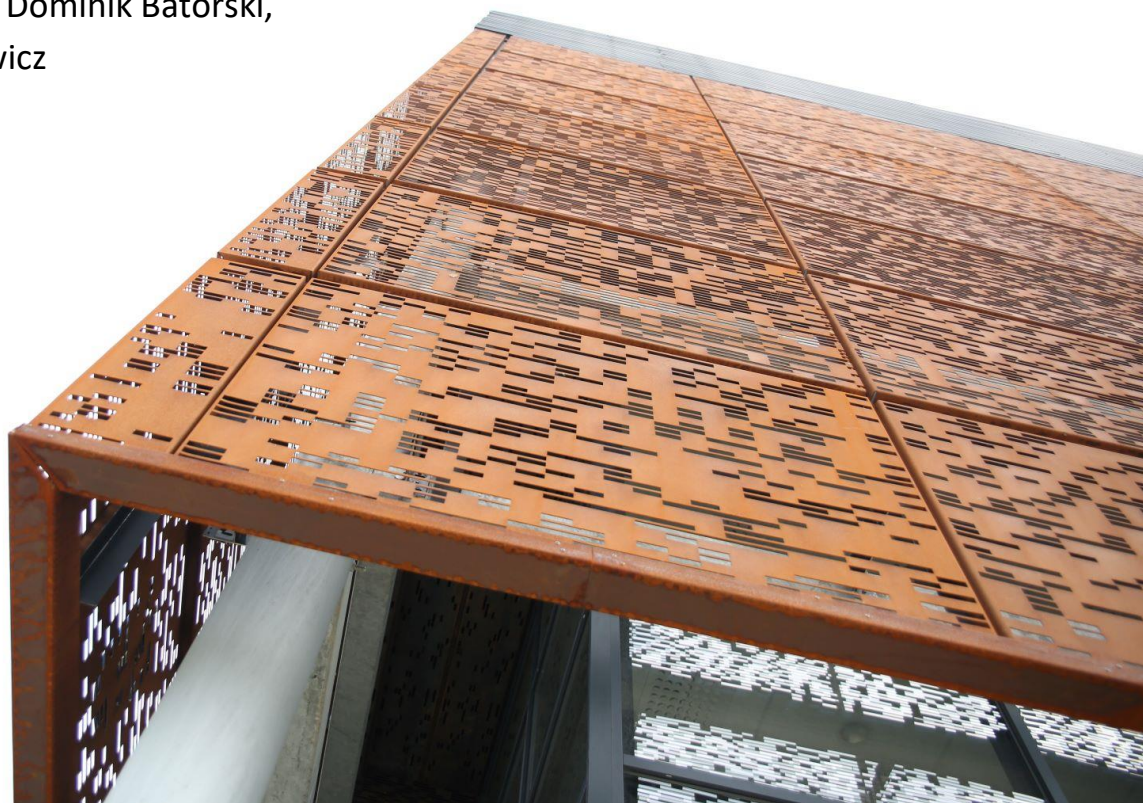




**DZIAŁALNOŚĆ INTERDYSCYPLINARNEGO
CENTRUM MODELOWANIA
MATEMATYCZNEGO I KOMPUTEROWEGO (ICM)
UNIwersYTETU WARSZAWSKIEGO
W OBSZARZE SZTUCZNEJ INTELIGENCJI I UCZENIA MASZYNOWEGO**

Potencjał ICM w realizacji Programu
“Inicjatywa doskonałości – uczelnia badawcza (2020-2026)”

