Prof. dr hab. Włodzisław Duch
Katedra Informatyki Stosowanej
i Laboratorium Neurokognitywne
Uniwersytet Mikołaja Kopernika, Toruń

Recenzja rozprawy doktorskiej mgr Łukasza Kwaśniewicza, „Metoda pomiaru wiarygodności wiadomości wykorzystująca elektroencefalografię ilościową i sztuczną inteligencję".

Recenzja opracowana została na prośbę Rady Naukowej Dyscypliny Informatyka Polsko-Japońskiej Akademii Technik Komputerowych w Warszawie.

Mgr Łukasz Kwaśniewicz przygotowała rozprawę doktorską na Wydziale Informatyki Polsko-Japońskiej Akademii Technik Komputerowych w Warszawie, pracując pod opieką dr hab. Grzegorza Wójcika, profesora UMCS i PJATK. Rozprawa liczy 126 stron, składa się z 5 rozdziałów i dwóch dodatków. Jej głównym tematem jest badanie zachodzących w mózgu procesów podczas oceny zaufania do otrzymanych wiadomości przy podejmowaniu decyzji. Zalew fałszywych wiadomości, ich bezkrytyczne powielanie w sieciach społecznościowych, oraz próby wyłudzenia informacji i pieniędzy są obecnie ogromnym problemem. Stąd potrzeba zbadania, czy pomiar aktywacji różnych obszarów mózgu pozwoli na przewidywanie ocen wiarygodności otrzymanych wiadomości. Takie badania prowadzono przy użyciu funkcjonalnego rezonansu już ponad dekadę temu, jednakże fMRI to technologia droga i mało praktyczna. Badania przy użyciu elektroencefalografii (EEG) opisane w tej pracy prowadzone są od kilku lat w grupie prof. Wójcika i należy je uznać za pionierskie.

Próba analizy stanów mózgu związanych z subiektywną oceną wiarygodności otrzymanej informacji wymaga zaprojektowania ściśle kontrolowanych warunków dla eksperymentu, wykonania badań EEG, użycia szeregu zaawansowanych technologii analizy sygnału a na końcu zastosowania metod klasyfikacji do oceny skuteczności przewidywań opracowanego modelu. Praca jest więc wysoce interdyscyplinarna, sytuuje się na pograniczu psychologii społecznej i poznawczej, neurobiologii i informatyki. Niewiele jest jeszcze prac stawiających sobie równie ambitne cele, badających tak złożone procesy.

Po ogólnym opisie celów pracy i podejścia do problemu, w rozdziale drugim dyskutowane jest pojęcie wiarygodności, relacje pomiędzy prawdziwością i wiarygodnością informacji. Jest

wiele przykładów pozornie wiarygodnych informacji, które okazały się prawdziwe, są też twierdzenia nierozstrzygalne na obecnym etapie wiedzy, a także twierdzenia matematyczne, które nie są rozstrzygalne w obrębie przyjętej aksjomatyki. Autor wyróżnił trzy rodzaje prawdy: postmodernistyczną, naukową i semantyczną. Znacznie więcej miejsca (7 stron) poświęcił pojęciu wiarygodności. Te rozważania są mu potrzebne do oceny różnych wymiarów wiarygodności wiadomości: struktury, zawartości, języka i sposobu dostarczenia lub prezentacji wiadomości. Każdy z tych wymiarów ma kilka aspektów, które go charakteryzują. Pomimo kilkudziesięciu lat badań nie mamy jednoznacznego sposobu charakteryzacji wiarygodnych informacji. W badaniach najczęściej stosuje się podejście ankietowe lub eksperymenty psychologiczne. Trwają też próby stworzenia narzędzi informatycznych do automatycznej oceny wiarygodności informacji. Opisane w tej rozprawie eksperymenty nie wykorzystywały takich narzędzi – to by wprowadziło dodatkowy poziom komplikacji, oceny wiarygodności narzędzi do oceny wiarygodności wiadomości.

