczonego na danych z 12 domostw, z modelami szczegółowymi, nauczonymi dla poszczególnych domów, zastosowano wskaźnik błędu średniokwadratowego. MSE, obliczany jako średnia kwadratów różnic między wartościami przewidywanymi przez model a rzeczywistymi danymi, posłużył jako miara odchylenia modelu ogólnego od wyników modeli szczegółowych.

W tym kontekście niższa wartość MSE wskazywała na lepszą zgodność modelu ogólnego z wynikami modeli szczegółowych, sugerując, że model ogólny skuteczniej generalizuje dane, zbliżając się do precyzji modeli trenowanych na danych z pojedynczych domostw.

Tabela 3.2: Porównanie wartości MSE dla dwóch modeli w kontekście poszczególnych domostw

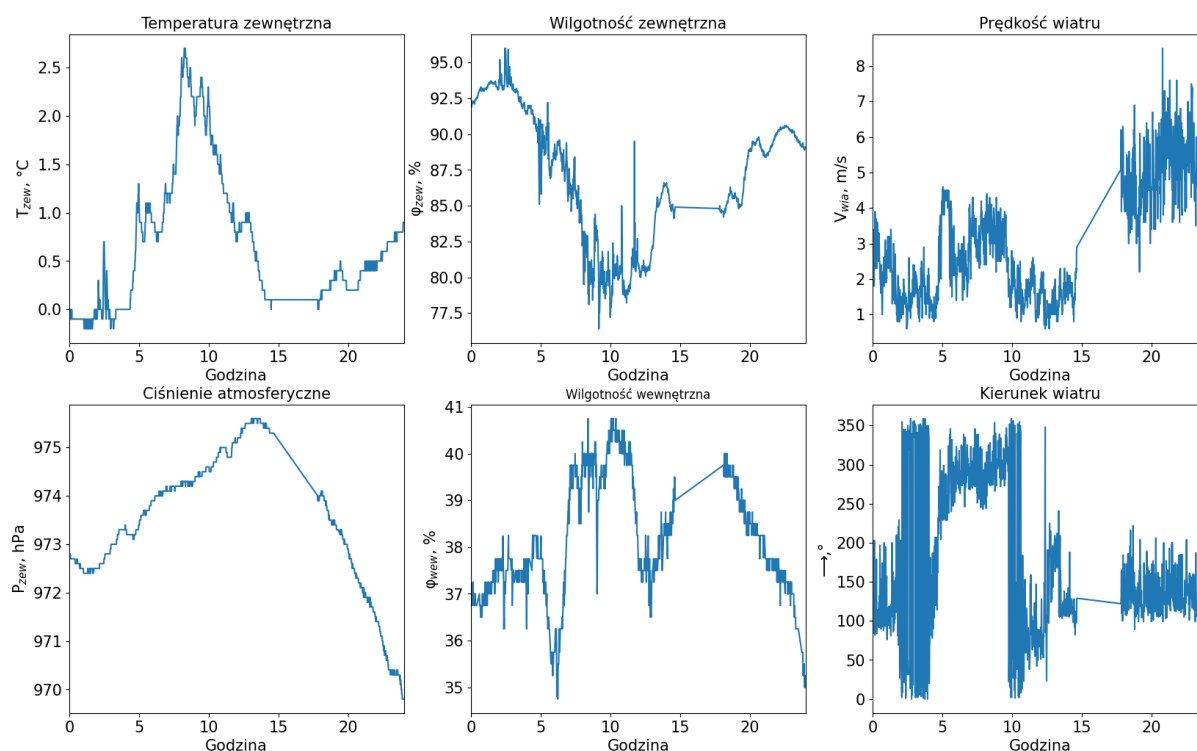
	DOMOSTWO											
	1	2	3	4	5	6	7	8	9	10	11	12
Model podstawowy	0.022	0.00740	0.037	0.00670	0.0110	0.0039	0.00840	0.00950	0.0072	0.220	0.01000	0.190
Model z dodatkowym wejściem	0.019	0.00053	0.034	0.00085	0.0013	0.0025	0.00055	0.00069	0.0027	0.075	0.00013	0.019

Analizując przedstawione dane w tabeli można zauważyć, że model uczący się danymi wraz z dodatkowym wejściem osiąga znacząco lepsze wyniki w porównaniu z modelem bazowy. Wartości błędu średniokwadratowego dla tego modelu są niższe dla każdego z domostw, co wskazuje na poprawę dokładności predykcji. Na przykład, dla domu nr 2, MSE zmniejszyło się z 0.0074 do 0.00053, co jest znaczącą poprawą. Podobne znaczące redukcje można zaobserwować w przypadku domu nr 12, gdzie MSE spadło z 0.19 do 0.019.

Rozdział 4

Wpływ warunków pogodowych

Według Sujith Mathew Iype[13] sztuczna inteligencja potrafi uczyć się ponad 125000 razy szybciej niż człowiek oraz potrafi identyfikować subtelne i zaawansowane wzorce, które mogą pozostać niezauważone przez ludzką percepcję. Dlatego też zdecydowano się przeprowadzić analizę decyzji podjętych przez sieć.



Rysunek 4.1: Porównanie warunków atmosferycznych na przestrzeni dnia

Rysunek przedstawia zestaw sześciu wykresów, które porównują warunki atmosferyczne w ciągu dnia. Każdy wykres zawiera oś X, która reprezentuje czas, oraz oś Y, która przedstawia różne zmienne meteorologiczne. Na pierwszym wykresie prezentowane są zmiany temperatury zewnętrznej na przestrzeni dnia. Temperatura oscyluje w zakresie od około -

0.5°C do 2.5°C, z maksymalnymi wartościami obserwowanymi w okolicach godziny ósmej. Charakterystyka tej zmienności temperatury sugeruje, że dane mogły być zbierane w okresie zimowym. Następny wykres ilustruje zmiany wilgotności powietrza na zewnątrz budynku, wyrażone w procentach. Wysoki poziom wilgotności, utrzymujący się przez cały dzień, może wskazywać na sezon zimowy lub jesienny. Prędkość wiatru, prezentowana na kolejnym wykresie, wykazuje znaczący wzrost w ciągu dnia, osiągając maksymalne wartości przekraczające $8 \frac{m}{s}$, co może być charakterystyczne dla warunków zimowych, kiedy wiatry są silniejsze. Wykres ciśnienia atmosferycznego pokazuje wzrost wartości w pierwszej połowie dnia, po czym następuje stopniowy spadek. Wilgotność wewnętrzna, prezentowana na piątym wykresie, wykazuje wahania z ogólnym trendem spadkowym w drugiej połowie dnia. Szósty wykres przedstawia kierunek wiatru, który jest zmienny w ciągu dnia, z wyraźnymi zmianami szczególnie w środkowej części dnia, co może świadczyć o niestabilnych warunkach atmosferycznych.

W celu zbadania wpływu wielkości modelu na wyniki oraz sprawdzenia ważności różnych wejść, zaproponowano w badaniu dwa modele o różnej złożoności architektury. Oba modele korzystały z tej samej funkcji harmonogramowania tempa uczenia, która redukowała szybkość uczenia po czwartej epoce, oraz z tych samych parametrów kompilacji, w tym optymalizatora Adam z początkową szybkością uczenia 0.001, funkcji straty MSE. Pierwszy model składał się z mniejszej liczby warstw i neuronów: warstwa normalizująca, trzy warstwy gęste z odpowiednio 64, 32 i 1 neuronem, używając funkcji aktywacji 'relu' dla pierwszych dwóch warstw i 'linear' dla warstwy wyjściowej. Drugi model był znacznie większy, zawierając więcej warstw i neuronów: warstwa normalizująca, sześć warstw gęstych o zwiększającej się liczbie neuronów: 1024, 512, 256, 128, 64, 32, zakończonych warstwą wyjściową z 1 neuronem, używając funkcji aktywacji 'relu' dla warstw ukrytych i 'linear' dla warstwy wyjściowej. Oba modele były trenowane przez 50 epok z rozmiarem partii równym 32. Celem porównania tych dwóch modeli było ustalenie, czy zwiększenie liczby warstw i neuronów w modelu wpłynie na jego predykcje.

