dr hab. inż. Piotr A. Kowalski, prof. AGH
Katedra Informatyki Stosowanej i Fizyki Komputerowej,
Wydział Fizyki i Informatyki Stosowanej,
Akademia Górniczo-Hutnicza w Krakowie,
al. Mickiewicza 30, 30-059 Kraków
email: pkowal@agh.edu.pl

oraz

Centrum Informatycznych Metod Analizy Danych
Instytut Badań Systemowych
Polskiej Akademii Nauk

Kraków, 5.01.2026

RECENZJA

rozprawy doktorskiej Pana mgra Szymona Niewiadomskiego pt.
„Fixing data quality issues in CMDB in IT and OT by using machine learning algorithms. Forecasting for IT procurement."

1. Uwagi ogólne

Prawną podstawą przygotowania recenzji rozprawy doktorskiej Pana mgra Szymona Niewiadomskiego jest pismo nr RDN-ITT/204/2025 z Politechniki Wrocławskiej podpisane przez Przewodniczącego Rady Naukowej Dyscypliny Informatyka Techniczna i Telekomunikacja Pana Profesora dr. hab. Wojciecha Bożejko, które otrzymałem wraz z egzemplarzem dysertacji 10 listopada 2025 r. Recenzja została przygotowana na podstawie rozprawy doktorskiej, która została zrealizowana pod kierunkiem Pana prof. PWr dr. hab. Grzegorza Mzyka. Recenzowana rozprawa doktorska przedstawiona została w postaci woluminu wydanego przez Politechnikę Wrocławską i jest efektem prac realizowanych w ramach tzw. doktoratu wdrożeniowego. Rozprawa doktorska jest rozpatrywana w dziedzinie nauk inżynieryjno-technicznych w dyscyplinie informatyka techniczna i telekomunikacja.

Przedłożona do oceny rozprawa doktorska została napisana w języku angielskim i obejmuje 106 stron, ma układ czterech rozdziałów merytorycznych: Introduction (s. 10), Data Cleaning (s. 31), Purchasing Strategy (s. 71) oraz Implementation (s. 85). Część zasadniczą uzupełniają cztery załączniki (Appendix A–D), rozpoczynające się odpowiednio na stronach 89, 93, 95 i 99, oraz bibliografia licząca 69 pozycji (zakres [1]–[69]). W pracy zamieszczono 22 rysunki (numeracja Fig. 1–3.9) oraz 10 tabel (numeracja Tab. 1–2.6).

Zakres merytoryczny dysertacji obejmuje dwa powiązane nurty: metody poprawy jakości danych (w szczególności detekcję nieprawidłowości) w systemach klasy Configuration Management Databases (CMDB) oraz zagadnienia prognozowania i optymalizacji decyzji zakupowych w obszarze IT.

W ramach głównych tez pracy Doktorant wskazuje, że algorytmy uczenia maszynowego, zaprojektowane w sposób zapewniający pełną transparentność, architekturę modułową oraz analitycznie uzasadnione mechanizmy detekcji błędów, mogą istotnie poprawiać jakość danych przechowywanych w bazach klasy CMDB zarówno w obszarach IT, jak i Operational Technology (OT). Zaproponowane metody — w tym metody jądrowe w szczególności rekurencyjna estymacja funkcji gęstości prawdopodobieństwa, miary entropii Jaccarda oraz interpretowalne sieci neuronowe wyprowadzone z pierwszych zasad — nie opierają się na nieprzejrzystych heurystykach ani bibliotekach typu „black-box", przeciwnie, są osadzone w jednoznacznie zdefiniowanych ramach matematycznych.

Wkład teoretyczny pracy obejmuje:

- sformułowanie nowych rekurencyjnych algorytmów uczenia, uwzględniających mechanizmy „zapominania" (forgetting) i „wycofania" (rollback) w nieparametrycznym ujęciu probabilistycznym;
- analityczne i eksperymentalne porównanie metod detekcji anomalii opartych na miarach odległości oraz korelacji w heterogenicznych danych konfiguracyjnych;
- opracowanie kryteriów walidacji odporności i wiarygodności modeli w środowiskach operacyjnych.