Bartosz Borucki, Dominik Batorski,
Marek Michalewicz



1. KRÓTKI WSTĘP O ICM	3
2. O SZTUCZNEJ INTELIGENCJI	4
3. DZIAŁALNOŚĆ I DOROBK ICM W OBSZARZE SZTUCZNEJ INTELIGENCJI I UCZENIA MASZYNOWEGO	5
3.1 Kompetencje	5
3.1.1 Główne kompetencje w zakresie AI	5
3.1.2 Główne kompetencje narzędziowe	6
3.1.3 Dodatkowe kompetencje techniczne	6
3.1.4 Główne obszary zastosowań	6
3.2 Dorobek ICM w badaniach sztucznej inteligencji i uczenia maszynowego	6
3.2.1 Sztuczna Inteligencja w medycznej diagnostyce obrazowej	7
3.2.2. Automatyczna diagnostyka ścięgni Achillesa na podstawie obrazowania Rezonansu Magnetycznego	7
3.2.3 Spersonalizowane planowanie zabiegów korekcji deformacji palucha koślawego	7
3.2.4. Rozproszony trening sieci neuronowych	7
3.2.5. Modelowanie konektomu nicienia C. Elegans	8
3.2.6. Zastosowanie systemów przetwarzających wiedzę prawniczą	8
3.2.7. Badanie zmian w prawie przy wykorzystaniu metod obliczeniowych	8
3.2.8. Przewidywanie pętli histerezy silnie skorelowanych układów	8
3.2.9. Przewidywanie energii silnie skorelowanych układów	9
3.2.10. Predykcja cytotoksyczności Mxenes	9
3.2.11. Predykcja energii i geometrii borofenu	9
3.2.12. Znajdowanie optymalnej geometrii nanorurek boru i azotków boru	9
3.2.13. Przewidywanie przekierowania samolotów z powodu ograniczenia widzialności na lotnisku docelowym (Projekt SESAR PJ.04)	10
3.2.14. Przewidywanie/bieżąca ocena stanu nawierzchni drogi startowej (Projekt SESAR PJ.03b)	10
3.2.15. Przewidywanie zjazdu z drogi startowej z wyprzedzeniem do 5 min przed przyziemieniem (Projekt SESAR PJ.02)	10
3.2.16. Rozwój metod selekcji wszystkich istotnych cech (Boruta)	11
3.2.17. Uogólniony klasyfikator paproci losowych (rFerns)	11
3.2.18. Obliczeniowe analizy mechanizmów neurochemicznych u szczura	11
3.2.19. Uczenie maszynowe przez atomizację obserwacji	12
3.2.20. Rozwój metod opartych o teorię informacji	12
3.2.21. Implementacja metody znajdowania wielowymiarowych cech przy użyciu biblioteki ParallelX	12
3.2.22. Zrównoleglanie głębokich sieci neuronowych na superkomputerach o różnych architekturach, bez kart graficznych	13
4. Infrastruktura	13
4.1. Oprogramowanie	13
5. Podsumowanie: Co ICM wnosi do programu Wyzwania Petabajtów w ramach "Inicjatywy doskonałości – uczelnia badawcza (2020-2026)" ?	14

1. Krótki wstęp o ICM¹

Interdyscyplinarne Centrum Modelowania Matematycznego i Komputerowego (ICM UW), powstało w 1993 roku jako podstawowa, ogólnouczelniana jednostka naukowa Uniwersytetu Warszawskiego. ICM zaangażowane jest w interdyscyplinarne badania naukowe oparte na modelowaniu matematycznym, symulacjach i modelowaniu komputerowym, obliczeniach w wielu i w wielkiej skali oraz dydaktyką w powyższych obszarach. Oprócz zaangażowania w obszarze opisanym w poniższym dokumencie, pracownicy ICM zajmują się Numeryczną Prognozą Pogody wraz z publikowaniem wyników prognoz w znanym serwisie pogodowym meteo.pl który odwiedzany jest blisko 200 milionów razy rocznie. Pracownicy ICM zajmują się problemami związanymi z lotnictwem cywilnym, modelowaniem procesów społecznych i sceintometriką – opartymi na dostępie do uprzywilejowanych wielkich danych. ICM zaangażowane jest również w negocjowanie licencji i zapewnienie dostępu polskich naukowców do literatury naukowej składającej się z ponad 8.000 tytułów czasopism oraz setek tysięcy książek, poprzez utrzymywanie Wirtualnej Biblioteki Nauki; tworzeniem otwartego dostępu do literatury naukowej, oraz wielu innymi problemami.

ICM zarządza dwoma centrami danych w Warszawie. Oddane do użytku w 2016 roku Centrum Technologii ICM (CT-ICM) w Białoleńcu ma ok. 10.000 m² powierzchni technicznej w której znajdują się dwa superkomputery światowej klasy: Petaflopowy Cray XC-40 (Okeanos) do tradycyjnych intensywnych obliczeń numerycznych, oraz klaster Huawei (Enigma) do analityki wielkich danych i obliczeń chmurowych. W serwerowni CT-ICM znajduje się również sprzęt do przechowywania danych (data storage) na ok. 20 PB danych. W nowo zbudowanej auli wykładowej dla 70-ciu słuchaczy znajduje się unikalny sprzęt wizualizacyjny z 16-oma monitorami i oprogramowaniem pozwalającym na transmisję wykładów czy obrazów z całego świata.