W ramach procesu weryfikacji skuteczności zastosowanych modeli sieci neuronowych została zastosowana walidacja krzyżowa, metodą holdout. Dane zostały podzielone w proporcji 80% do 20% na zestaw treningowy i testowy. Ta strategia podziału danych miała na celu zapewnienie solidnej bazy do nauki modeli oraz efektywnej oceny ich wydajności. Model numer 1 wykazał się niższą skutecznością w porównaniu do modelu numer 2. Świadczy o tym wartość błędu średniokwadratowego, która dla modelu pierwszego wyniosła 0.0668, natomiast dla modelu drugiego jej wartość była znacząco niższa, a mianowicie 0.0237. Ta różnica w wartościach MSE wskazuje na wyższą precyzję i efektywność modelu numer 2 w procesie uczenia i oceny na podstawie dostępnych danych.

4.1 Permutacyjna ważność cech

Permutacyjna ważność cech, to technika stosowana w uczeniu maszynowym do oceny znaczenia poszczególnych wejść dla modelu predykcyjnego. Metoda ta jest stosowana zarówno dla modeli klasyfikacyjnych jak i regresyjnych. Proces ten rozpoczyna się od trenowania modelu na oryginalnym zestawie danych, co pozwala na ustalenie bazowej wydajności modelu. Następnie przeprowadza się permutację każdej cechy z osobna w zbiorze testowym, losowo mieszając jej wartości, podczas gdy wszystkie inne cechy pozostają niezmienione. Po dokonaniu permutacji model jest ponownie oceniany na zmodyfikowanym zbiorze danych. Wyniki tej oceny są następnie porównywane z wynikami uzyskanymi na oryginalnym, niezmodyfikowanym zbiorze. Różnica w wydajności modelu, taka jak spadek dokładności w klasyfikacji lub wzrost błędu średniokwadratowego w regresji, jest wykorzystywana do oceny ważności danej cechy. Im większy spadek wydajności, tym większa uważana jest ważność tej cechy dla modelu.

W kontekście wykorzystania PFI do oceny ważności cech w modelu istotne jest zwrócenie uwagi na kwestię korelacji między danymi. PFI opiera się na permutacji pojedynczych cech, co oznacza zmianę wartości jednej zmiennej niezależnie od pozostałych. W przypadku, gdy dane są silnie skorelowane, taka metoda permutacji może prowadzić do błędnych wniosków. W celu implementacji techniki PFI w pracy został wykorzystany obiekt `PermutationImportance` z biblioteki `sklearn`. Zastosowanie tego dedykowanego narzędzia umożliwiło nie tylko dokładną, ale i wydajną realizację tej metody, co pozwoliło na jej skuteczną integrację z procesem badawczym.

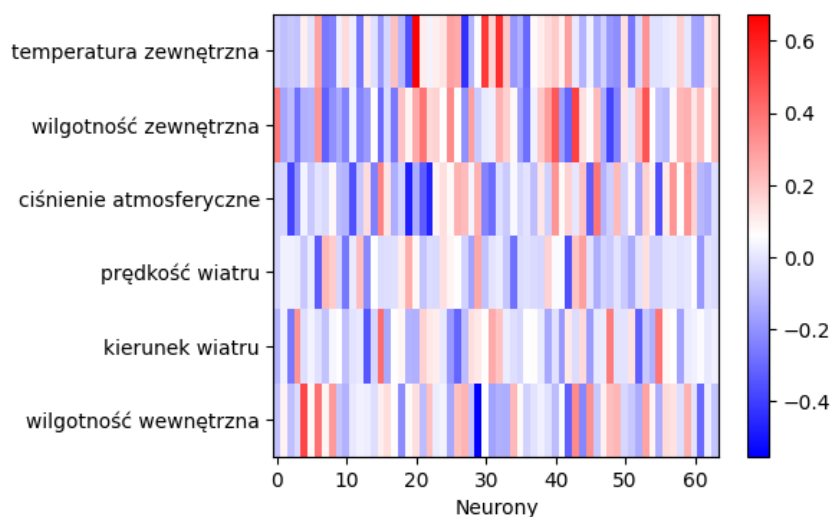
Tabela 4.1: Porównanie ważności wejść dla metody PFI

	Temperatura zewnętrzna	Wilgotność zewnętrzna	Ciśnienie atmosferyczne	Prędkość wiatru	Kierunek wiatru	Wilgotność wewnętrzna
Model uproszczony	0.39	1.00	0.66	0.15	0.01	0.44
Model rozszerzony	0.64	1.00	0.54	0.04	0.19	0.81

Tabela 4.1 porównuje znormalizowane wartości wejść dla dwóch wariantów modelu dla metody PFI. W modelu uproszczonym, najwyższą ważność ma wilgotność zewnętrzna, co sugeruje, że jest to główny predyktor w tym modelu. Temperatura zewnętrzna, wilgotność wewnętrzna i ciśnienie atmosferyczne również odgrywają znaczące role, podczas gdy kierunek i prędkość wiatru ma stosunkowo niższą ważność. W modelu rozszerzonym obserwujemy wzrost ważności Temperatury i wilgotności. prędkość i kierunek wiatru mają sumarycznie większy wpływ w porównaniu do modelu krótkiego, lecz ich ważności jest odwrotna, w modelu rozszerzonym ważniejszy okazał się kierunek, a w uproszczonym prędkość. Ciśnienie atmosferyczne doznaje znaczącego obniżenia ważności, zostając cechą o trzeciej najmniejszej ważności.

4.2 Badanie wag wejściowych pierwszej warstwy

W analizie modeli uczenia maszynowego, interpretacja wag pierwszej warstwy może służyć jako metoda określania ważności cech wejściowych. Ponieważ wagi w pierwszej warstwie sieci neuronowej są bezpośrednio połączone z cechami wejściowymi, wartości tych wag mogą dostarczać informacji o znaczeniu poszczególnych cech dla predykcji modelu. Wysoka wartość wagowa sugeruje, że zmiana wartości tej cechy wejściowej może mieć istotny wpływ na wynik modelu. Jednakże, ta metoda interpretacji może być mniej efektywna w przypadku bardziej złożonych, głębokich sieci neuronowych. W takich modelach, liczne warstwy i skomplikowane struktury, w tym nieliniowe aktywacje i interakcje między neuronami, mogą skutkować tym, że bezpośredni wpływ pojedynczych wag jest trudniejszy do zrozumienia.

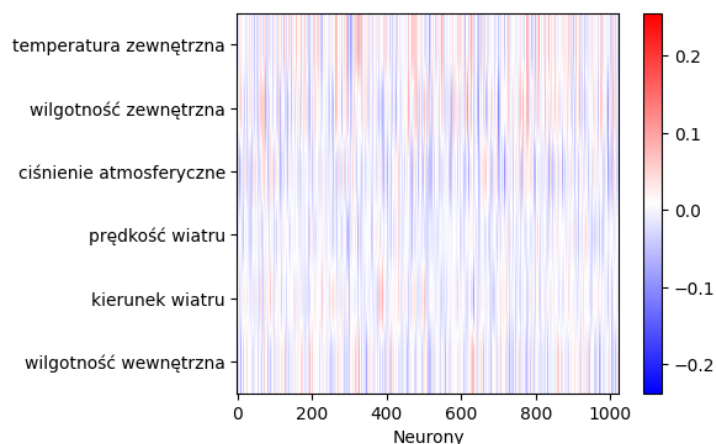


Rysunek 4.2: Mapa ciepła wag pierwszej warstwy dla modelu o uproszczonej architekturze

Mapa ciepła wag pierwszej warstwy sieci neuronowej stanowi cenne narzędzie analityczne, pozwalające na wizualną interpretację i zrozumienie wpływu dużej liczby cech na proces uczenia. Jest to instrument szczególnie użyteczny w kontekście wysokowymiarowych zbiorów danych, gdzie tradycyjne metody analizy mogą okazać się niewystarczające. W przedstawionym przypadku, analiza mapy ciepła nie ujawnia istotnych anomalii w rozkładzie wag. Obserwuje się jedynie sporadycznie wartości, które odstają od średnich wag, lecz nie osiągają one poziomu znacząco wpływającego na wyniki modelu. Można zauważyć, że kierunek i prędkość wiatru są reprezentowane przez najmniej intensywne kolory. To sugeruje, że te dwie zmienne mają mniejszy wpływ na model w porównaniu do innych badanych czynników.

Dla analizy bardziej złożonych modeli sieci neuronowych, metoda interpretacji wag pierwszej warstwy może okazać się nieefektywna. Ze względu na zwiększoną głębokość i złożoność architektury, wagi w pierwszej warstwie tracą bezpośrednią i jednoznaczną

interpretowalność, która jest charakterystyczna dla prostszych modeli. W modelach rozbudowanych, cechy wejściowe przechodzą przez wiele warstw transformacji, co skutkuje utratą bezpośredniego powiązania między wagami pierwszej warstwy a wynikowymi decyzjami modelu. W efekcie, interpretacja tych wag może nie odzwierciedlać faktycznego wpływu poszczególnych cech na decyzje modelu, co jest spowodowane nakładaniem się, transformacją i połączeniem informacji w kolejnych warstwach sieci.