Rozdział trzeci opisuje elektroencefalograficzne metody badania mózgu, podstawowe wiadomości na temat neuronów i ogólnej budowy mózgu oraz problemy z artefaktami związanymi z fizjologią, bodźcami z otoczenia i technicznymi aspektami pomiarów EEG. Nie zawsze opisy są tu ścisłe. Podział na pasma częstotliwości EEG jest wygodny, ale jest sprawą umowną. Stwierdzenie „Zaobserwowano, iż w mózgu człowieka występuje 5 rodzajów fal mózgowych, których podział związany jest z zakresem ich częstotliwości ..." robi wrażenie, jakby to były oddzielne zjawiska. Beta jest od 14-26 Hz, a gamma od 30 do 100 Hz – nie ma fal o częstotliwości 26 do 30 Hz? Fale μ mają takie częstotliwości jak fale alfa, ale nazywamy je inaczej ze względu na lokalizację źródeł. Są wynikiem aktywności neuronów kory ruchowej. Stwierdzenie, że są obecne „nad korą czuciowo – ruchową" jest co najmniej niezręczne. Rys. 3.4 przedstawia schematyczny potencjał czynnościowy, ale jego źródłem jest Wikipedia a nie ref. [3], a prawdziwy obserwowany potencjał odbiega od tego wzorca, jak pokazuje to artykuł w Wiki. Nie bardzo wiem, co ma oznaczać przetwarzanie sygnałów „w czasie ciągłym, jednakże przy początkowym przekształceniu ich w sygnały czasu dyskretnego".

W podrozdziale 3.9 opisano problemy z pomiarami potencjałów wywołanych ERP, głównie na podstawie książki „Applied Event-Related Potential Data Analysis" Stevena Lucka (darmowa zaktualizowana wersja tej książki jest dostępna w LibreText) i książkę zredagowaną przez N. Kamela i A.S. Malika „EEG/ERP analysis: methods and applications". Byłoby lepiej odwołać się do rozdziału Z. Jianga, a nie całej książki. Zamieszczone tłumaczenie fragmentu

tego rozdziału jest mylące: uśrednienie dla jednego komponentu ERP to nie „co do jednego komponentu". Użycie wyrażenia „liczenie średniej z blokadą odpowiedzi" nie oddaje sensu wyrażenia „responsed-locked averages", chodzi o liczenie średniej, zsynchronizowanej z czasem reakcji. „Blokowanie" używane jest często w sensie „synchronizacji", np. fazy, czy czasu. „Grand average ERP waveform" zostało przetłumaczona jako „średnia między przedmiotami (ang. Great Average)". Niestety podobne niezręczności są w wielu miejscach. Lepiej było napisać pracę doktorską po angielsku niż męczyć się z przekładaniem na język polski.

W dalszej części rozdziału jest przegląd modeli lokalizacji źródeł i rozwiązań problemu odwrotnego. Tą część niełatwo się czyta, Autor kilka razy nawraca do wyjaśnienia, że jest problem prosty i odwrotny, odwołuje się do liniowego równania (3.10), pisząc, że jest to problem nieliniowy, ale nie wyjaśniając dlaczego. Wzmianki o zastosowaniach pomieszane są z opisem szeregu metod rekonstrukcji i lokalizacji źródeł. Jest wiele rodzajów szumu neuronalnego, ale określenie „szum biologiczny" obejmującego wszystkie z nich pojawia się chyba tylko w jednym artykule (Jatoi i inn., rozdział w książce pod red. Kamel, Malik, 2015).

Rozdział czwarty opisuje projekty własnych eksperymentów. Kandydat rozpisał się tu na temat przydatności badań oceny wiarygodności, stawiając sobie za cel określenie aktywowanych obszarów mózgu, zbadania czynników, które sprzyjają ich aktywacji, oraz korelacji takich ocen z aktywnością mierzoną za pomocą EEG. Wcześniej badania kontrastu pomiędzy zaufaniem, niepewnością i nieufnością za pomocą rezonansu funkcjonalnego pokazały zmiany aktywności w wielu obszarach kory przedczołowej, czołowej, skroniowej i kory zakrętu obręczy. W badaniu Pameli Douglas i inn. (nie Douglasa) z 2013 roku wykorzystano zarówno fMRI, jak i 256-kanałowe EEG. Analiza falkowa i wykorzystanie składowych ICA pozwoliło na stworzenie prostego modelu przewidywania ocen wiarygodności. Ta praca nie spowodował jednak większego zainteresowania trudną tematyką oceny wiarygodności. Inne badania omówione w tym rozdziale, wykonane przy użyciu fMRI i PET, pokazały aktywację w okolicach bieguna skroniowego górnego, zakrętu wrzecionowatego i paru innych struktur, których aktywność związana jest z rozumieniem języka i analizą emocji związaną z wypowiedziami. Opis struktur mózgu, aktywowanych w tych eksperymentach, miesza nazwy angielskie z polskimi, np. „... calcarine - gdzie pierwotna kora wzrokowa jest skoncentrowana (BA17)". Calcarine sulcus to bruzda ostrogowa, czyli zakręt potyliczno-skroniowy przyśrodkowy. Nie można powiedzieć, że kora wzrokowa jest skoncentrowana w jednym obszarze Brodmanna. Pre-SMA to przeddodatkowe pole ruchowe, a nie nadrzędny obszar motoryczny. Sam udo-

stępniam w sieci słowniczek neuroanatomiczny, bo tłumaczenia takich nazw zawsze sprawiają kłopoty.