W obszarze usług typu High Performance Computing (HPC) i usług chmurowych, ICM wspomaga ok. 1000 użytkowników z całej Polski korzystających z naszych superkomputerów i infrastruktury obliczeniowej, sieciowej oraz przechowywania wielkich danych.

ICM zatrudnia obecnie 110 pracowników.

¹ Pewne części tego dokumentu zostały przedstawione w niezmienionej formie w dwóch innych dokumentach: 1. *Dorobek Interdyscyplinarnego Centrum Modelowania Matematycznego i Komputerowego (ICM) Uniwersytetu Warszawskiego w obszarze Nauk Biologicznych i Medycznych* oraz 2. *Convergence of Big Data with High Performance Computing and Artificial Intelligence*. Powtarzamy je tutaj dla pełności prezentacji i argumentacji.

2. O Sztucznej Inteligencji

Pomimo, iż termin Sztuczna Inteligencja (*ang. Artificial Intelligence, AI*) jest dziś nadużywany w bardzo szerokim zasięgu przypadków, możemy dla uproszczenia rozumieć AI jako użycie maszyn (czy też komputerów), w celu naśladowania kognitywnych funkcji umysłu człowieka – takich jak n.p. rozpoznawanie przedmiotów, sensu tekstu, uczenie się czy rozwiązywanie problemów. W szczególności, znaczna część takich dziedzin jak uczenie maszynowe (*ang. Machine Learning, ML*), analityka dużych zbiorów danych (*ang. Big Data*), eksploracja danych (*ang. data mining*), czy bardziej ogólnie - nauka o danych (*ang. data science*), mocno zazębiają się z ogólnikowym terminem AI. Same dane natomiast stanowią bez wątpienia fundament dla większości rozwiązań AI.

ICM jako centrum superkomputerowe i ośrodek badawczy zawsze skupiał się na zagadnieniach związanych z danymi. Od problematyki technicznej, przez narzędziową, po algorytmy, jak również od uczenia maszynowego, po analitykę dużych danych, w ICM istniał znaczący dorobek w tych obszarach już na długo przed ostatnio donoszonymi sukcesami, erupcją zainteresowania i rozprzestrzenieniem się popularnego poglądu jakoby AI było remedium i wytrychem do rozwiązywania wszelkich problemów.

Kluczową rolę w rozwoju i wdrażaniu AI odgrywa synergia kilku niezbędnych komponentów: **danych, systemów ich składowania, platform**

oprogramowania oraz obliczeń. Bez dostępu do samych danych nie jest możliwe np. uczenie modeli sztucznej inteligencji. Bez odpowiedniej możliwości zarządzania dużą ilością danych i bez adekwatnej wydajności systemów składowania nie jest możliwe zasilenie AI danymi. Bez odpowiedniego oprogramowania i platform zapewniających dostęp i uruchamianie narzędzi nie jest możliwa praca z danymi i samo tworzenie modeli. Bez wysokowydajnej warstwy obliczeniowej - sprzętu i zoptymalizowanego oprogramowania - nie jest możliwa bądź realna praktyczna realizacja modeli. Dopiero powiązanie tych komponentów, wspierane odpowiednimi kompetencjami pozwala skutecznie budować Sztuczną Inteligencję.

ICM jest w stanie odpowiedzieć całościowo na to zapotrzebowanie.

Posiadając zarówno interdyscyplinarną wiedzę merytoryczną, doświadczenie, jak i adekwatne komponenty zasobów danych, oprogramowania i sprzętu mamy wyjątkowe możliwości do rozwoju AI i wyjścia z ofertą zarówno komplementarną, jak i szczegółową.

3. Działalność i dorobek ICM w obszarze Sztucznej Inteligencji i uczenia maszynowego²

3.1 Kompetencje

Kompetencje zespołów ICM zajmujących się tematyką sztucznej inteligencji znajdują się zarówno w warstwie merytorycznej badań aplikacyjnych i dziedzinowych, jak również w warstwie technicznej - narzędziowej i infrastrukturalnej.