W części pierwszej (rozdział 2) przedstawiono transparentny, modułowy łańcuch przetwarzania wspierający cykliczny (miesięczny) przegląd jakości danych, w którym – przy typowym ograniczeniu budżetu weryfikacji do ok. 100 rekordów miesięcznie – generuje się ranking podejrzanych wpisów do inspekcji eksperckiej. Zaproponowane podejście opiera się na automatycznej konstrukcji cech (w tym na tokenizacji n-gramowej), nieparametrycznej estymacji gęstości (KDE) dla klas rekordów poprawnych i błędnych, doborze cech z uwzględnieniem jednocześnie ich zdolności dyskryminacyjnej i redundancji informacyjnej (m.in. przez miary typu Jaccard/Rajski), a następnie na klasyfikatorze sieci neuronowej o płytkiej architekturze, mającym modelować nieliniowe zależności pomiędzy cechami.

W części drugiej (rozdział 3) sformułowano problem optymalizacji strategii zakupowej zasobów IT/OT (w szczególności magazynów danych) przy ograniczeniach popytu, ceny, dostępności oraz kosztów procesowych zależnych od progów zakupowych. Autor proponuje algorytm genetyczny, obejmujący m.in. zasady inicjalizacji populacji, selekcji, krzyżowania, mutacji i warunek stopu, wspierany prostymi estymatorami funkcji popytu i ceny. Rozdział 4 zawiera wskazówki wdrożeniowe (organizacyjne i architektoniczne) dla implementacji rozwiązania.

2. Ocena rozprawy

a. Uwagi krytyczno-polemiczne:

W niniejszej cześci pokrótce przedstawione zostaną główne mankamenty dysertacji. Ze względu na istotność tejże części recenzji niniejsze uwagi zostaną przedstawione w punktach, do których łatwiej będzie się odnieść Doktorantowi.

1. Rozprawa obejmuje dwie zasadnicze linie badawcze (jakość danych CMDB oraz optymalizacja zakupów). Choć Autor wskazuje wspólny kontekst zarządzania usługami IT, w tekście przydałoby się mocniejsze, bardziej formalne uzasadnienie spójności problemowej oraz wskazanie, które elementy części „Purchasing Strategy" wynikają bezpośrednio z potrzeb i danych CMDB (np. przez powiązanie prognoz popytu z miarami jakości danych lub procesem capacity management).
2. W rozdziale 2 Doktorant wprowadza techniki nieparametrycznych estymatorów jądrowych funkcji gęstości prawdopodobieństwa, lecz nie precyzuje On takich zagadnień jak np. w jaki sposób taktowane są dane wielowymiarowe? Czy używana jest koncepcja jądra produktowego czy radialnego? W jaki sposób dokonano wyboru metody wyznaczania parametrów wygładzania(motywacja)? Dość naturalnym jest użycie mało kosztowej metody modyfikacji parametru wygładzania, czego nie zrobiono – dlaczego?
3. W rozdziale 2 znacząca część ewaluacji opiera się na danych syntetycznych (np. eksperyment z numerami seryjnymi) oraz na danych anonimowych o strukturze zgodnej z produkcyjną. Z perspektywy pracy doktorskiej warto byłoby uzupełnić tekst o bardziej systematyczną prezentację walidacji na danych operacyjnych: charakterystykę typów błędów obserwowanych w CMDB, sposób etykietowania oraz stabilność wyników w czasie.
4. Autor deklaruje generację bardzo dużej liczby cech (potencjalnie „milionów"). W pracy opisano filtrację (wariancja, częstość), jednak brakuje precyzyjnych informacji o złożoności obliczeniowej pełnego potoku w typowym cyklu miesięcznym (czas/zasoby dla M oraz M') oraz o wpływie doboru słownika n-gramów (np. wielkość słownika) na koszty i jakość detekcji.
5. Wskaźnik ufności oparty na mnożeniu wskaźników cech zakłada niezależność par cech, a następnie stosuje się redukcję redundancji oraz klasyfikator sieciowy. Wskazane byłoby doprecyzowanie, w jakim sensie zachowana jest „interpretowalność" rozwiązania po przejściu na sieć neuronową oraz jakie narzędzia interpretacji (np. Analiza wrażliwości/ablacje) przewiduje Autor dla wdrożeń krytycznych.
6. W eksperymentach jako główną miarę jakości raportowana jest precyzja w top-K (P@K), co jest uzasadnione scenariuszem rankingowym. Dla pełniejszego obrazu przydatne byłyby jednak również miary odzwierciedlające pokrycie błędów (np. recall w ramach budżetu inspekcji, krzywe precision–recall) oraz analiza wpływu odsetka błędów w populacji na wyniki.

