1. Michał MAJ[1,2], 2. Damian PLISZCZUK[2], 3. Tomasz CIEPLAK[3], 4. Łukasz MACURA[2]

Akademia WSEI(1), Centrum Badawczo-Rozwojowe Netrix S.A.(2), Politechnika Lubelska(3)
ORCID: 1. 0000-0002-7604-8559; 2. 0000-0002-5727-979X; 3. 0000-0002-2712-6098; 4. 0000-0001-8657-3472

doi:10.15199/48.2025.03.21

Destylacja wiedzy w głębokim uczeniu za pomocą sieci multimodalnych

Streszczenie. *Artykuł opisuje zastosowanie destylacji wiedzy w głębokim uczeniu z wykorzystaniem sieci multimodalnych do stworzenia inteligentnego urządzenia. Opracowany model jednocześnie wykrywa osobę na obrazie, klasyfikuje jej emocje oraz sprawdza, czy jest zarejestrowanym użytkownikiem. Dzięki destylacji wiedzy uzyskano mniejszy model o zbliżonej skuteczności, który działa w czasie rzeczywistym na urządzeniach wbudowanych. Wyniki pokazują, że techniki multimodalne i destylacja wiedzy poprawiają wydajność modeli.*

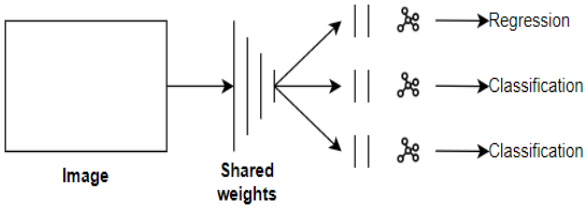
Abstract*. The article describes the application of knowledge distillation in deep learning using multimodal networks to create an intelligent device. The developed model simultaneously detects a person in an image, classifies their emotions, and verifies if they are a registered user. Thanks to knowledge distillation, a smaller model with similar effectiveness was obtained, which operates in real-time on embedded devices. The results show that multimodal techniques and knowledge distillation improve model performance. (**Knowledge distillation in deep learning via multimodal networks**).*

Słowa kluczowe: Destylacja wiedzy, sieci multimodalne, wykrywanie emocji, systemy wbudowane.
Keywords: Knowledge distillation, multimodal networks, emotion detection, embedded systems.

Wstęp

Destylacja wiedzy jest technika stosowaną w głębokim uczeniu. Jej celem jest przeniesienia wiedzy z większego, bardziej złożonego modelu, zwanego „nauczycielem" do mniejszego, bardziej wydajnego modelu - „ucznia"[1]. Umożliwia to uzyskanie temu drugiemu skuteczności zbliżonej do modelu nauczyciela, a jednocześnie wymaga mniej zasobów [2].

W przypadku tej techniki warto rozróżnić dwa etapy. Pierwszy z nich to etap ucznia. W tym przypadku trening odbywa się na dużym zbierze danych z wykorzystaniem mocy obliczeniowej kart graficznych. Można więc zauważyć, że nie ma potrzeby rozwiązywania danego problemu w czasie rzeczywistym. Drugim etapem jest etap inferencji na rzeczywistych danych. Model w takim przypadku nie ma już dostępu do danych uczących. Dodatkowo większość obliczeń będzie wykonywana na CPU i wymagane jest działanie w czasie rzeczywistym [3-5]. Stąd też pojawiła się idea destylacji wiedzy, gdzie na potrzeby treningu budowany jest większy model wymagający dużej mocy obliczeniowej. Następnie na podstawie tego modelu budowany jest mniejszy, bardziej wydajny model. Warto także zaznaczyć, że do badań zostanie użyty model multimodalny (Rys. 1).