Rysunek 4.3: Mapa ciepła wag pierwszej warstwy dla modelu o rozszerzonej architekturze

Na przedstawionej mapie ciepła, ilustrującej wagi przypisane poszczególnym neuronom w warstwie wejściowej sieci neuronowej o rozmiarze 1024 jednostek, zaobserwować można istotne wzorce, choć sama mapa staje się trudna do interpretacji ze względu na dużą gęstość neuronów. Mimo tej wizualnej złożoności, zauważyć można, że temperatura zewnętrzna jest reprezentowana przez znacząco większą liczbę wag o dodatnich wartościach. W przeciwieństwie, ciśnienie atmosferyczne wykazuje tendencję do posiadania większej ilości wag o wartościach ujemnych. Wagi odpowiadające kierunkowi i prędkości wiatru wydają się charakteryzować przewagą wartości bliskich zeru, co sugeruje mniejszy wpływ tej cechy na model.

W celu dokładnej analizy ważności poszczególnych wag wejściowych sieci neuronowej przeprowadzono obliczenie średniej wartości bezwzględnej wag dla każdego z wejść. Procedura ta umożliwiła identyfikację względnej ważności cech poprzez porównanie ich przeciętnego wpływu na aktywację neuronów w modelu. Następnie, aby umożliwić porównywalność wyników niezależnie od ich pierwotnej skali, dokonano normalizacji obliczonych średnich wag.

Tabela 4.2: Porównanie ważności wejść dla metody PFI

	Temperatura zewnętrzna	Wilgotność zewnętrzna	Ciśnienie atmosferyczne	Prędkość wiatru	Kierunek wiatru	Wilgotność wewnętrzna
Model uproszczony	0.82	1.00	0.78	0.47	0.50	0.69
Model rozszerzony	0.83	1.00	0.83	0.48	0.58	0.73

Analiza przeprowadzona z wykorzystaniem mapy cieplnej okazała się w tym przypadku być wyjątkowo efektywnym narzędziem, pozwalającym na wizualne wyodrębnienie cech o dużej wadze dla modelu. Obserwacje dotyczące temperatury zewnętrznej i ciśnienia atmosferycznego pokrywają się z wynikami otrzymanymi z analizy wag wejść do pierwszej warstwy. Obserwacje dotyczące prędkości i kierunku wiatru na mapie cieplnej wykazały zgodność z obliczonymi wartościami. Obie te dwie zmienne okazały się najmniej znaczącymi wejściami do sieci neuronowej.

Zarówno model uproszczony, jak i model rozszerzony wykazały bardzo podobne predykcje. Jedyną zauważalną różnicą między nimi jest to, że w modelu rozszerzonym temperatura zewnętrzna uzyskała wartość mniejszą niż ciśnienie atmosferyczne, co jest zgodne z obserwacjami w modelu uproszczonym. W obu modelach wilgotność zewnętrzna okazała się najważniejszym wejściem. Wilgotność wewnętrzna, mimo że nie jest najważniejszym wejściem, znalazła się na czwartym miejscu pod względem ważności.

Analiza ważności wejść pierwszej warstwy sieci neuronowej, przeprowadzona za pomocą metody badania wag wejściowych, ujawnia różnice w porównaniu z wynikami uzyskanymi metodą Permutation Feature Importance. Zauważalną różnicą między PFI a badaniem wag pierwszej warstwy jest to, że dla metody PFI wyniki dla obu modeli różniły się bardziej. W przypadku metody PFI największa różnica wynosiła 0.37 dla wilgotności wewnętrznej, podczas gdy dla metody badania wag wejściowych ta różnica wynosiła tylko 0.16. Mimo tych różnic, wyniki dla obu metod są bardzo podobne. Warto jednak zauważyć, że dla modelu uproszczonego średnia ważność jest znacząco mniejsza niż dla modeli z metody badania wag wejściowych.

4.3 LIME

W kontekście zrozumienia modeli uczenia maszynowego, metoda Local Interpretable Model-agnostic Explanations jest istotnym narzędziem, które umożliwia interpretację decyzji modelu na poziomie lokalnym. Metoda ta wyróżnia się spośród innych, takich jak PFI, dzięki swojemu unikalnemu podejściu skoncentrowanemu na pojedynczych instancjach danych. Podczas gdy PFI dąży do zapewnienia ogólnego zrozumienia wpływu cech na model przez zbieranie i analizowanie informacji z różnych instancji, LIME skupia się na wyjaśnieniu, w jaki sposób model dokonuje przewidywań dla konkretnej, wybranej próbki danych.

Podstawowym elementem metody LIME jest modyfikacja danych i tworzenie na ich podstawie uproszczonego modelu, który ma za zadanie odwzorować zachowanie oryginalnego, skomplikowanego modelu, ale tylko w ograniczonym, lokalnym obszarze wokół analizowanej instancji. Proces ten rozpoczyna się od stworzenia zbioru danych przez modyfikację wybranej instancji, co skutkuje powstaniem podobnych, ale nieidentycznych przykładów. Następnie, na tych zmodyfikowanych danych, oblicza się przewidywania za pomocą oryginalnego modelu. Kluczowym krokiem jest trenowanie prostego modelu, takiego jak regresja liniowa, na podstawie tych przewidywań. Model ten służy do aproksymacji wpływu zmian w danych na przewidywania modelu. W ten sposób, analizując współczynniki modelu liniowego, można zrozumieć, które cechy miały największy wpływ na przewidywania dla danej instancji.

W ramach pracy, w celu dokładnej oceny skuteczności metody LIME, zdecydowano się na zastosowanie jej na losowo wybranych próbkach ze zbioru testowego.

Tabela 4.3: Porównanie ważności wejść dla metody LIME

	Temperatura zewnętrzna	Wilgotność zewnętrzna	Ciśnienie atmosferyczne	Prędkość wiatru	Kierunek wiatru	Wilgotność wewnętrzna
Model uproszczony	1.00	0.08	0.68	0.39	0.04	0.81
	1.00	0.37	0.58	0.28	0.22	0.15
	0.82	0.14	1.00	0.34	0.05	0.62
Model rozszerzony	1.00	0.79	0.41	0.20	0.25	0.08
	1.00	0.16	0.14	0.11	0.09	0.76
	0.60	1.00	0.20	0.30	0.70	0.50

Tabela przedstawia wyniki sześciu losowo wybranych prób oceny ważności wejść, z zastosowaniem zarówno modelu uproszczonego, jak i rozszerzonego. W celu ułatwienia porównania, wyniki te zostały znormalizowane i zaokrąglone do dwóch miejsc po przecinku. Analiza danych wykazała, że w dwóch na trzy przypadki, zarówno w modelu uproszczonym, jak i rozszerzonym, najważniejszym wejściem okazała się być tempera-

tura zewnętrzna. Ponadto, w pięciu z sześciu analizowanych przypadków, kierunek wiatru został oceniony jako najmniej znaczący. Szczególnie interesująca jest obserwacja z pierwszej próby, gdzie temperatura zewnętrzna otrzymała znacznie niższą wagę, podczas gdy ciśnienie atmosferyczne, które zazwyczaj charakteryzowało się niższymi wagami, w tej próbie uzyskało najwyższą. Z kolei prędkość wiatru, niezależnie od zastosowanego modelu, konsekwentnie otrzymywała niską wagę w każdej z prób.

Mimo, iż LIME jest metodą zasadniczo skoncentrowaną na dostarczaniu interpretacji lokalnych, istnieje możliwość adaptacji jej do generowania wniosków o charakterze bardziej ogólnym. Można to osiągnąć poprzez zastosowanie metody LIME wielokrotnie na różnych próbkach danych, a następnie wyciągnięcie średniej. Zaproponowano wybranie czterech zestawów ze zbioru testowego, zawierających odpowiednio 10, 50, 100, i 1000 próbek.