7. W rozdziale 2.2 przedstawiono podejście all-pairs dla detekcji anomalii VIN, z omówieniem kosztu $O(N^2)$. W pracy wskazano możliwości ograniczenia kosztu (bramkowanie, wstępne przefiltrowanie), jednak w kontekście wdrożeń dla dużych CMDB przydałoby się bardziej konkretne porównanie z metodami skalowalnymi (np. przybliżone najbliższe sąsiedztwo, klastrowanie) oraz kryteria wyboru pomiędzy podejściami globalnymi i lokalnymi.
8. W części dotyczącej błędów strukturalnych (CMDB jako graf) formalizacja jest poprawna, natomiast ocena empiryczna tych modułów jest w pracy ograniczona. Wskazane byłoby doprecyzowanie, jakie dane (schemat Σ, sygnały operacyjne) są w praktyce dostępne i jak mierzyć skuteczność propozycji „add-edge / unlink / relink" w warunkach ograniczonego nadzoru.
9. W rozdziale 3 przyjęto konkretne postacie estymatorów ceny i popytu (w tym AR(1)). Warto byłoby uzasadnić zakres ich stosowalności oraz przedyskutować ryzyka modelowania: niestacjonarność, zdarzenia skokowe, opóźnienia dostaw i niepewność prognoz. Część tych zagadnień Autor wskazuje jako „open problems", jednak ich wpływ na stabilność strategii zakupowej mógłby zostać mocniej zaakcentowany w części zasadniczej.

Podkreślając pozytywną wartość merytoryczną pracy oraz ciekawy dobór tematyki, należy zauważyć, że sam układ pracy wydaje się mniej przemyślany. Struktura sprawia wrażenie przygotowanej bez należytej staranności, co nieco utrudnia płynność odbioru całości. Być może jest to pokłosie wdrożeniowego trybu postępowania doktorskiego, w którym rezygnuje się z pełnej i kompletnej dysertacji naukowej na rzecz skrótowego opisu przekładając ciężar na aspekt aplikacyjny. Niemniej jednak, staranniejsze zaplanowanie i uporządkowanie poszczególnych części znacząco podniosłoby odbiór tak wartościowego opracowania. W szczególności dość problematycznym jest układ, w którym poszczególne rozdziały są niezbalansowane w kontekście ich długości oraz nagminnie używane są wypunktowania, które raczej prowadzą do skrócenia woluminu, który nie ma (w tym przypadku) rozwleczonego charakteru. Dodatkowo dość męczącym dla czytelnika jest bardzo granularny opis dysertacji polegający na wprowadzeniu – w opinii recenzenta – nadmiernej liczby podpunktów, które w niektórych przypadkach obejmują treścią kilka linii tekstu.

W szczególności z racji na obowiązki recenzenta, muszę w tym miejscu wskazać jeszcze kilka uwag związanych z pojedynczymi niedociągnięciami stylistycznymi, czy dość skąpym opisem pewnych części pracy. Doktorant również nie ustrzegł się nielicznych mankamentów natury technicznej, takich jak błędy interpunkcyjne, czy też w pracy występują nieliczne usterki edytorskie np. odwołania typu „equation (??)" oraz literówki w opisach rysunków itp. jednak w dużej mierze nie rzutują one na ocenę pracy.

b. Ocena ogólna.