Rys.1. Przykład ostatniej warstwy sieci multimodalnej

Cel

Głównym założeniem projektu jest opracowanie algorytmów dla prototypu inteligentnego urządzenia konsjerż, które ma pełnić funkcję osobistego doradcy klienta w postaci totemu. Będzie to końcowym elementem systemu, z którym użytkownik będzie miał bezpośredni kontakt (Rys. 2). W ramach jednego modelu głębokiego uczenia zbudowano rozwiązanie związane z sieciami multimodalnymi. Głównym założeniem było wykrycie osoby na obrazie i określenie emocji tej osoby na podstawie mimiki twarzy oraz sklasyfikowanie czy osoba istnieje już w bazie danych. Wyuczony model powinien wykryć osobę na podstawie zdjęcia oraz emocji towarzyszących osobie z nagrania wideo. Więc jeden model odpowiadał za klasyfikację osoby na obrazie, sklasyfikowanie jej emocji i zwrócenie informacji, czy dana osoba jest już zarejestrowanym użytkownikiem (Rys. 1). Dodatkowo wykorzystano techniki destylacji wiedzy, aby zapewnić działanie modelu w czasie rzeczywistym na systemach wbudowanych takich jak Raspberry Pi oraz Jetson Nano.

Rys.2. Obecna iteracja konsjerża

Model został przeszkolony na obrazach i filmach zarejestrowanych podczas debaty politycznej w dniu 17 czerwca 2020 roku, która była częścią wyborów prezydenckich w Polsce w 2020 roku. Natomiast informację, kim jest dana osoba dodano ręcznie – tak jakby użytkownik sam wpisywał swoje dane. Na potrzeby badania przygotowano cztery modele, trzy pojedyncze, gdzie każdy osobno był odpowiedzialny za wykrycie osoby, detekcje emocji i potwierdzenie czy dana osoba istnieje w bazie danych (Rys. 3). Dodatkowo powstał czwarty model, który wykonywał wszystkie powyższe czynności za pomocą sieci

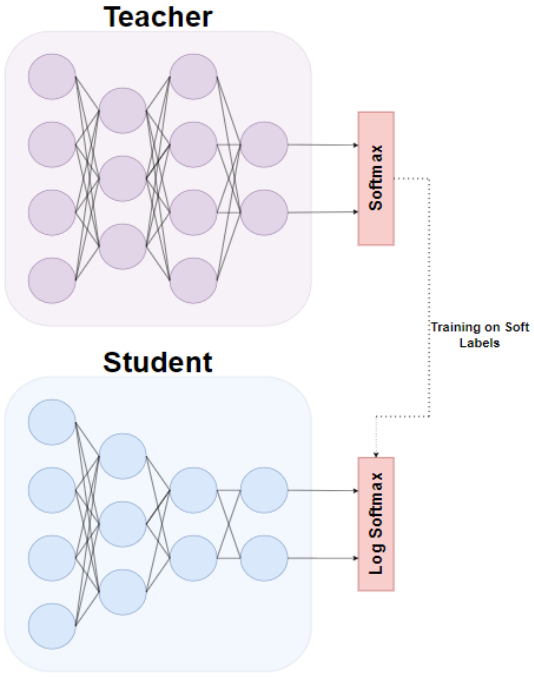
multimodalnych. Warto także dodać, że we wszystkich czterech przypadkach zastosowano destylację wiedzy i porównano czasy i skuteczność inferencji [6-10].

Rys.3. Przykład wykrywania osoby i jej emocji

Metody

W przeciwieństwie do transferu wiedzy, gdzie używany jest ten sam model do treningu i inferencji, a następnie podmieniana jest „głowa" – wyjście z wyuczonej sieci, destylacja wiedzy wykorzystuje dwa osobne modele. W pierwszy przypadku współdzielone są pomiędzy modelami wagi, natomiast w tym drugi generalizacja. Więc jest to jedna z form optymalizacji modelu.

W związku z tym celem jest przygotowanie mniejszego modelu, który będzie w stanie wykonywać te same zadania co większy model, mogąc przy tym wykonywać swoją pracę w czasie rzeczywistym. Aby można było tego dokonać jednym z elementów jest trening na miękkich etykietach (ang. Soft Labels). Oznacza, to że kiedy model dokonuje predykcji, otrzymuje się wartość na ile procent w skali 0-1 dana klasa jest przewidziana. Więc dla powyższego wyjścia (Rys. 2) miękka etykieta ma wartość 0.92. W tradycyjnych sieciach neuronowych stosowany jest następnie one-hot enconding, aby otrzymać twardą etykietę. Natomiast podczas trenowania modelu ucznia w destylacji wiedzy korzysta się z miękkich etykiet (Rys. 4).