Tabela 4.4: Porównanie uśrednionych ważności wejść dla metody LIME dla wielu prób

	Ilość próbek	Temperatura zewnętrzna	Wilgotność zewnętrzna	Ciśnienie atmosferyczne	Prędkość wiatru	Kierunek wiatru	Wilgotność wewnętrzna
Model uproszczony	10	0.70	1.00	0.12	0.08	0.42	0.39
	50	1.00	0.24	0.27	0.00	0.08	0.09
	100	1.00	0.43	0.46	0.09	0.09	0.07
	1000	1.00	0.37	0.44	0.12	0.13	0.07
Model rozszerzony	10	1.00	0.53	0.52	0.15	0.19	0.00
	50	1.00	0.39	0.01	0.10	0.06	0.00
	100	1.00	0.25	0.26	0.19	0.12	0.03
	1000	1.00	0.61	0.26	0.19	0.08	0.00

W analizie wyników zauważono istotne odstępstwa dla uproszczonego modelu, w którym liczba próbek wynosiła 10. Wskazuje to na istotną rozbieżność w porównaniu z rezultatami uzyskanymi dla innych wielkości próbek. W związku z powyższym, dane z tego konkretnego przypadku zostały wyłączone z procesu interpretacji wyników. Takie odstępstwa mogą być rezultatem losowego pobierania próbek oraz tego, że liczba 10 jest relatywnie niewielka, co może prowadzić do niereprezentatywności danych.

Zarówno w modelu uproszczonym, jak i modelu rozszerzonego, temperatura zewnętrzna jest systematycznie uznawana za cechę o największym wpływie na przewidywania modelu, co wskazuje na jej przewodnią rolę w analizowanych danych meteorologicznych. Niezmiennie wysokie wartości ważności tej zmiennej świadczą o jej dominującej pozycji w obu modelach, niezależnie od wielkości próby. Dla obu modeli, wilgotność jak i ciśnienie wykazują stabilne wartości, tylko w pojedynczych próbach ciśnienie staje się dominującą zmienną. Kierunek jak i prędkość wiatru dla obu modeli pozostaje cechą o niskiej ważności. Wilgotność wewnętrzna konsekwentnie otrzymuje jedne z najniższych wartości

ważności.

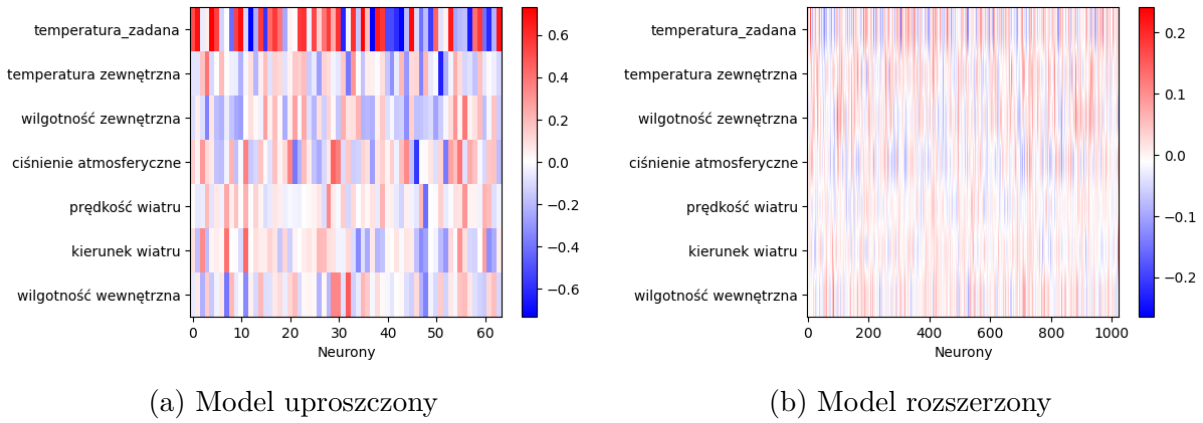
4.4 Ocena skuteczności metod

W celu zweryfikowania poprawności implementacji metod sprawdzania ważności wejść do sieci neuronowej, zaproponowano model, który został rozszerzony o dodatkowe wejście – temperaturę zadaną. Temperatura zadana stanowiła oczekiwaną wartość, jaką piec starał się utrzymać, co sprawia, że każda z metod powinna przypisywać jej kluczowe znaczenie. Ustawiona była na poziomie 21 stopni celsjusza w godzinach 6-23 oraz 19 stopni celsjusza w godzinach 23-6. Piec był sterowany za pomocą regulacji dwupołożeniowej, której odbiornik był umiejscowiony w pomieszczeniu, w którym brakowało jakiegokolwiek źródła ciepła, takiego jak grzejnik.

Tabela 4.5: Wartości MSE modeli uproszczonych i rozszerzonych z uwzględnieniem dodatkowego parametru

	Bez dodatkowego parametru	Z dodatkowym parametrem
Model uproszczony	0.0706	0.0691
Model rozszerzony	0.0262	0.0230

Analizując tabelę 4.5 zauważono, że zarówno model uproszczony, jak i model rozszerzony z dodatkowym parametrem - temperaturą oczekiwaną, osiągały niższe wartości błędu średniokwadratowego w porównaniu do modeli bez tego dodatkowego wejścia. Ten wynik sugeruje, że uwzględnienie temperatury oczekiwanej w strukturze modelu przyczyniło się do poprawy dokładności prognozowania na zbiorze testowym.



Rysunek 4.4: Mapy cieplne modeli z dodatkowym wejściem

Analizując mapy cieplne 4.4a i 4.4b, bez trudu zauważono, że temperatura zadana w obu przypadkach osiąga najbardziej wyraźne kolory praktycznie przy każdym neuro- nie wejściowym. Warto również zaznaczyć, że zarówno prędkość, jak i kierunek wiatru osiągają wartości najbliższe zera. Natomiast pozostałe wejścia, takie jak temperatura ze- wnętrzna, wilgotność zewnętrzna i wewnętrzna, wykazują zróżnicowane wartości, jednak nigdy nie osiągają one tak intensywnych kolorów jak temperatura zadana, co pokrywa się z założeniami jak i poprzednimi wynikami.

Tabela 4.6: Porównanie ważności wejść dla wszystkich metod

	Metoda	Temperatura zadana	Temperatura zewnętrzna	Wilgotność zewnętrzna	Ciśnienie atmosferyczne	Prędkość wiatru	Kierunek wiatru	Wilgotność wewnętrzna
Model uproszczony	PFI	1.00	0.50	0.00	0.48	0.86	0.87	0.46
	Wagi wejściowe	1.00	0.34	0.35	0.41	0.22	0.31	0.31
	LIME _{sr(100)}	1.00	0.11	0.40	0.34	0.09	0.05	0.01
Model rozszerzony	PFI	1.00	0.62	0.00	0.13	0.58	0.73	0.11
	Wagi wejściowe	1.00	0.54	0.61	0.59	0.33	0.34	0.49
	LIME _{sr(100)}	1.00	0.17	0.11	0.07	0.02	0.00	0.01

Analizując tabelę 4.6 stwierdzono skuteczność implementacji metod dla obu modeli. Dla każdej z metod temperatura zadana uzyskała najwyższą wartość. Po wprowadzeniu dodatkowego parametru zauważono, że metoda LIME, zastosowana na 100 próbkach, przypisała podobne wagi cech dla obu modeli. Zaobserwowano jednakże, obserwowano większe zmiany w przypadku wilgotności zewnętrznej i ciśnienia atmosferycznego, gdzie te cechy uzyskały znacznie większą ważność w modelu o bardziej złożonej architekturze. Metoda wag wejściowych utrzymała identyczną hierarchię ważności, z tą różnicą, że dla modelu rozszerzonego wszystkie wagi były większe. Metoda PFI, podobnie jak metoda LIME, wykazała bardzo podobne wyniki dla obu modeli, z jedyną różnicą w przypadku ciśnienia atmosferycznego i wilgotności wewnętrznej, gdzie modele rozszerzone uzyskały niższe wartości.

Analizując średnie wartości ważności dla poszczególnych wejść i pomijając temperaturę zadną, można zauważyć, że kluczowymi czynnikami są temperatura zewnętrzna i kierunek wiatru, obie osiągające średnią ważność na poziomie 0.38. Następnie prędkość wiatru zajmuje ważne miejsce, uzyskując średnią ważność wynoszącą 0.35. Ciśnienie atmosferyczne prezentuje się również jako istotny parametr, uzyskując średnią ważność na poziomie 0.34. Na końcu, wilgotność zewnętrzna i wilgotność wewnętrzna mają średnie ważności wynoszące odpowiednio 0.25 i 0.2. Warto jednak podkreślić, że obliczenia średnich wartości ważności mogą nie dostarczyć pełnego obrazu sytuacji., ponieważ metoda PFI wykazała znacznie różniące się wartości ważności dla prędkości i kierunku wiatru, wahające się od 0.58 do 0.87, natomiast metoda LIME przypisała prędkości i kierunku wiatru wartości od 0.00 do 0.09. Różnice te sugerują istotne rozbieżności w interpretacji ważności poszczególnych parametrów między zastosowanymi metodami analizy.