3.1.1 Główne kompetencje w zakresie AI

- uczenie maszynowe (ML),
- klasyfikacja nadzorowana i nienadzorowana,
- analityka BigData,
- analityka small data,
- rozpoznawanie wzorców,
- inżynieria cech i wielowymiarowe metody selekcji cech,
- metody regresyjne,
- metody statystyczne,
- sieci neuronowe (płytkie),
- głębokie sieci neuronowe,
- głębokie sieci konwolucyjne (CNN),
- głębokie sieci rekurencyjne (RNN),
- metody grafowe,
- kaskadowanie klasyfikatorów,
- teoria informacji,
- przetwarzanie języka naturalnego (NLP),
- modele oparte o równania różniczkowe zwyczajne i cząstkowe,
- modele oparte o równania różniczkowo-całkowe,
- algorytmy optymalizacji,
- widzenie maszynowe (CV),
- cechy obrazowe

² Dorobek ICM w niezbędnym i dopełniającym obszarze wielkich danych (Big Data) został przedstawiony w dokumencie „Convergence of Big Data with High Performance Computing and Artificial Intelligence” ICM, 23 January 2020

3.1.2 Główne kompetencje narzędziowe

- Python,
- NumPy,
- SciPy,
- Caffé,
- TensorFlow,
- PyTorch,
- Keras,
- Anaconda,
- MLflow,
- Scikit-learn,
- Pandas,
- R,
- Java,
- C/C++,
- CUDA,
- Julia,
- Statistica,
- Horovod,
- Gensim,
- Nltk,
- Mallet,
- Snowball stemmer,
- Polimorfologik,
- SPARQL,
- Matlab,
- FEniCS
- i inne.

3.1.3 Dodatkowe kompetencje techniczne

- programowanie równoległe,
- programowanie GPU,
- edge computing.

3.1.4 Główne obszary zastosowań

- analiza obrazowa,
- obrazowanie medyczne,
- diagnostyka ruchowa i śledzenie ruchu,
- biologia,
- sieci,
- genomika,
- bioinformatyka,
- prawo,
- pogoda i klimat
- awiacja
- fizyka.

3.2 Dorobek ICM w badaniach sztucznej inteligencji i uczenia maszynowego

Poniższe zestawienie prezentuje dorobek poszczególnych grup oraz pracowników ICM w zakresie tematyki sztucznej inteligencji, uczenia maszynowego oraz analityki danych. Jest to tematyczne zestawienie zadań, będących zarówno projektami B+R finansowanymi ze źródeł zewnętrznych, jak i projektów realizowanych wewnętrznie i we współpracy zewnętrznej z udziałem ICM.

3.2.1 Sztuczna Inteligencja w medycznej diagnostyce obrazowej

Opracowywanie i wykorzystanie technik głębokiego uczenia się (DL) oraz uczenia maszynowego (ML) w problemach związanych z obrazowaniem medycznym i radiologią. Od poprawy jakości obrazów, poprzez segmentację, detekcję, automatyzację diagnostyki, wspomaganie decyzji, aż po strukturyzację raportowania, optymalizację pracy radiologicznej (np. triage) i automatyzację analizy i generowania opisów badań. Głównie w zakresie MRI, CT i USG, w większości obszarów anatomicznych. Obecne główne obszary zainteresowania: ortopedia / kończyna dolna, klatka piersiowa / oraz płuca. Zespół realizujący: B.Borucki, N.Kapiński, J.Zieliński, J.Nowosielski, K.Nowiński (ICM UW)

3.2.2. Automatyczna diagnostyka ścięgna Achillesa na podstawie obrazowania Rezonansu Magnetycznego

Dzięki wykorzystaniu konwolucyjnych sieci neuronowych (CNN) oraz metod uczenia maszynowego i statystyki, opracowane zostało narzędzie pozwalające zautomatyzować ocenę badania rezonansu magnetycznego (MR) ścięgna Achillesa. Na podstawie danych obrazowych generowane są wskaźniki numeryczne opisujące ilościowo stan poszczególnych ocenianych cech obrazowanych tkanek. Lekarze mogą wykorzystać punktową informację diagnostyczną, a także monitorować procesy gojenia i obserwować zachowanie procesu na tle statystyk.

Zespół realizujący: B.Borucki, N.Kapiński, J.Zieliński, J.Nowosielski, K.Nowiński, P.Regulski (ICM UW)

3.2.3 Spersonalizowane planowanie zabiegów korekcji deformacji palucha koślawego

Wykorzystaliśmy klasyczne metody uczenia maszynowego do analizy przestrzeni cech opisujących geometrię układu kostnego stopy. W efekcie opracowane zostały metody pozwalające nie tylko automatycznie ocenić parametry deformacji geometrycznych, ale też metody diagnostyczne i planowania zabiegów operacyjnych. Opracowane wzorce deformacji oraz analizy podtypów zabiegów chirurgicznych, wraz z klasyfikacją pozwalają lepiej dobrać metodę operacyjną w personalizacji leczenia.