Rys.4. Przykład treningu na miękkich etykietach

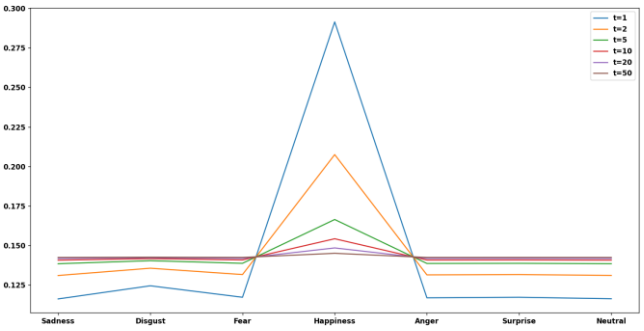
W procesie nauki używa się funkcji *Temperature Softmax* (Softmax z temperaturą). Różni się ona od klasycznej wersji tym, że wynik predykcji danej klasy dzieli się najpierw przez pewną wartość temperatury T.

$$Temperature\ Softmax = \frac{exp(x_i/T)}{\sum_i exp(x_i/T)} \quad (1)$$

Temperatura w funkcji Softmax odgrywa istotną rolę w zmiękczaniu rozkładów prawdopodobieństwa generowanych przez model nauczyciela. Wyższa wartość temperatury powoduje, że rozkład prawdopodobieństwa staje się bardziej płaski, a różnice między prawdopodobieństwami poszczególnych klas są mniejsze. Dzięki temu model uczeń może lepiej uchwycić subtelne relacje między klasami, co jest szczególnie ważne w przypadku zadań, gdzie klasy są do siebie podobne, jak na przykład w klasyfikacji emocji.

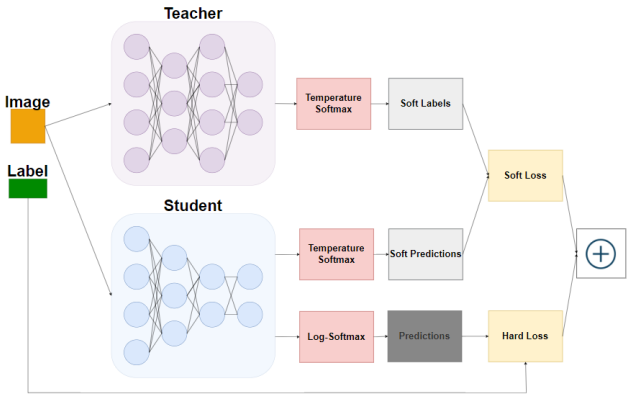
Jednak zbyt wysoka wartość temperatury może prowadzić do utraty istotnych informacji, ponieważ rozkład prawdopodobieństwa staje się zbyt jednolity, co utrudnia modelowi ucznia naukę kluczowych wzorców. Z drugiej strony, zbyt niska wartość sprawia, że funkcja Softmax zbliża się do standardowej wersji, a korzyści płynące z wykorzystania miękkich etykiet są ograniczone. Dlatego istotne jest znalezienie optymalnej wartości temperatury, która zapewni najlepszy kompromis między zmiękczeniem rozkładu a zachowaniem ważnych informacji dla procesu uczenia.

Jest to technika powiązana z miękkimi etykietami, ponieważ pozwala zauważyć czy dana klasa, która została wskazana podczas predykcji jest podobna do innej klasy (Rys. 5). Natomiast w tym przypadku, można zauważyć, że dla T=5 najbliższa do wykrytej emocji jest emocja „wstręt". Więc jest to bardziej prawdopodobne, że może to być emocja „wstręt" niż emocja „smutek". Stąd można też stwierdzić, że dana emocja bardzie wygląda na „wstręt" niż na „smutek". Dzięki temu zbudowana została hierarchia wartości, pozwala ona na zachowanie pewnej wiedzy na temat miękkich etykiet i przekazania jej od nauczyciela do ucznia.