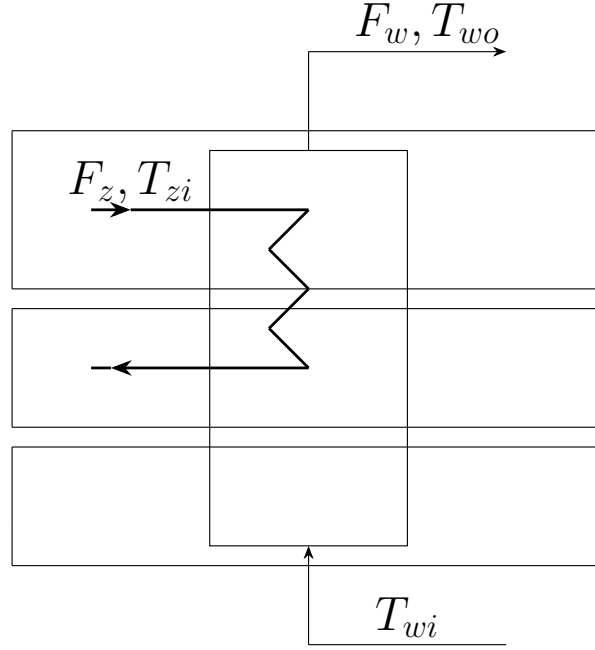
Rozdział 5

Modelowanie dynamiki temperatury wody użytkowej

Do celów modelowania zdecydowano się na wykorzystanie trójwarstwowego modelu CWU. W ramach tego modelu przyjęto założenie idealnego mieszania w zakresie warstwy, co oznacza, że temperatura w całej warstwie jest jednolita. Dodatkowo, zakładano, że każda z warstw traci ciepło do otoczenia oraz wymienia je z sąsiednimi warstwami. Model trójwarstwowo wyposażono w połączenia do systemu grzewczego w drugiej i trzeciej warstwie. W celu uproszczenia modelu pominięto wpływ objętości grzałki na proces wymiany ciepła oraz pominięto wpływ obudowy zbiornika na wymianę ciepła z otoczeniem. Ponadto, zastosowano czujnik temperatury, który mierzy temperaturę w drugiej warstwie. Informacja ta jest przekazywana do regulatora, a sygnał z regulatora steruje systemem grzewczym.

Postanowiono wykorzystać nauczony model sieci neuronowej w praktyce, używając jego prognozy jako do modelu zasobnika. W celu uwzględnienia korelacji tygodniowej oraz sezonowej przepływu wody, zdecydowano się na wybór modelu typu C wraz z dodatkowym parametrem. Symulacje przepływu wody zostały przeprowadzone dla letniego wtorku. W ramach udoskonalenia modelu, podjęto decyzję o modyfikacji prognoz dotyczących zużycia wody. Wszystkie wartości ujemne zostały przekształcone na zero, ze względu na ich nierealność w kontekście pomiaru zużycia wody.

Do zasobnika został zaimplementowany sterownik proporcjonalno różniczkujący. Został on zastosowany ze względu na szybkość reagowania na zmiany wejścia oraz eliminację uchybu w stanie ustalonym. Regulator został nastrojony wykorzystując metodę Zieglera-Nicholsa. Polega ona na doprowadzeniu układu do granicy stabilności wykorzystując tylko regulator proporcjonalny, następnie na podstawie wzmocnienia krytycznego oraz okresu oscylacji wyznacza się wartości parametrów regulatora. Na podstawie tej metody wybrano wzmocnienie części proporcjonalnej równe 7.8 oraz 40.25 dla części różniczkującej



Rysunek 5.1: Trójwarstwowy model zasobnika

$$\frac{dT_{wo}^3}{dt} = b_1^3 F_z (T_{zi} - T_{wo}^3) - b_2^3 F_w (T_{wo}^3 - T_{wo}^2) - b_3^4 (T_{wo}^3 - T_{ot}) \quad (5.1)$$

$$\frac{dT_{wo}^2}{dt} = b_1^2 F_z (T_{zi} - T_{wo}^2) - b_2^2 F_w (T_{wo}^2 - T_{wo}^1) - b_3^2 (T_{wo}^2 - T_{ot}) - b_4^2 (T_{wo}^2 - T_{wo}^1) + b_5^2 (T_{wo}^3 - T_{wo}^2) \quad (5.2)$$

$$\frac{dT_{wo}^1}{dt} = -b_2^1 F_w (T_{wo}^1 - T_{wi}) - b_3^1 (T_{wo}^1 - T_{ot}) + b_5^1 (T_{wo}^2 - T_{wo}^1) \quad (5.3)$$

$$\frac{dT_{zi}}{dt} = p_1 Q_g - p_2 F_z (T_{zi} - T_{wo}^3) - p_3 (T_{zi} - T_{ot}) \quad (5.4)$$

- Q_p - Ciepło systemu grzewczego
- T_{ot} - Temperatura otoczenia
- T_{wo}^n - Temperatura wyjściowa
- T_{wi} - Temperatura wejściowa
- T_{zi} - Temperatura systemu grzewczego
- F_w - Przepływ wody przez zasobnik
- F_z - Przepływ wody przez system grzewczy
- b^n - współczynnik wymiany ciepła
- n - numer warstwy

5.1 Wyniki symulacji

Modelowanie zasobnika CWU w modelu trójwarstwowym zostało przeprowadzone przy użyciu metody Eulera, która jest jedną z najprostszych metod numerycznych służących do rozwiązywania układów równań różniczkowych. Metoda ta polega na iteracyjnym przybliżaniu rozwiązania, bazując na przybliżeniu wartości pochodnej funkcji w każdym kroku. Wzór opisujący metodę Eulera jest następujący:

$$y_{n+1} = y_n + \Delta t \cdot f(t_n, y_n), \quad (5.5)$$

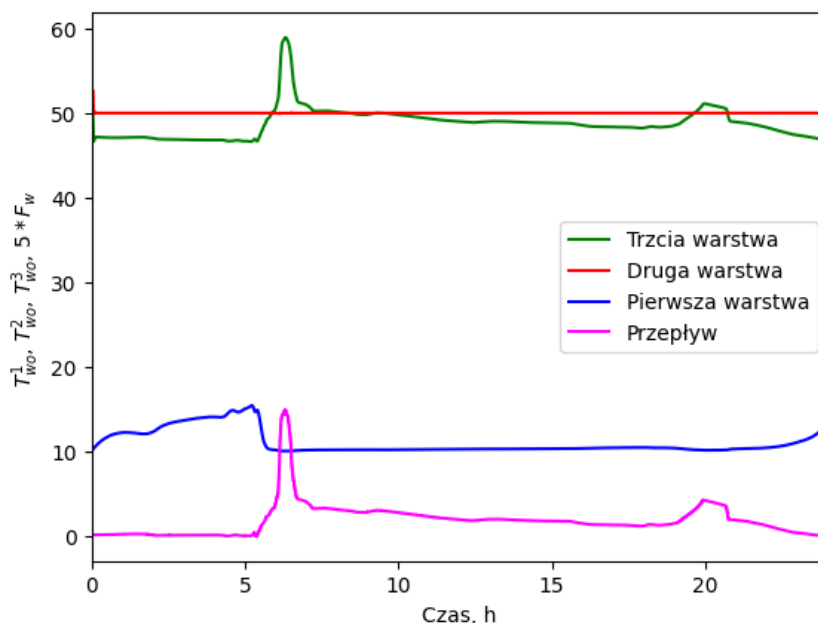
gdzie:

- y_n - wartość przybliżona rozwiązania w chwili t_n
- Δt - krok czasowy
- $f(t_n, y_n)$ - funkcja określająca tempo zmiany y w punkcie (t_n, y_n)

W symulacji użyto poniższych parametrów:

Tabela 5.1: Współczynniki użyte podczas modelowania

Warstwa/ parametr	n=1	n=2	n=3
b_1	0	0.2	0.025
b_2	0.71	0.065	0.067
b_3	0.00006	0.0019	0.0054
b_4	0.0001	0.00079	0.0053
b_5	0.00001	0.0001	0
p_1	0.17		
p_2	0.014		
p_3	0.0048		
T_{wi}	10		
T_{ot}	25.5		
F_z	0.5		
Q_p	8000		