Zespół realizujący: B.Borucki, M.Łyczek, K.Nowiński (ICM UW)

3.2.4. Rozproszony trening sieci neuronowych

Celem prac było wykorzystanie biblioteki PCJ (implementującej model programowania PGAS w maszynie wirtualnej Javy) do zrównoleglenia obliczeń związanych z uczeniem sieci neuronowych. W toku badań

zaimplementowano rozproszony synchroniczny i asynchroniczny algorytm obliczenia gradientu zstępującego (bibliotekę PCJ wykorzystując do implementacji komunikacji, obliczenia wewnątrz węzłów delegując do Tensorflow) oraz zbadano wydajność rozwiązania - porównując je z tandemem: Horovod+Tensorflow.

Realizacja: P. Bała, Ł. Górski (ICM UW), M. Nowicki (WMiI UMK)

3.2.5. Modelowanie konektomu nicienia C. Elegans

W ramach prac zaimplementowano szkielet aplikacji (framework) odpowiedzialny za uruchomienie rozproszonego algorytmu ewolucyjnego (ewolucji różnicowej) z wykorzystaniem biblioteki PCJ (w modelu PGAS). Szkielet ten został następnie wykorzystany do przeprowadzenia obliczeń - optymalizacji parametrów modelu konektomu nicienia C. Elegans - predykcje tego modelu odpowiadały danym eksperymentalnym.

Realizacja: P. Bała, F. Rakowski (ICM UW), Ł. Górski (WMiI UMK / ICM UW),

3.2.6. Zastosowanie systemów przetwarzających wiedzę prawniczą

W ramach prac brano udział w rozwoju systemu komputerowego opartego o wiedzę (knowledge-base system), w założeniu mającego wspierać urzędników patentowych w ich rutynowych obowiązkach. W toku prac opracowano nową metodę wyszukiwania najbardziej relewantnych dla danego przepisu prawa orzeczeń (z wykorzystaniem mechanizmów modelowania tematycznego - pakiet Mallet) oraz opracowano metodologię służącą do weryfikacji i walidacji niektórych aspektów ontologii przechowującej wiedzę prawniczą.

Realizacja: S. Ramakrishna, A. Paschke (FU Berlin), Ł. Górski (WPiA / WMiI UMK)

3.2.7. Badanie zmian w prawie przy wykorzystaniu metod obliczeniowych

Celem prac jest zbadanie możliwości wykorzystania metod obliczeniowych do zbadania zmian w prawie. W ramach do tej pory prowadzonych prac badawczych dokonano klasteryzacji kolejnych nowelizacji kodeksu cywilnego, wykorzystując przy tym różne metody ich reprezentacji (doc2vec, word2vec). Obecnie rozpoczęto badania nad możliwością detekcji różnych linii orzeczniczych na podstawie danych zgromadzonych w bazach judykatów.

Realizacja: Ł. Górski (ICM UW)

3.2.8. Przewidywanie pętli histerezy silnie skorelowanych układów

Celem projektu było zbudowanie modelu opartego na algorytmach uczenia maszynowego w celu modelowania kształtu pętli histerezy i własności

magnetycznych układów o silnych oddziaływaniach długo-zasięgowych. Testowano dwa podejścia RNN i SLTM w celu modelowania procesu magnetyzacji w układzie oraz klasycznych metod uczenia maszynowego do wyznaczenia punktów charakterystycznych pętli. Dane pochodziły z symulacji MCMC.

Realizacja: M.Marchwiany (ICM UW), A.Majhofer (Wydział Fizyki UW)

3.2.9. Przewidywanie energii silnie skorelowanych układów

Celem projektu było zbudowanie modelu predykcji energii silnie skorelowanych układów w celu wykorzystania do akceleracji symulacji MCMC. Dane pochodziły z symulacji MCMC.

Realizacja: M.Marchwiany (ICM UW), A.Majhofer (Wydział Fizyki UW)

3.2.10. Predykcja cytotoksyczności Mxenes

Celem projektu było opisanie cytotoksyczności Mxenes za pomocą uczenia maszynowego. Pracowano na dwóch zbiorach danych: eksperymentalnym (zmierzone własności próbek) i teoretycznym (zmierzona cytotoksyczność dla związku + teoretycznie policzona geometria). Zbudowany model ma dobrą zdolność predykcyjną na zbiorze doświadczalnym. Dzięki feature selection powstała hipoteza o przyczynach cytotoksyczności Mxenes.

Realizacja: M.Marchwiany (ICM UW), M.Popielska(Birowska), J.Majewski, A.Jastrzębska (Wydział Fizyki UW)

3.2.11. Predykcja energii i geometrii borofenu

Celem projektu było znalezienie zależności pomiędzy defektami sieci, a energią i niepłaskością borofenu. Dane pochodziły z symulacji.