Rys.5. Przykład funkcji Softmax z różnymi wartościami temperatury

Kolejnym elementem wpływającym na skuteczne przekazanie wiedzy jest odpowiednia funkcja straty dla destylacji wiedzy. Funkcja straty dla destylacji wiedzy łączy dwa elementy, aby efektywnie przekazać wiedzę z modelu nauczyciela do modelu ucznia. Pierwszy element to twarda strata, która mierzy różnicę między przewidywaniami ucznia a rzeczywistymi etykietami klas, zazwyczaj za pomocą entropii krzyżowej. Drugi element to miękka strata, która ocenia różnicę między zmiękczonymi rozkładami prawdopodobieństwa (miękkimi etykietami) generowanymi przez nauczyciela i ucznia, często za pomocą dywergencji Kullbacka-Leiblera. Więc wynik funkcji straty będzie stanowił wynik treningu „ucznia" lub dywergencja pomiędzy treningiem „nauczyciela" a treningiem „ucznia". W pierwszym przypadku będzie to Hard Loss a w drugim Soft Loss, wynikiem tych działań będzie suma tych wartości (Rys. 6).



Rys.6. Przykładowe zastosowanie funkcji straty

Warto podkreślić, że optymalne wartości parametrów destylacji mogą różnić się w zależności od specyfiki zadania, struktury modelu oraz jakości danych treningowych. Dlatego ważne jest przeprowadzenie odpowiednich eksperymentów w celu dostosowania tych parametrów do konkretnego zastosowania.

Ponadto, wpływ parametrów destylacji jest szczególnie istotny w kontekście sieci multimodalnych. W przypadku takich modeli, które integrują informacje z różnych źródeł (np. obrazu i wideo), odpowiednie ustawienie temperatury i współczynnika α może znacząco wpłynąć na zdolność modelu ucznia do efektywnego łączenia tych informacji. Zauważono, że w sieciach multimodalnych nieco wyższa wartość temperatury pomagała w lepszym uchwyceniu złożonych zależności między modalnościami.

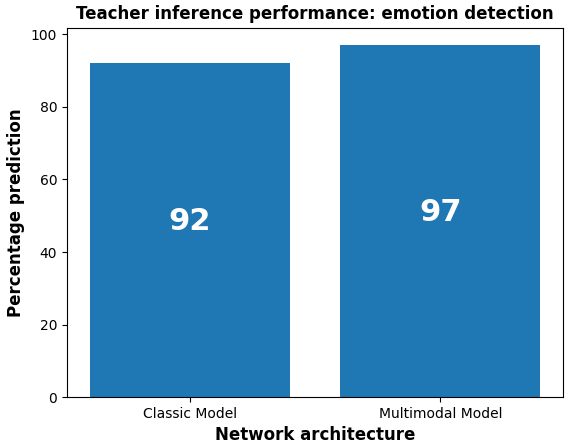
Korzystając z tych dwóch elementów nauczono cztery modelu „ucznia" bazując na czterech modelach „nauczyciela". Trening był wykonany w technice offline. Więc widzę przekazywano z wcześniej wytrenowanych większych modeli to modeli mniejszych. Warto zaznaczyć, że dla modelu multimodalnego zarówno miękkie etykiety jak i funkcja strat zostały zastosowane osobno dla obu wyjść klasyfikacji jak i jednego wyjścia regresji. Można więc określić, że model posiada wiele głów.