Podrozdział 4.3 opisuje pierwszy projekt własnego eksperymentu. Pojęcie „projektu wiadomości", które ma być podstawą do podejmowanie decyzji, nie zostało zdefiniowane. Pilotażowy eksperyment polega na przedstawieniu japońskich znaków ideograficznych i propozycjach ich tłumaczeń. Zadaniem badanej osoby jest podjąć decyzję, czy podano poprawne znaczenie. Uczestnicy nie znali japońskiego, więc ich decyzje były czysto intuicyjne, podejmowane na podstawie krótkiej lub dłuższej wiadomości komentującej dany znak.

Z pomiarów EEG wyliczany jest średni ładunek w danym obszarze (MEC). Procedura obliczania MEC nie została jednak dokładniej opisana, a w oryginalnych pracach (ref. 154-157) również brakuje informacji, jak obliczane są prądy i ładunki z całych obszarów Brodmanna. Mamy wysokiej jakości sygnał z 256 elektrod, ale w rozprawie ani w publikacjach nie znalazłem informacji o liczbie źródeł, ani ich umiejscowieniu. Czy wybrano jedno źródło na całe pole Brodmanna, niezależnie od jego wielkości? Ajk te sygnały były uśredniane? Cała analiza opiera się na ocenie aktywności pól Brodmanna mierzonej za pomocą ładunku i prądów, ale same równania nie wyjaśniają jakiej użytej procedury.

W drugim eksperymencie badana jest różnica pomiędzy oceną wiarygodności wiadomości przy braku lub posiadaniu pewnej wiedzy pozwalającej podjąć decyzję. Jedynym czynnikiem różnicującym zadania była długość wiadomości. Po nauczeniu się znaczenia 3 prostych znaków kanji uczestnicy mieli za zadanie odpowiedzieć na pytania, czy tłumaczenie danego znaku jest poprawne, mieli więc pełną wiedzę. W drugiej grupie 80 pytań były nowe znaki kanji i krótkie opisy, a w trzeciej grupie nowe znaki i długie opisy, sugerujące właściwy wybór odpowiedzi. Tym razem postawiono aż 6 hipotez i omówiono sposoby ich weryfikacji. Szósta hipoteza to związek decyzji z aktywnością lewych pól BA8 i BA9 w płatach czołowych. Problem w tym, że takie obszary są zaangażowane w kilkanaście różnych funkcji związanych z tym zdaniem: ocenę niepewności, pamięć krótkotrwałą, automatyczne reakcje, wykrywanie błędów, uwagę werbalną, wnioskowanie o intencjach innych osób, wnioskowanie dedukcyjne z wyobrażeń przestrzennych, rozumowanie indukcyjne, przypisywanie intencji. Nie jest więc jasne, o czym właściwie świadczy aktywność tych obszarów w kontekście projektowanych eksperymentów. Statystyczne różnice w ocenie wielkości ładunku MES zaobserwowano w pilotażowym eksperymencie w 18 polach Brodmanna. Jak duże były to różnice? W rozprawie czytamy, że podzbiór tych pól, wybrany przez model oceny wiarygodności, jest zaprezento-

wany „w poniższej tabeli", ale ta tabela jest 14 stron dalej. W dodatku cytowany „Handbook of Transcranial Magnetic Stimulation" zawiera tylko jedną wzmiankę o polach Brodmanna i trudno to połączyć z prezentowaną w pracy tabelą. Z uwagi na stronie 70 wynika, że ładunek elektryczny MEC przepływający przez pola Brodmanna zlokalizowane pod elektrodami na korze mózgowej w CPTR obliczany jest na podstawie ERP, o czym wcześniej nie wspomniano. Nie wiadomo też, co oznacza CPTR, bo ani w pracy, ani w publikacji [156], ani szukając w Internecie, takiego skrótu nie znalazłem.