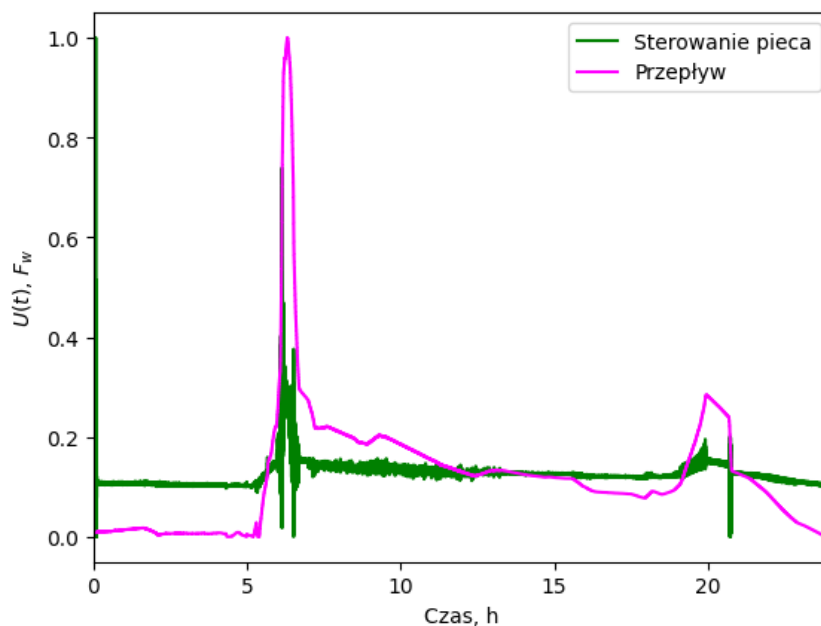


Rysunek 5.2: Przebiegi temperatur w warstwach oraz przepływu wody

W trakcie symulacji ustawiono temperaturę zadaną na poziomie 50°C. Warstwa regulowana, czyli druga, utrzymywała się stabilnie na tym poziomie przez cały czas trwania symulacji. Jedynie na początku obserwowano pewne odchylenie, które wynikało z nieidealnego dostosowania początkowych wartości wymaganych przez metodę Eulera. Sterowanie piecem było początkowo ustawione na 50%, co skutkowało chwilowymi wahaniami temperatury warstwy drugiej.

Temperatura warstwy wejściowej utrzymywała się między 10°C a 16°C. Od godziny 0 do godziny 15, gdy przepływ był bliski zera, obserwowano nagrzewanie wody w warstwie pierwszej. W momencie gwałtownego zużycia wody około godziny 6, temperatura w warstwie pierwszej spadła, a w warstwie trzeciej osiągnęła maksymalną wartość około 58°C.

Gdy zużycie wody ustabilizowało się na poziomie około 0.5 l/min, temperatura warstwy pierwszej utrzymywała się na stałym poziomie 10°C, a temperatura warstwy trzeciej stopniowo spadała. Około godziny 19, podczas zwiększonego zużycia wody, temperatura warstwy trzeciej wzrosła w sposób zbliżony do wzrostu zużycia wody. Zmniejszenie zużycia wody użytkowej w ostatniej godzinie symulacji skutkowało wzrostem temperatury pierwszej warstwy i jednoczesnym obniżeniem temperatury ostatniej warstwy.



Rysunek 5.3: Porównanie wartości sterowania oraz wartości przepływu

Analizując wykres 5.3, można zauważyć początkowe duże wartości sterowania, związane z ich przybliżonym ustawieniem wartości początkowych sterowania. Bezpośrednio po tych początkowych fluktuacjach, sterowanie stabilizuje się na poziomie około 10% i utrzymuje tę wartość do godziny 6. W tym czasie występuje zwiększone zużycie wody, co skutkuje dynamicznymi wahaniami wartości sterowania pieca, oscylującymi od 0% aż do 45%. Następnie, w miarę upływu czasu, wartość sterowania stabilizuje się w okolicach 18%. Warto zauważyć, że nagły spadek przepływu wody w okolicach godziny 20:30 jest bezpośrednio odzwierciedlony na wykresie poprzez gwałtowny spadek wartości sterowania.

5.2 Optymalizacja

Funkcja kosztów

Funkcja kosztów została zdefiniowana jako ilość energii wydzielana przez piec w okresie badawczym. Wzór tej funkcji można zapisać jako:

$$G = \int_0^T p_1 Q_g \cdot u(t) dt, \quad (5.6)$$

gdzie:

- $u(t)$ - Sterowanie systemem grzewczym w chwili t
- Q_g - Ciepło systemu grzewczego
- T - czas trwania badanego okresu

Funkcja komfortu

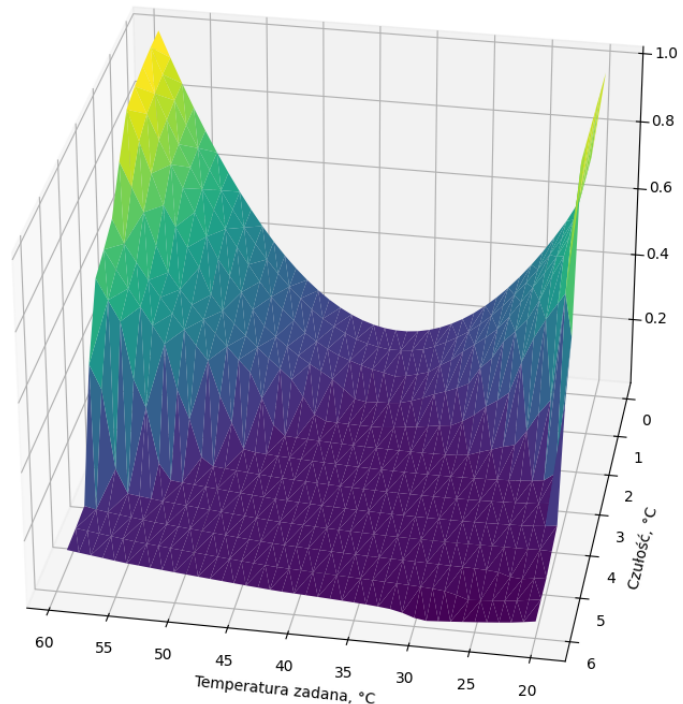
Funkcja komfortu określa w jakim stopniu temperatura rzeczywista różni się od wartości oczekiwanej, z uwzględnieniem czułości. Wzór tej funkcji można zapisać jako:

$$J = \int_0^T (T_{wo}(t) - T_{wym})^2 \left| \frac{\text{sign}(T_{wo}(t) - T_{wym} - \delta) + \text{sign}(T_{wo}(t) - T_{wym} + \delta)}{2} \right| dt, \quad (5.7)$$

gdzie:

- $T_{wo}(t)$ - Temperatura wyjściowa w chwili t
- T_{wym} - Temperatura zadana
- δ - Czułość
- T - czas trwania badanego okresu

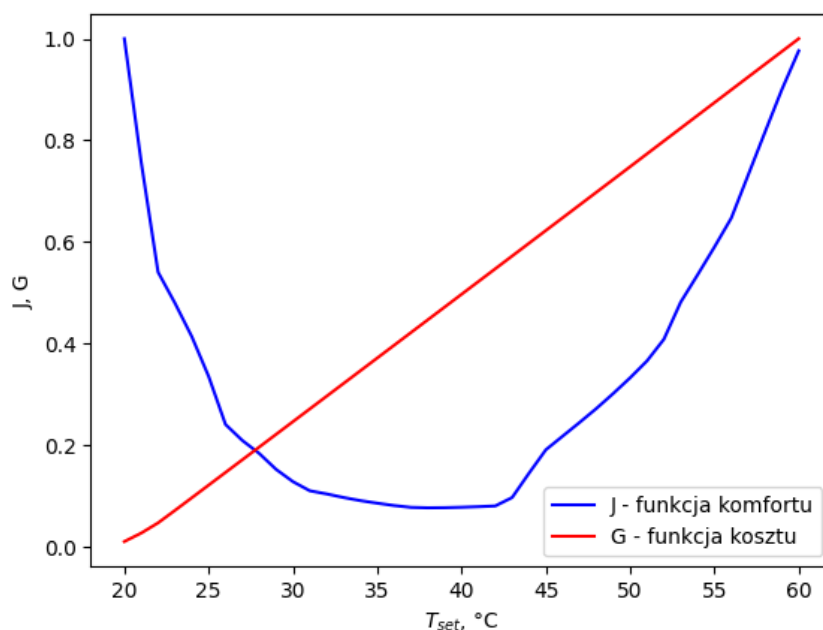
Wartość funkcji komfortu jest uzależniona od dwóch kluczowych parametrów: temperatury wymaganej T_{wym} oraz czułości δ . W związku z tym, postanowiono dokładniej zbadać, w jaki sposób zmiana tych parametrów wpływa na ostateczną wartość funkcji komfortu dla danego układu.