Wyniki

W ramach przeprowadzonych prac badawczych użyto Raspberry Pi 4B z 4 gigabajtami pamięci RAM oraz Jetson Nano firmy Nvidia, również posiadający 4 gigabajtami pamięci RAM. Korzystając z tych urządzeń przetestowano czas wnioskowania każdego z czterech modeli „nauczyciela" oraz czterech modeli „ucznia". Testy umożliwiły sprawdzenie skuteczności działania tych modeli. Zbiór testowy użyty do sprawdzenia został utworzony z 20% uzyskanych danych. Wczytywano kolejne pliki wideo i na tej podstawie badano skuteczność modeli w wykrywaniu osoby i towarzyszących jej emocji. Odrębnie porównano każdy z modeli klasycznych oraz model nauczany techniką multimodalną. Testy przeprowadzono na obu urządzeniach. Wyniki skuteczności inferencji uzyskane przez model nauczyciela prezentują się następująco (Rys. 7).

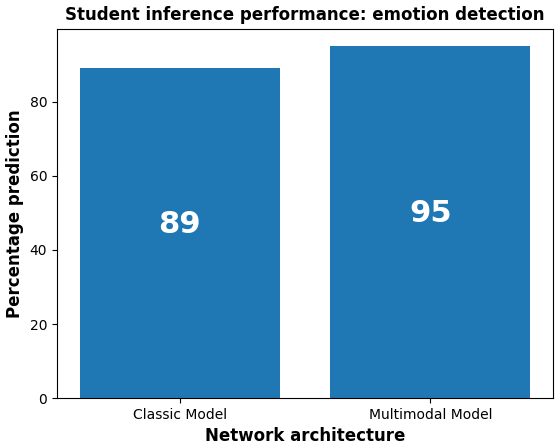
Tabela 1. Czas inferencji w sekundach dla modelu nauczyciela

Nauczyciel / Urządzenie	Wykrywanie emocji	Wykrywanie twarzy	Wykrywanie użytkowniak	Multimodalny model
Raspberry	0.33s	0.09s	0.06s	0.29s
JETSON	0.24s	0.08s	0.06s	0.15s



Rys.7. Porównanie skuteczności modeli nauczyciela w wykryciu emocji

Natomiast wynik uzyskany przez model ucznia również w zadaniu związanym z klasyfikacją emocji przedstawiono na rysunku 8.



Rys.8. Porównanie skuteczności modeli ucznia w wykryciu emocji

Dodatkowo zbadano czas potrzeby na wykonanie inferencji. Tabela 1 zawiera informację o czasie inferencji dla poszczególnych modeli nauczyciela. Natomiast Tabela 2 zawiera informację o czasie inferencji dla poszczególnych modeli ucznia. Niższa wartość oznacza, że więcej klatek obrazu na sekundę można przetworzyć, więc im mniej czasu jest wymagane na inferencję, tym system wydajniej działa.

Tabela 2. Czas inferencji w sekundach dla modelu ucznia

Uczeń / Urządzenie	Wykrywanie emocji	Wykrywanie twarzy	Wykrywanie użytkowniak	Multimodalny model
Raspberry	0.21s	0.07s	0.03s	0.19s
JETSON	0.17s	0.06s	0.01s	0.11s

Podsumowanie

Jak można zauważyć, modele nauczyciela radzą sobie lepiej, jeśli chodzi o procentowy wynik inferencji. W szczególności jest to widoczne w przypadku modeli klasycznych, gdzie wynik uzyskany przez model „ucznia" był mniejszy niż 90% (Rys. 7). Natomiast model „ucznia" sieci multimodalnych zachował swoje właściwości i uśredniona różnica w inferencji wyniosła jedynie 2%, ale nadal była większa niż 92%. Spowodowane jest to tym, że technika multimodalna umożliwia osiągnięcie lepszych rezultatów niż

pojedyncze klasyczne modele, łącząc informacje z różnych modalności i pozwalając na lepsze zrozumienie danych.

Warto także dodać, że porównując wyniki z Tabeli 1 oraz Tabeli 2, można zauważyć, że czas inferencji modelu „ucznia" jest zawsze lepszy niż modelu „nauczyciela". Jest to istotne w kontekście zastosowań w systemach wbudowanych, gdzie zasoby sprzętowe są ograniczone, a czas reakcji systemu ma znaczenie. Z badań wynika także, że model multimodalny jest wolniejszy niż pojedynczy model klasyczny. Natomiast należy wziąć pod uwagę, że w tym przypadku ten model wykonuje pracę trzech klasycznych modeli jednocześnie. Więc suma czasów tych oddzielnych modeli jest zawsze większa niż jednego modelu multimodalnego, co ostatecznie przekłada się na większą efektywność całego systemu.