Nie bardzo rozumiem czemu użycie MEC ze wszystkich pól Brodmanna we wszystkich przedziałach czasowych do zdefiniowania niezależnych zmiennych miałoby prowadzić do przetrenowania modelu. To zależy od klasyfikatora i sposobu jego uczenia, a nie wielkości danych treningowych. Problemem jest raczej rozważanie bardzo wielkiej liczby modeli, które prowadzi do znalezienia przypadkowej zgodności modeli. Na podstawie różnic wartości MEC dla pól Brodmanna w przypadku decyzji zgodnej lub odrzucającej prawdziwość wiadomości, wybrano 5 pól dla każdego rozważanego przedziału czasowego. Jeśli obliczamy zmiany ładunku w czasie, to powinno się dać wyznaczyć początek reakcji na prezentowane bodźce. Tymczasem rozważane są wszystkie przedziały czasowe. Wyliczenia liczby przedziałów czasowych dla przedziałów 25 ms, w okienku od 0 do 900 milisekund dały 194 interwały do sprawdzenia, a potem 2261 przedziałów czasowych – nie wiem, skąd Autor wziął takie liczby. Użycie podzbiorów dla 5 wybranych pól Brodmanna i 2261 przedziałów daje 70091 modeli. Selekcja najlepszych modeli z takiej liczby musi prowadzić do przypadkowych wyników. Selekcja modeli to duży dział statystyki, o którym praca nie wspomina. Nie podano żadnych informacji o sposobie zastosowania procedury bootstrap do wyboru modelu. Wyniki z Tab. 4.2 są na poziomie 79% dokładności, ale testy 10-krotnej kroswalidacji pokazały słabsze wyniki, na poziomie 68%. W tym przypadku używamy 90% danych do treningu, a nie 75%, a dokładność mocno spada. To wskazuje na zbyt optymistyczne oceny związane z podziałem danych na zbiór treningowy i walidacyjny. Zastosowanie nowego testu F Causeura do testowania istotności efektów było dobrym pomysłem, ale mamy tylko opis teorii, a nie wyników jej zastosowania.

Rysunki 4.9-4.12 przedstawiają uśrednione wartości ERP dla pojedynczych elektrod. Jak były obliczane na poziomie elektrod dla całych obszarów Brodmanna? Tylko dla elektrody 118, dla krótkich notek, widać wyraźniejsze różnice między zaufaniem i brakiem zaufania. Spodziewałem się podobnych wykresów MEC i porównania wyników klasyfikacji przy użyciu

ERP i MEC, ale takich wyników nie ma. Dla różnych przedziałów czasowych 5 wybranych pól się różni. W sumie w modelach użytych do klasyfikacji obszarów wykazujących istotne różnice było 18 pól, a w najlepszych klasyfikatorach było tylko 7.

W tekście pojawia się uwaga, że różnice aktywności pola BA29 przyczynia się do weryfikacji hipotezy 4, czyli znaczącej różnicy modeli przewidujących decyzje uczestników, którzy często wybierają długie wiadomości w porównaniu do modeli innych uczestników. Pole Brodmana 29 nie pojawia się jednak w Tab. 4.2 (podpis informuje tylko, że są to wyniki klasyfikacji decyzji), ani Tab. 4.3. Jest to niewielki, głęboko schowany obszar, łączący tylny koniec kory zakrętu obręczy z zakrętem przyhipokampowym. Ocena aktywności tego obszaru za pomocą EEG nie może być wiarygodna. Model wprost pokaże, jaki jest wkład aktywności dipoli umiejscowionych w tym obszarze do mierzonego sygnału na powierzchni głowy.

W podsumowaniu eksperymentu pilotażowego autor podkreśla, że wyniki dają podstawę dla rozwoju metody pomiarowej wiarygodności wiadomości. Nie wyobrażam sobie jednak, by dało się te wyniki odtworzyć, gdyż nie udostępniono danych, oprogramowania, a opis teoretyczny samej procedury („potoki i algorytmu", jak pisze Autor) jest niezbyt dokładny. Konkluzja o konieczności dalszej weryfikacji tych wyników jest na pewno bardzo słuszna.

Omawianie wyników głównego eksperymentu zaczyna się od twierdzenia „Klasyfikator regresji logistycznej został zaimplementowany w języku R". Ponieważ jest to standardowa metoda w bibliotece R, co było tu do implementacji? Dłuższe wiadomości pobudzają mózg, a efekty takiego torowania powinny być widoczne w aktywacjach struktur nawet bez podejmowania decyzji. W sekcji 4.4 autor wspomniał, że do rozróżnienia czytania krótkiej i długiej notatki na poziomie 92% wystarczy sygnał z dwóch pól Brodmanna. Nie dziwi więc, że są różnice modeli dla długich i krótkich wiadomości. Wybrane pola Brodmanna się różnią, ale nie wiemy, jak duże były to różnice, na ile są znaczące, na ile mogą być efektem wyboru specyficznych danych, czy takie same pola wybierane są w kroswalidacji. Przydałyby się histogramy to ilustrujący. Opis pól Brodmana użytych w wybranych modelach jest bardzo skrótowy, każde z nich ma wiele funkcji, zależnie od kontekstu badania, w którym opisywano jego funkcje. Model używający tylko sygnałów z pól BA8 i BA9 dał słabe wyniki, ale trudno uznać te pola za decydujące w podejmowaniu decyzji. Szczególnie obszar BA9 bierze udział w bardzo wielu procesach.