Rysunek 5.4: Wartość funkcji komfortu w zależności od temperatury zadanej oraz czułości

Analizując wykres 5.4, można zauważyć, że wraz ze wzrostem wartości czułości, wartość funkcji komfortu maleje niezależnie od wybranej temperatury zadanej. Dla czułości przekraczającej 5 stopni Celsjusza, funkcja komfortu nie osiąga wartości większej niż 0.1 po znormalizowaniu.

Najniższą wartość komfortu uzyskano dla czułości równej 6 i temperatury zadanej wynoszącej 29 stopni Celsjusza. Dla niewielkich wartości czułości można zauważyć szybko rosnącą wartość funkcji komfortu w miarę oddalania się temperatury zadanej od wartości około 36 stopni Celsjusza, gdzie wraz ze zwiększeniem czułości tempo narastania maleje.



Rysunek 5.5: Wartość funkcji komfortu i kosztów w zależności od temperatury zadanej

W kontekście procesu optymalizacji, przyjęto wartość czułości równą 2. Stwierdzono, że dla ustalonej czułości, funkcja komfortu najniższą wartość osiąga dla temperatury zadanej równej 41°C. W miarę zmiany tej temperatury, wartość funkcji komfortu systematycznie rośnie. Zauważono również, że funkcja kosztów jest funkcją liniową w całym badanym zakresie temperatur, gdzie jej wartość wzrasta wraz z rosnącą temperaturą zadaną.

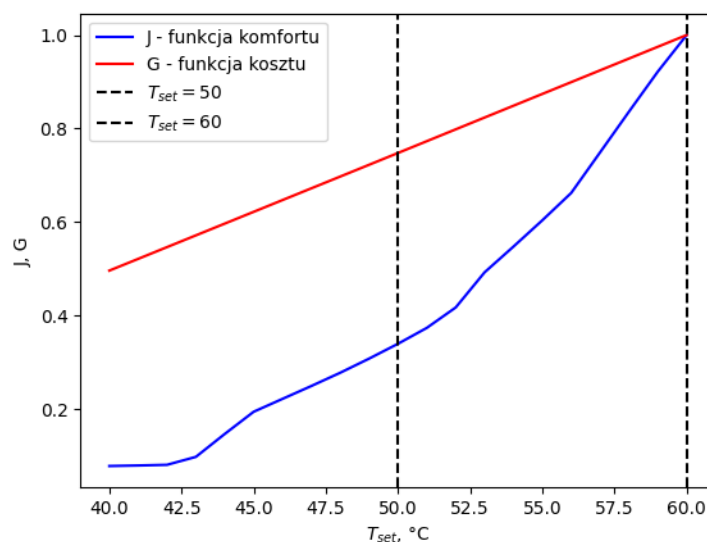
W celu optymalizacji zastosowano funkcję celu, która przypisuje wagi każdej z funkcji, a których suma, podzielona przez sumę wag, jest minimalizowana. Przeprowadzono także analizę wpływu zmian wag na optymalną temperaturę w ramach procesu optymalizacyjnego.

Tabela 5.2: Porównanie optymalnej temperatury wyjściowej w zależności od wag funkcji celu

Waga funkcji komfortu	Waga funkcji kosztów	Optymalna temperatura zadana
0.5	0.5	29
0.3	0.7	26
0.3	0.9	26
0.1	0.9	22
0.7	0.3	31
0.9	0.3	31
0.9	0.1	36

W wyniku analizy zauważono, że wraz ze zwiększaniem ważności funkcji kosztów, optymalna temperatura zadana systematycznie maleje, osiągając wartość 22 stopni Celsjusza, gdy funkcji kosztów przypisuje się wagę dziewięciokrotnie większą niż funkcji komfortu. Odpowiednio, zwiększenie ważności funkcji komfortu działa odwrotnie na optymalną temperaturę zadaną – wraz z podniesieniem jej wagi, optymalna temperatura wzrasta, osiągając 36 stopni Celsjusza, gdy funkcja komfortu ma dziewięciokrotnie większą wagę niż funkcja kosztów.

Zgodnie z Rozporządzeniem Ministra Infrastruktury [14, 15], instalacja ciepłej wody powinna uzyskiwać temperaturę w zakresie od 55 do 60. Istnieje konieczność przestrzegania tych norm, ponieważ jest to związane między innymi z ograniczeniem rozwoju bakterii Legionella. Bakterie te rozwijają się najlepiej w temperaturze pomiędzy 20 a 45 stopni Celsjusza, a giną gdy temperatura utrzymuje się na poziomie 50°C przez 2 godziny[16]. Z uwagi na wymagania sanitarno-epidemiologiczne i konieczność optymalizacji systemu, postanowiono przeprowadzić optymalizację temperatury wyjściowej, przy jednoczesnym utrzymaniu temperatury zadanej w zakresie od 50 do 60 stopni Celsjusza. Choć ta temperatura jest niższa niż wartość ustawowo zalecana, odnosi się ona do warstwy drugiej, która charakteryzuje się niższą temperaturą niż warstwa wyjściowa w momentach zwiększonego przepływu.



Rysunek 5.6: Wartość funkcji komfortu i kosztów z zaznaczonymi ograniczeniami

Po wprowadzeniu opisanych ograniczeń zaobserwowano, że obie funkcje (kosztów i komfortu) wykazują tendencję wzrostową w całym zakresie badanych temperatur. Oznacza to, że obie funkcje przyjmują najniższe wartości na początku analizowanego przedziału temperatur. Dodatkowo przeprowadzono badanie, w którym ustawiono czułość funkcji komfortu na 4 stopnie Celsjusza. Mimo wprowadzenia tej zmiany funkcja komfortu nie zmieniła swojej charakterystyki. Po wprowadzeniu opisanych ograniczeń ustalono, że optymalną temperaturą zadana jest wartość $T_{set} = 50^{\circ}\text{C}$.

Rozdział 6

Podsumowanie i wnioski

W kontekście przyszłych prac nad projektem zaleca się rozważenie nauczania modelu, który został zaproponowany przez Bakker et al. i porównanie wyników tego modelu z rezultatami modelu opisanego w rozdziale 3. Mimo że model Bakker’a nie został pierwotnie opracowany z myślą o przewidywaniu zużycia wody, to istnieje potencjał, że oba te zjawiska mogą wykazywać podobne charakterystyki. Podobieństwo to może wynikać z podobnych korelacji czasowych oraz sezonowych. Przyszłe prace badawcze mogą więc skoncentrować się na adaptacji modelu Bakker’a do przewidywania zużycia wody.

Wartym uwagi rozwiązaniem, które może poprawić działanie modelu jest zastosowanie modelu LSTM (Long Short-Term Memory), który jest specjalnym rodzajem rekurencyjnej sieci neuronowej skupiającym się na rozwiązaniu problemu zanikającego gradientu. LSTM jest szczególnie efektywny w zadaniach dotyczących przetwarzania szeregów czasowych, dzięki zdolności do długotrwałego przechowywania i aktualizacji informacji.

W celu poprawy wyników badań nad ważnością warunków pogodowych proponowane jest zastosowanie sieci neuronowej posiadającej trzy wyjścia w celu uwzględnienia danych z różnych źródeł. Dostarczone dane zawierały dane z trzech czujników. Dane zebrane z czujnika zastosowana nad kuchenką gazową oraz na kominku zostały pominięte w badaniach. Użycie tych informacji mogłoby popawić wyniki dla każdej z metod sprawdzania ważności wejść.

Zastosowanie metody Holdout, w której zbiór danych jest dzielony na zbiór treningowy i testowy, mogło być mniej adekwatne w sytuacji, gdzie porównanie jakości przewidywań w warunkach nie znanych modelowi jest mniej ważne od dokładności decyzyjności w warunkach znanych. Lepszym wyborem była by metoda Resubstytucji, która zapewnia pełne wykorzystanie danych do nauki modelu jak i do jego testowania co przekłada się na pozytywne obciążenie oraz na dokładniejszą naukę na dostępnych danych, lecz nie obrazu decyzyjności modelu w nieznanych warunkach.