Na podstawie tego można więc stwierdzić, że zastosowanie technik multimodalnych przynosi korzyści zarówno ze względu na wydajność modelu, jak i procentowy wynik inferencji. Dodatkowo zastosowanie destylacji wiedzy pozwala w sieciach multimodalnych na uzyskanie lepszej wydajności kosztem niewielkiej straty procentowej inferencji. Jest to akceptowalny kompromis, szczególnie gdy celem jest implementacja modeli na urządzeniach o ograniczonej mocy obliczeniowej, takich jak Raspberry Pi czy Jetson Nano.

Wyniki te wskazują, że destylacja wiedzy jest efektywną metodą redukcji rozmiaru i złożoności modeli głębokiego uczenia bez znacznej utraty ich skuteczności. Dzięki temu możliwe jest wdrażanie zaawansowanych algorytmów sztucznej inteligencji w środowiskach o ograniczonych zasobach, co otwiera nowe możliwości dla zastosowań w realnym świecie.

Ponadto, zaobserwowana niewielka różnica w skuteczności pomiędzy modelami nauczyciela i ucznia sugeruje, że model ucznia jest w stanie efektywnie uczyć się od modelu nauczyciela, zachowując informacje potrzebne do wykonywania zadań takich jak wykrywanie twarzy, klasyfikacja emocji czy identyfikacja użytkownika. To pokazuje potencjał destylacji wiedzy jako narzędzia do tworzenia lekkich i efektywnych modeli.

Należy jednak zwrócić uwagę na pewne ograniczenia przeprowadzonych badań. Po pierwsze, trening i testowanie modeli zostało przeprowadzone na specyficznym zbiorze danych pochodzących z debaty politycznej, co może wpływać na ogólność uzyskanych wyników. W przyszłości warto byłoby rozszerzyć eksperymenty na bardziej zróżnicowane dane, aby sprawdzić, czy obserwowane korzyści utrzymują się w innych kontekstach.

Po drugie, chociaż modele ucznia wykazują zwiększoną wydajność, to jednak pewna utrata dokładności jest nieunikniona. W zależności od zastosowania, nawet niewielki spadek skuteczności może być istotny. Dlatego ważne jest dokładne zrozumienie wymagań konkretnego systemu i ewentualne dostosowanie parametrów destylacji, takich jak temperatura czy współczynnik α, aby zoptymalizować równowagę pomiędzy wydajnością a dokładnością.

Kolejnym aspektem wartym rozważenia jest możliwość dalszej optymalizacji modeli poprzez zastosowanie innych technik, takich jak kwantyzacja sieci neuronowych. Mogłoby to dodatkowo zmniejszyć zapotrzebowanie na zasoby obliczeniowe, jednocześnie minimalizując utratę dokładności.

Przeprowadzone badania potwierdzają, że połączenie destylacji wiedzy z sieciami multimodalnymi stanowi obiecujące podejście do tworzenia wydajnych i skutecznych modeli głębokiego uczenia, zdolnych do działania na urządzeniach o ograniczonej mocy obliczeniowej. Otwiera to drzwi do szerokiego zakresu zastosowań, od inteligentnych urządzeń konsumenckich po systemy monitoringu czy interfejsy człowiek-maszyna, gdzie szybka i precyzyjna analiza danych w czasie rzeczywistym jest bardzo ważna.

W przyszłych pracach warto skupić się na eksploracji innych architektur sieci multimodalnych, a także na integracji dodatkowych modalności, takich jak dźwięk czy tekst. Umożliwiłoby to na zwiększenia zdolności modelu do rozumienia kontekstu i interakcji z użytkownikiem. Ponadto, dalsze badania nad optymalizacją procesu destylacji wiedzy mogą przyczynić się do jeszcze lepszego zachowania równowagi pomiędzy wydajnością a dokładnością modeli.