Realizatorzy: M.Marchwiany (ICM UW), N.Gonzalez-Szwacki, T.Tarkowski (Wydział Fizyki UW)

3.2.12. Znajdowanie optymalnej geometrii nanorurek boru i azotków boru

Celem projektu jest znalezienie optymalnej geometrii nanorurek boru. Zbudowany zostanie model AI przewidujący energię na podstawie geometrii. Zostanie on użyty jako szybka metoda szacowania energii podczas procesu optymalizacji geometrii. Dane pochodzą z symulacji.

Realizatorzy: M.Marchwiany (ICM UW), N.Gonzalez-Szwacki (Wydział Fizyki UW)

3.2.13. Przewidywanie przekierowania samolotów z powodu ograniczenia widzialności na lotnisku docelowym (Projekt SESAR PJ.04)

Narzędzie, przewidujące ryzyko przekierowania samolotu na lotnisko zapasowe, w przypadku ograniczonego zasięgu widzialności wzdłuż drogi startowej na lotnisku docelowym, zostało opracowane przy wykorzystaniu konwolucyjnych sieci neuronowych (CNN), danych o planach lotów, minimach statków powietrznych i ich załóg, a także danych pogodowych z lotniskowych stacji pogodowych i prognoz pogody. Dzięki alertom o przekierowaniu danego statku powietrznego, generowanym z godzinnym wyprzedzeniem, kontrolerzy ruchu lotniczego dysponują zapasem czasu na przygotowanie nowej trasy przelotu, a lotniska i linie lotnicze na wdrożenie planów awaryjnych z odpowiednim wyprzedzeniem.

Realizacja: K. Rutkowski, K. Piwoński (ICM UW)

3.2.14. Przewidywanie/bieżąca ocena stanu nawierzchni drogi startowej (Projekt SESAR PJ.03b)

Ocena stanu nawierzchni drogi startowej bez konieczności prowadzenia inspekcji (a tym samym zajmowania drogi startowej) została opracowana przy użyciu wielu dostępnych źródeł danych takich jak: parametry pogodowe AWOS, dane radarowe, dane z czujników systemu Ice Alert. U podstaw rozwiązania leży system oparty o uczenia maszynowe (Gradient Boosted Trees). W zastosowaniu operacyjnym przewidywane korzyści z użycia systemu to głównie podniesienie poziomu bezpieczeństwa poprzez dostęp do bardziej spójnej i aktualnej informacji o drodze startowej oraz możliwość lepszego planowania operacji utrzymania. System jest zgodny z nowym standardem ICAO GRF.

Realizacja: W. Rosiński, M. Hallay-Suszek, K. Rutkowski, J. Kopeć (ICM UW)

3.2.15. Przewidywanie zjazdu z drogi startowej z wyprzedzeniem do 5 min przed przyziemieniem (Projekt SESAR PJ.02)

W oparciu o techniki Random Forest oraz Gradient Boosted Trees opracowano narzędzie oceniające wybór zjazdu z drogi startowej przez statki powietrzne w końcowej fazie zbliżania. Narzędzie to wykorzystuje różne klasy danych, między innymi bieżące dane pogodowe z lotniska oraz dane radarowe o trajektorii statku powietrznego. Uzyskane wyniki są w każdym wypadku istotnie lepsze od obecnie używanych metod opartych o klasy wagowe. Poprawiona znajomość zjazdu może posłużyć do poprawy systemów AMAN/DMAN czy coupled AMAN/DMAN oraz wszędzie tam gdzie istotna jest znajomość Arrival Runway Occupancy Time.

Realizacja: K. Rutkowski, W. Rosiński, J. Kopeć (ICM UW)

3.2.16. Rozwój metod selekcji wszystkich istotnych cech (Boruta)

Od pierwszego wydania w 2010 roku, metoda jest badana i rozwijana; wydano 6 wersji wzorcowej implementacji (<https://CRAN.R-project.org/package=Boruta>), ulepszając dokładność i szybkość działania pierwotnego algorytmu, jak również rozbudowując go o inne źródła istotności zmiennych. W 2014 w artykule "Robustness of Random Forest-based gene selection methods" porównano algorytm Boruta z innymi metodami wykorzystując pionierską analizę stabilności selekcji cech. Realizacja: M. Kursa (ICM UW)