Bibliografia

- [1] A. Laperrière S. Edwards I. Beausoleil-Morrison. „Representative hot water draw profiles at high temporal resolution for simulating the performance of solar thermal systems”. W: *Solar Energy* 111 (2015), s. 43–52.
- [2] K. Gamage L. Gelažanskas. „Forecasting hot water consumption in residential houses”. W: *Energies* 8 (2015), s. 12702 –1271.
- [3] N.D. Pflugradt. *Modeling water and energy consumption in households*. 2016. URL: <https://nbn-resolving.org/urn:nbn:de:bsz:ch1-qucosa-209036> (term. wiz. 08.01.2024).
- [4] Carriere Claquin i Rocaries. *Modelling and application of adaptive control to a gas heater*. Glasgow, UK: Proceedings of IEEE International Conference on Control i Applications, 1994.
- [5] S. Alizadeh. „An Experimental and Numerical Study of Thermal Stratification in a Horizontal Cylindrical Solar Storage”. W: *Solar Energy* (1999).
- [6] J. Streicher M. Y. Haller. „Comparative Analysis of Thermal Energy Storage Stratification Efficiency”. W: *Thermal Energy Storage for Efficiency and Sustainability* (2009).
- [7] Stack Overflow. *Developer Survey*. 2023. URL: <https://survey.stackoverflow.co/2023> (term. wiz. 17.01.2024).
- [8] Google Colab. *Welcome to Colaboratory*. 2023. URL: <https://cloud.google.com/tpu/docs/intro-to-tpu> (term. wiz. 08.01.2024).
- [9] Google Cloud. *Introduction to Cloud TPU*. 2023. URL: <https://cloud.google.com/tpu/docs/intro-to-tpu> (term. wiz. 08.01.2024).
- [10] R.G. Palmer J. Hertz A. Krogh. *Wstęp do teorii obliczeń neuronowych*. Warszawa: WNT, 1993.
- [11] S. Osowski. *Sieci neuronowe w ujęciu algorytmicznym*. Warszawa: WNT, 1996.
- [12] M.J. Booysen M.J. Ritchie J.A.A. Engelbrecht. „A probabilistic hot water usage model and simulator for use in residential energy management”. W: *Energy & Buildings* 8 (2021).

- [13] Sujith Mathew Iype. *Human Brain vs Artificial Intelligence Systems*. 2018. URL: <https://ignitarium.com/human-brain-vs-existing-artificial-intelligence-systems/> (term. wiz. 21.01.2024).
- [14] *Rozporządzenie Ministra Infrastruktury z dnia 12 kwietnia 2002 r. w sprawie warunków technicznych, jakim powinny odpowiadać budynki i ich usytuowanie*. Dziennik Ustaw Rzeczypospolitej Polskiej. Numer 75, pozycja 690. 2002.
- [15] *Obwieszczenie Ministra Infrastruktury i Rozwoju z dnia 17 lipca 2015 r. w sprawie ogłoszenia jednolitego tekstu rozporządzenia Ministra Infrastruktury w sprawie warunków technicznych, jakim powinny odpowiadać budynki i ich usytuowanie*. Dziennik Ustaw Rzeczypospolitej Polskiej. Pozycja 1422. 2015.
- [16] Vaillant. *Temperatura CWU (cieplej wody użytkowej) – zalecenia i optymalizacja*. 2022. URL: <https://www.vaillant.pl/klienci-indywidualni/porady-i-wiedza/poradnik/inne/temperatura-cwu/> (term. wiz. 25.01.2024).

Dodatki

Spis skrótów i symboli

ASHRAE American Society of Heating, Refrigerating and Air-Conditioning Engineer

ARIMA Autoregressive integrated moving average

MISO Multiple Input, Single Output

MSE Mean squared error

CUDA Compute Unified Device Architecture

GPU Graphics processing unit

CWU Ciepła woda użytkowa

CPU Central processing Unit

LIME Local Interpretable Model-agnostic Explanations

PFI Permutation Feature Importance

LSTM Long Short-Term Memory

$f(\cdot)$ Funkcja aktywacji neuronu

y sygnał wyjściowy neuronu

x sygnał wejściowy neuronu

v pobudzenie neuronu

w_j Waga neuronu

b Składnik stały

$T_{zi}^{n,m}$ Temperatura systemu grzewczego

T_{wo}^n Temepatura warsty

T_{ot} Temperatura otoczenia

F_w Przepływ wody w zbiorniku

F_z Przepływ wody systemu grzewczego

Q_p Ciepło systemu grzewczego

Q_{zm} Ciepło dostarczane do zasobnika

Q_m TODO

Q_w Ciepło całego zbiornika

Q_{wm} Ciepło warstwy

Q_u Ciepło upuszczające zbiornik DO POPRAWY TODO

Q_s Straty ciepła

A Powierzchnia wymiany ciepła DO POPRAWY TODO

d powierzchnia wymiany ciepła DO POPRAWY TODO

V Objętość

C_w Ciepło właściwe

δ TODO

λ Współczynnik wymiany ciepła

n numer warstwy

m numer źródła ciepła

ϕ_{zew} wilgotność zewnętrzna

ϕ_{wew} wilgotność wewnętrzna

\longrightarrow Kierunek wiatru

V_{wia} Prędkość wiatru

T_{zew} Temperatura zewnętrzna

P_{zew} Ciśnienie zewnętrzne

Lista dodatkowych plików, uzupełniających tekst pracy

1. Dane

- (a) Dane_1 - Dane użyte do nauki sieci neuronowej przewidującej życie wody oraz ich wersja znormalizowana wersja.
- (b) Dane_2 - Dane użyte do nauki modelu służącego do oceny ważności wejść.

2. Skrypty

- (a) Przygotowanie_danych.IPYNB - Skrypt przetwarzający dane Dane_1
- (b) Przeply.IPYNB - Skrypt uczący modele typu A, B, C, D oraz wyświetlający wyniki symulacji.
- (c) Przeply_dodatkowy.IPYNB - Skrypt uczący model typu C dodając do niego dane historyczne
- (d) Pogoda.IPYNB - Skrypt uczący model oraz badający ważność wejść trzema metodami.
- (e) Symulacja.IPYNB - Skrypt symulujący model trzywarstwowy zbiornika CWU wraz z obliczaniem wartości komfortu i kosztów

3. Modele

- (a) Modele typu A, B, C, C+d, D przewidujące zużycie wody
- (b) Modele o uproszczonej i rozszerzonej architekturze, służące do sprawdzenia ważności wejść

Spis rysunków

3.1	Model neuronu	9
3.2	Porównanie wartości przepływów w różnych ramach czasowych.	11
3.3	Porównanie przepływów dla przykładowych domów w dniu 05/02/2018 . .	12
3.4	Fragment skryptu przetwarzającego dane.	13
3.5	Porównanie symulacji dla modeli typu A	16
3.6	Porównanie symulacji dla modeli typu B	17
3.7	Porównanie symulacji dla modeli typu C	18
3.8	Porównanie symulacji dla modeli typu D	19
3.9	Porównanie modelu ogólnego z modelami szczególnymi	21
3.10	Porównanie modelu po dodaniu kolejnego wejścia do sieci	22
4.1	Porównanie warunków atmosferycznych na przestrzeni dnia	25
4.2	Mapa ciepła wag pierwszej warstwy dla modelu o uproszczonej architekturze	28
4.3	Mapa ciepła wag pierwszej warstwy dla modelu o rozszerzonej architekturze	29
4.4	Mapy cieplne modeli z dodatkowym wejściem	34
5.1	Trójwarstwowy model zasobnika	38
5.2	Przebiegi temperatur w warstwach oraz przepływy wody	40
5.3	Porównanie wartości sterowania oraz wartości przepływu	41
5.4	Wartość funkcji komfortu w zależności od temperatury zadanej oraz czułości	42
5.5	Wartość funkcji komfortu i kosztów w zależności od temeperatury zadanej .	43
5.6	Wartość funkcji komfortu i kosztów z zaznaczonymi ograniczeniami	45

Spis tabel

3.1	Porównanie wartości MSE dla różnych modeli przy zróżnicowanej ilości danych	20
3.2	Porównanie wartości MSE dla dwóch modeli w kontekście poszczególnych domostw	23
4.1	Porównanie ważności wejść dla metody PFI	27
4.2	Porównanie ważności wejść dla metody PFI	30
4.3	Porównanie ważności wejść dla metody LIME	31
4.4	Porównanie uśrednionych ważności wejść dla metody LIME dla wielu prób	32
4.5	Wartości MSE modeli uproszczonych i rozszerzonych z uwzględnieniem dodatkowego parametru	33
4.6	Porównanie ważności wejść dla wszystkich metod	34
5.1	Współczynniki użyte podczas modelowania	39
5.2	Porównanie optymalnej temperatury wyjściowej w zależności od wag funkcji celu	44