Praca podejmuje istotny z naukowego i aplikacyjnego punktu widzenia problem zapewnienia jakości danych w CMDB, który ma bezpośrednie przełożenie na niezawodność procesów zarządzania usługami IT oraz odporność cybernetyczną. Na uwagę zasługuje przyjęcie perspektywy „human-in-the-loop", w której celem algorytmu jest generowanie rankingu rekordów do weryfikacji, a nie bezwarunkowa automatyzacja decyzji. Takie ustawienie problemu jest zgodne z realnymi ograniczeniami organizacyjnymi i umożliwia iteracyjną adaptację modelu.

Do mocnych stron rozprawy należy zaliczyć: (i) konsekwentne dążenie do transparentności metod (jawne zdefiniowanie potoku cech, estymatorów i kryteriów selekcji), (ii) zestawienie ujęcia probabilistyczno-nieparametrycznego (KDE, wskaźniki dyskryminacji) z ujęciem sieciowym dla modelowania zależności nieliniowych, (iii) omówienie redundancji cech miarami entropijnymi oraz (iv) uwzględnienie kontekstu wdrożeniowego (on-premise, ograniczenia bezpieczeństwa, utrzymanie).

W rozdziale 2 przedstawiono wyniki eksperymentów wskazujące przewagę klasyfikatora sieciowego nad podejściem opartym o współczynnik ufności w zakresie precyzji detekcji. Uzupełnieniem jest analiza ekonomiczna, w której – przy przyjętym koszcie weryfikacji 100 USD/rekord i 10% odsetku błędów – koszt wykrycia pojedynczego błędu może ulec istotnemu obniżeniu, a jakość CMDB w kolejnych cyklach kontroli rośnie szybciej niż przy inspekcji losowej (przy tym samym budżecie 100 rekordów miesięcznie).

Na pozytywną ocenę zasługuje także rozdział 2.2, w którym przeanalizowano podejście oparte na dystansach, w tym eksperyment all-pairs dla danych VIN (N = 1000), pozwalający interpretować częstość występowania rekordu w parach o dużym dystansie jako heurystykę wskazującą kandydatów błędów. W części poświęconej zakupom (rozdział 3) algorytm genetyczny jest opisany w sposób kompletny i zweryfikowany na scenariuszach brzegowych, a porównania ze strategiami standardowymi wskazują możliwość uzyskania oszczędności rzędu 5–20%.

Rozdział 4 i załączniki stanowią użyteczne uzupełnienie pracy: zawierają zalecenia dotyczące wdrożenia i utrzymania rozwiązania, samoocenę w duchu Trustworthy AI oraz udokumentowanie wykorzystania generatywnych modeli językowych w procesie przygotowania tekstu (Appendix B), co zwiększa „przejrzystość warsztatu Autora".

3. Podsumowanie

Podsumowując, rozprawa przedstawia zestaw metod i analiz dotyczących poprawy jakości danych w CMDB oraz wspomagania decyzji zakupowych w obszarze IT/OT. Autor wykazuje się znajomością literatury, poprawnym warsztatem matematycznym i umiejętnością przekładu problemu praktycznego na modele i procedury obliczeniowe. Zgłoszone uwagi mają w większości charakter doprecyzowujący i edytorski i nie podważają zasadniczej wartości naukowej uzyskanych

rezultatów. W myśl wymagań sprecyzowanych w „Prawo o szkolnictwie wyższym i nauce" (Dz. U. 2024, poz. 1571) można stwierdzić, iż recenzowana rozprawa doktorska autorstwa Pana mgra Szymona Niewiadomskiego prezentuje ogólną wiedzę teoretyczną kandydata w dyscyplinie informatyka techniczna i telekomunikacja oraz umiejętność samodzielnego prowadzenia pracy naukowej, a jej przedmiotem jest oryginalne rozwiązanie problemu naukowego oraz oryginalne rozwiązanie w zakresie zastosowania wyników własnych badań naukowych w sferze gospodarczej.

Konkludując uważam, że rozprawa doktorska Pana mgra Szymona Niewiadomskiego spełnia wymagania stawiane w odpowiednich przepisach rozprawom doktorskim i wobec tego stawiam wniosek o jej dopuszczenie do dalszych, przewidzianych Ustawą, etapów postępowania o nadanie stopnia doktora.

dr hab. inż. Piotr A. Kowalski, prof. AGH