3.2.17. Uogólniony klasyfikator paproci losowych (rFerns)

W 2012 ukazał się artykuł "A Comparison of Random Forests and Ferns on Recognition of Instruments in Jazz Recordings" wprowadzający uogólnienie znanej z obliczeniowej analizy obrazu metody paproci losowych, zdolne do pracy z arbitralnymi zbiorami danych, w kontekście analizy danych audio. W 2014 ukazał się ogólniejszy artykuł "rFerns: An implementation of the random ferns method for general-purpose machine learning" w którym przedstawiono dualistyczną interpretację tego algorytmu jako pomostu między metodami stochastycznymi i Bayesowskimi. Zaproponowano też wykorzystanie klasyfikatora jako wydajnego i nieobciążonego źródła ważności zmiennych, zarówno z myślą o metodzie Boruta jak i niezależnych zastosowaniach; badania w tym kierunku zaowocowały prototypową metodą "naive-wrapper", która została opisana w artykule "Efficient all relevant feature selection with Random Ferns" i jest dalej rozwijana.

Zaproponowano także wersję algorytmu dedykowanego do klasyfikacji wieloetykietowej (multilabel), opisaną w artykule "Multi-label Ferns for Efficient Recognition of Musical Instruments in Recordings"; wydaje się jednak że rozszerzenie to nie jest konkurencyjne poza pewnymi specyficznymi danymi wejściowymi.

Zbudowano także bardzo wydajną implementację metody, dostępną jako pakiet R rFerns (<https://CRAN.R-project.org/package=rFerns>). Implementacja ta dobrze wykorzystuje maszyny wielordzeniowe, wykazano dobrą skalowalność do 96 rdzeni.

Realizacja: M. Kursa (ICM UW), A. Wieczorkowska (PJATK, badania audio).

3.2.18. Obliczeniowe analizy mechanizmów neurochemicznych u szczura

Od 2017 prowadzone są prace nad zastosowaniem nowoczesnych metod statystycznych do analizy danych z przekrojowych pomiarów wielostrukturalnych, ze szczególnym uwzględnieniem prób rekonstrukcji sieci zależności.

Ich dotychczasowym owocem są prace “Non-parametric analysis of neurochemical effects and Arc expression in amphetamine-induced 50-kHz ultrasonic vocalization” i “Inter-individual differences in serotonin and glutamate co-transmission reflect differentiation in context-induced conditioned 50-kHz USVs response after morphine withdrawal”; kolejna w przygotowaniu, trwają też prace eksperymentalne rozwijające dotychczasowe ustalenia.

Realizacja: M. Kursa (ICM UW), A.Hamed (IBD PAN).

3.2.19. Uczenie maszynowe przez atomizację obserwacji

Od konkursu FlowCAP II trwają prace nad analizą danych z eksperymentów cytometrii przepływowej. W wyniku tych prac powstała metoda FlowForest (<https://github.com/flowforest/flowforest>), opisana w artykule “Fully automatic classification of flow cytometry data”. Została ona wykorzystana przy udziale w konkursach, zrelacjonowanych w artykułach “Critical assessment of automated flow cytometry data analysis techniques” i “A benchmark for evaluation of algorithms for identification of cellular correlates of clinical outcomes”. Obecnie trwają prace nad generalizacją tego podejścia na dane z klasy małej liczby złożonych obserwacji, kiedy to wejściem do analizy są obiekty o dużej zawartości informacji, ale jest ich za mało by wytworzyć użyteczną abstrakcję tradycyjnymi podejściami; w szczególności, przeprowadzono owocne badania pilotażowe na danych elektroencefalograficznych.

Realizacja: M. Kursa, B. Piotrowski (ICM UW), S. Walkowiak (udział w konkursie FlowCAP IV).

3.2.20. Rozwój metod opartych o teorię informacji

Od 2018 trwają prace nad pakietem praznik, kolekcją wydajnych implementacji popularnych metod selekcji i oceny zmiennych opartych o teorię informacji (<https://CRAN.R-project.org/package=praznik>). Kod ten zyskał uznanie w środowisku za stabilność, wydajność i skalowalność. Artykuł opisujący pakiet znajduje się obecnie w recenzji.

Realizacja: M. Kursa (ICM UW).