Modele były oparte na standaryzowanych wartościach MEC. Przedziały czasu to 105 do 330 ms dla krótkich wiadomości, a 830-855 dla długich. Zwykle szuka się różnic w kompo-

nentach potencjałów wywołanych, jest na ten temat obszerna literatura, ale takiego porównania nie zrobiono. MES jest nową techniką, której wiarygodność trudno jest ocenić. Tabele A1-A4 podają przedziały czasowe, dla których otrzymano istotne różnice w przypadku zaufania i braku zaufania, ale nie ma podanych istotności tych różnic, wariancji dla wszystkich badanych. W niektórych przypadkach przedziały są bardzo krótkie, albo jest tylko jedna liczba, np. E72 ma wartość 4. Nie wiadomo co to znaczy. Różnice w krótkich przedziałach nie mają zapewne realnego znaczenia.

Ostateczne wnioski opisane są w rozdziale 5. Przytoczone dokładności odnoszą się do zbioru walidacyjnego, a nie kroswalidacji. Rozważania na temat możliwości praktycznych zastosowań badań EEG do oceny wiarygodności złożonych wiadomości wydają się zbyt optymistyczne. Złożoność wiedzy, potrzebnej do rozstrzygnięcia, czy mamy do czynienia z oszustwem wymagałaby mapy siatki pojęciowej zarówno odbiorcy, jak i głębszej analizy samego prezentowanego tekstu. Takie analizy tekstu zrobiono w niedawnej pracy Miani, A., Hills, T., & Bangerter, A. (2022). Interconnectedness and (in)coherence as a signature of conspiracy worldviews. *Science Advances*, *8*(43). Jak to jednak przełożyć na badania EEG?

Redakcja pracy budzi wiele zastrzeżeń. Tłumaczenie „attrition rate" jako „stopnia ścieralności" jest dobre w materiałoznawstwie, ale nie w naukach behawioralnych, to jest stopień wykruszania się uczestników. Określenia precyzyjność i recall to w polskiej literaturze precyzja i czułość, często podaje się też swoistość. Cytowania nie zawsze są poprawne, np. [115] JO et al Rinne, lub „według Hadamarda [158]" to niewłaściwy odnośnik. W wielu miejscach rozjechało się formatowanie. Tym samym wielkościom – potencjałom na skórze głowy, prądom, macierzy wiodącej (zwanej zwykle macierzą przejścia, leadfield matrix) - przypisane są nawet w tym samym rozdziale różne symbole. Macierz referencyjna H w równaniu (3.13) nie została zdefiniowana. Skrót MEC pojawił się w końcowym podrozdziale 3.9.5, ale jego rozwinięcie jest 10 stron dalej. Ma to być zapewne ładunek w obszarze Brodmanna. Takich błędów jest niestety więcej.

Oprócz samej pracy doktorskiej mgr Łukasz Kwasniewicz jest współautorem 10 prac naukowych zgodnych z ogólną tematyką przedstawioną w rozprawie, opublikowanych w czasopismach o wysokiej randze, połowa z nich w „Frontiers in neuroinformatics". Jego wkład do tych publikacji nie został jednak określony.

Podsumowując, praca doktorska dotyczy aktualnej i ważnej tematyki analizy sygnałów EEG w zastosowaniu do ocen wiarygodności informacji. Autor podjął się beznadziejnie trud-

nego zadania, niewiele osób odważyło się zająć tym tematem. Zaprojektował proste eksperymenty w celu wyizolowania czynników mających wpływ na decyzję, wykonał odpowiednie pomiary na stosunkowo dużej grupie nadanych i przeanalizował wyniki, stosując nowe metody, opracowane w grupie prof. Wójcika.

W recenzji przedstawiłem wiele uwag krytycznych. Pomimo tych zastrzeżeń uważam jednak, iż osiągnięte wyniki są wystarczająco interesujące, by uzasadnić wniosek o dopuszczenie Autora do dalszych etapów przewodu doktorskiego.

Prof. Włodzisław Duch,

Toruń, 4/11/2022