Ostatecznie, uzyskane wyniki podkreślają znaczenie dostosowywania zaawansowanych technik głębokiego uczenia do potrzeb systemów wbudowanych. Dzięki temu możliwe jest tworzenie inteligentnych urządzeń, które są nie tylko efektywne pod względem obliczeniowym, ale także potrafią dostarczyć użytkownikom wartościowych funkcjonalności w czasie rzeczywistym. To istotny krok w kierunku powszechnej integracji sztucznej inteligencji w codziennym życiu i technologii przyszłości.

***Autorzy**: mgr inż. Michał Maj, Lubelska Akademia WSEI, Wydział Transportu i Informatyki, ul. Projektowa 4, 20-209 Lublin, E-mail: michal.maj@wsei.pl; mgr Damian Pliszczuk, Netrix Link, ul. Związkowa 26, 20-148 Lublin, Email: damian.pliszczuk@netrix.com.pl; dr inż. Tomasz Cieplak, Politechnika Lubelska, Katedra Organizacji Przedsiębiorstwa, ul. Nadbystrzycka 38, 20-618 Lublin; dr Łukasz Maciura, Netrix Link, ul. Związkowa 26, 20-148 Lublin, Email: lukasz.maciura@netrix.com.pl.*

LITERATURA

[1] Gou J., Yu B., MaybankS. J. , TaoD., Knowledge Distillation: A Survey, Int J Comput Vis, 129 (2020), No. 6, 1789–1819

[2] Rybak G, Kozłowski E, Król K, Rymarczyk T, Sulimierska A, Dmowski A, Bednarczuk P. Algorithms for Optimizing Energy Consumption for Fermentation Processes in Biogas Production*. Energies,* 16 (2023); No. 24, 7972

[3] Baran B, Kozłowski E, Majerek D, Rymarczyk T, Soleimani M, Wójcik D. Application of Machine Learning Algorithms to the Discretization Problem in Wearable Electrical Tomography Imaging for Bladder Tracking*, Sensors,* 23 (2023); No. 3,1553

[4] Przysucha B, Wójcik D, Rymarczyk T, Król K, Kozłowski E, Gąsior M. Analysis of Reconstruction Energy Efficiency in EIT and ECT 3D Tomography Based on Elastic Net*. Energies,* 16 (2023); No. 3, 1490

[5] Kozłowski E., Borucka A., Oleszczuk P., Jałowiec T., Evaluation of the maintenance system readiness using the semi-Markov model taking into account hidden factors, *Eksploatacja i Niezawodność – Maintenance and Reliability,* 25 (2023) ;No. 4, 172857

[6] Panskyi T, Korzeniewska E, Firych-Nowacka A. Educational Data Clustering in Secondary School Sensor-Based Engineering Courses Using Active Learning Approaches*. Applied Sciences,* 14 (2024), No.12, 5071

[7] Kulisz M., Kłosowski G., Rymarczyk T., Hoła A., Niderla K., Sikora J., The use of the multi-sequential LSTM in electrical tomography for masonry wall moisture detection*, Measurement,* 234 (2024) 114860*.*

[8] Kulisz M, Kłosowski G, Rymarczyk T, Słoniec J, Gauda K, Cwynar W. Optimizing the Neural Network Loss Function in Electrical Tomography to Increase Energy Efficiency in Industrial Reactors*. Energies,* 17 (2024); No. 3, 681

[9] Kłosowski G, Rymarczyk T, Niderla K, Kulisz M, Skowron Ł, Soleimani M. Using an LSTM network to monitor industrial reactors using electrical capacitance and impedance tomography – a hybrid approach, *Eksploatacja i Niezawodnosc – Maintenance and Reliability*, 25 (2023). No.1

[10] Król, K., Rymarczyk, T., Niderla, K., & Kozłowski, E., Sensor platform of industrial tomography for diagnostics and control of technological processes. *Informatyka, Automatyka, Pomiary W Gospodarce I Ochronie Środowiska,* 13 (2023), No.1, 33–37