3.2.21. Implementacja metody znajdowania wielowymiarowych cech przy użyciu biblioteki ParallelX

Student *Inżynierii Obliczeniowej* w ICM, Karol Niedzialewski w swojej pracy magisterskiej z 2019 roku zaimplementował bibliotekę ParallelX do zrównoleglania kodów na komputerach wielkiej mocy obliczeniowej w kontekście wykrywania cech wielowymiarowych w zbiorach genomicznych. Jego praca “Multidimensional Feature Selection and High Performance

ParalleX” która ukazała się w Springer Nature Computer Science 24-go października 2019 była już ściągana 1244 razy z portalu wydawcy Springer Nature. Równoległe obliczenia przeprowadzone przez Karola Niedzielewskiego na 6,144 rdzeniach obliczeniowych superkomputera Okeanos przyspieszyły program w wersji serialnej z 6.5 godzin do 3 minut w wersji równoległej.

Realizacja: K. Niedzielewski pod kierunkiem M. Michalewicza

3.2.22. Zrównoleganie głębokich sieci neuronowych na superkomputerach o różnych architekturach, bez kart graficznych

Zbigniew Jankowski, student *Inżynierii Obliczeniowej* w ICM w swojej pracy magisterskiej *“Skalowalność głębokich sieci neuronowych na superkomputerach o różnych architekturach”* pokazał, że framework CaffeNet można z wielkim powodzeniem zaimplementować na superkomputerze Ocean (bez kart graficznych) oraz na 32 węzłach komputera Stratus w A*STAR w Singapurze, który składa się wyłącznie z procesorów Knights Landing, każdy o 64 rdzeniach obliczeniowych.

Realizacja: Z. Jankowski pod kierunkiem M. Michalewicza

4. Infrastruktura

Unikalna infrastruktura obliczeniowa ICM została opisana w szczegółach w dwóch dokumentach:

1. *Dorobek Interdyscyplinarnego Centrum Modelowania Matematycznego i Komputerowego (ICM) Uniwersytetu Warszawskiego w obszarze Nauk Biologicznych i Medycznych*, 9 styczeń 2020; oraz
2. *Convergence of Big Data with High Performance Computing and Artificial Intelligence*, 23 styczeń 2020.

4.1. Oprogramowanie

Na komputerach ICM zainstalowane jest oprogramowanie i biblioteki potrzebne do uczenia maszynowego. Listę oprogramowania zawarto powyżej w paragrafie 3.1.2 *Główne kompetencje narzędziowe*

Zasoby te są rozszerzane o unikalne w skali światowej rozwiązania dedykowane do konkretnych platform obliczeniowych m.in.:

- Urica XC - oprogramowanie do analizy wielkich danych (ang. Big Data), przetwarzania grafów i uczenia głębokich sieci neuronowych (ang. Deep Learning) na superkomputerze Cray XC-40,
- Biblioteka SOL - biblioteka do uczenia głębokich sieci neuronowych (ang. Deep Learning) na ośmiu kartach do obliczeń wektorowych na unikalnym komputerze wektorowym NEC Aurora Tsubasa,
- Oprogramowanie Trovares do obliczeń w obszarze tzw. grafów cech (property albo attribute graphs)

ICM będzie przeprowadził w najbliższym czasie szkolenia z powyższych rozwiązań dla użytkowników³.

5. Podsumowanie: Co ICM wnosi do programu Wyzwania Petabajtów w ramach „Inicjatywy doskonałości – uczelnia badawcza (2020-2026)” ?

ICM od lat zaangażowane jest w badania naukowe i rozwojowe przy zastosowaniu metod sztucznej inteligencji, uczenia maszynowego i złożonych metod analizy danych.

ICM posiada bogate doświadczenie we współpracy w zakresie rozwiązań technologicznych do realizacji zadań związanych ze sztuczną inteligencją, a w szczególności integracji rozwiązań w warstwie danych, obliczeń i analityki.

ICM:

- realizuje ekspertyzy i konsultacje w zakresie technologii i obszarów opisanych w rozdziale „Kompetencje”.
- udostępnia na elastycznych zasadach infrastrukturę danych i obliczeń oraz przestrzeń centrum danych, a także infrastrukturę analityki i wizualizacji danych.
- realizuje projekty B+R w skali krajowej, europejskiej i międzynarodowej.

³ Oprogramowanie Urica XC, SOL, język Julia, i inne odpowiednie tematy będą przedmiotem specjalnych wykładów i warsztatów dla użytkowników w ramach nadchodzącej konferencji „Supercomputing Frontiers Europe 2020”, 23-26 marzec 2020 <https://supercomputingfrontiers.eu/2020/> . O SOL i programowaniu komputera wektorowego można się będzie dowiedzieć na konferencji NEC Users Group Meeting w maju tego roku w Warszawie (współorganizowanego przez ICM).