Załącznik 3 Katowice 2022-11-29

Autoreferat

1. Imię i nazwisko

Rafał Deja

Akademia WSB
ul. Cieplaka 1C, 41-300 Dąbrowa Górnicza
mail: rdeja@wsb.edu.pl

2. Posiadane dyplomy, stopnie naukowe

- Stopień doktora nauk technicznych w zakresie informatyki uzyskany w Instytucie Podstaw Informatyki Polskiej Akademii Nauk w Warszawie, 2001 rok.
  Tytuł rozprawy doktorskiej: Zastosowanie teorii zbiorów przybliżonych w analizie konfliktów

- Stopień magistra inżyniera informatyki uzyskany na Politechnice Śląskiej, Wydział Informatyki, Elektroniki i Automatyki, Gliwice, kierunek Informatyka, rok 1992;

3. Informacja o dotychczasowym zatrudnieniu w jednostkach naukowych lub artystycznych

2008-teraz Katedra Transportu i Informatyki, Akademia WSB, ul. Cieplaka 1C, 41-300 Dąbrowa Górnicza

1993-1996 Politechnika Śląska, Krasińskiego 8, 40-019 Katowice

4. Omówienie osiągnięć naukowych

4.1 Tytuł osiągnięcia naukowego

Jako podstawowe osiągniecie naukowe przedstawiam cykl dziesięciu publikacji monotematycznych. Wszystkie wskazane artykuły zostały zrealizowane po uzyskaniu przeze mnie tytułu doktora. Tytuł osiągnięcia naukowego:

Metody sztucznej inteligencji w modelowaniu dynamiki procesów chorobowych z danych i wiedzy dziedzinowej.

4.2 Publikacje

A1. Deja, Rafal, Wojciech Froelich, "Forecasting Basal Insulin for the Clinical Therapy of Juvenile Diabetes at Onset." *Procedia Computer Science* 207 (2022): 138-144.
Wkład autorów:
RD: zaproponowałem koncepcję rozwiązania, opracowałem dane i przeprowadziłem eksperymenty, przygotowałem pierwszą wersję artykułu.
WF: opracowanie manuskryptu artykułu

A2. Deja, Rafal, Wojciech Froelich, and Grazyna Deja. "Rule-based Medical Treatment Graph for the Modeling of Hypo-and Hyperglycemia at Onset." *Procedia Computer Science* 192 (2021): 1393-1400.
Wkład autorów:
RD: zaproponowałem koncepcję rozwiązania, opracowałem algorytm i przeprowadziłem eksperymenty, przygotowałem pierwszą wersję artykułu.
WF: opracowanie manuskryptu artykułu
GD: weryfikacja medyczna, interpretacja wyników.

A3. Deja, Rafal, Wojciech Froelich, and Grazyna Deja. "Mining clinical pathways for daily insulin therapy of diabetic children." *International Journal of Applied Mathematics and Computer Science* 31.1 (2021).
Wkład autorów:
RD: zaproponowałem koncepcję rozwiązania, zebrałem dane, opracowałem algorytm i przeprowadziłem eksperymenty, przygotowałem pierwszą wersję artykułu.
WF: opracowanie manuskryptu artykułu, formalne ujęcie nowości merytorycznej
GD: weryfikacja medyczna, interpretacja wyników.

A4. Deja, Rafał, Wojciech Froelich, Grażyna Deja, and Alicja Wakulicz-Deja. "Hybrid approach to the generation of medical guidelines for insulin therapy for children." *Information Sciences* 384 (2017): 157-173.
Wkład autorów:
RD: zaproponowałem koncepcję rozwiązania, która w toku dyskusji ze współautorami ulegała modyfikacjom i rozszerzeniom, opracowałem algorytm i przeprowadziłem eksperymenty, przygotowałem pierwszą wersję artykułu.
WF: opracowanie ostatecznej koncepcji rozwiązania, opracowanie manuskryptu artykułu
GD: weryfikacja medyczna, interpretacja wyników.
AWD: udział w ostatecznej redakcji artykułu.

A5. Deja, Rafal. Applying Roughication to Support Establishing Intensive Insulin Therapy at Onset of T1D. In: Czarnowski, I., Howlett, R., Jain, L. (eds) Intelligent Decision Technologies 2017. IDT 2017. Smart Innovation, Systems and Technologies, vol 72. pp. 265-272. Springer, Cham, 2018.
Wkład autora: 100%

A6. Deja, Rafal. Building Medical Guideline for Intensive Insulin Therapy of Children with T1D at Onset. In: Nguyen, N., Iliadis, L., Manolopoulos, Y., Trawiński, B. (eds) Computational Collective Intelligence. ICCCI 2016. Lecture Notes in Computer Science(), vol 9876. Springer, Cham. 2016
Wkład autora: 100%

A7. Deja, Rafał, Wojciech Froelich, and Grazyna Deja. "Differential sequential patterns supporting insulin therapy of new-onset type 1 diabetes." *BioMedical Engineering OnLine* 14.1 (2015): 1-11.
Wkład autora.
RD: zaproponowałem koncepcję rozwiązania, zebrałem dane, opracowałem algorytm i przeprowadziłem eksperymenty, przygotowałem pierwszą wersję artykułu
WF: opracowanie manuskryptu artykułu
GD: weryfikacja medyczna, interpretacja wyników.

A8. Froelich, Wojciech, Rafał Deja, and Grażyna Deja. "Mining therapeutic patterns from clinical data for juvenile diabetes." *Fundamenta Informaticae* 127.1-4 (2013): 513-528.
Wkład autora:
RD: opracowanie danych, implementacja algorytmów i przeprowadzenie eksperymentów, przygotowanie artykułu.
WF: zaproponowanie koncepcji rozwiązania, opracowanie algorytmu, opracowanie ostatecznej wersji manuskryptu
GD: weryfikacja medyczna, interpretacja wyników.

A9. Deja, Rafal. Comparison of Rules Synthesis Methods Accuracy in the System of Type 1 Diabetes Prediction. In: Kapczyński, A., Tkacz, E., Rostanski, M. (eds) Internet - Technical Developments and Applications 2. Advances in Intelligent and Soft Computing, vol 118, pp 13-44. Springer, Berlin, Heidelberg. 2012
Wkład autora: 100%

A10. Deja, Rafal. Applying Rough Set Theory to the System of Type 1 Diabetes Prediction. In: Tkacz, E., Kapczynski, A. (eds) *Internet – Technical Development and Applications. Advances in Intelligent and Soft Computing*, vol 64. pp.119-129, Springer, Berlin, Heidelberg, 2009.
Wkład autora: 100%

4.3 Omówienie celu naukowego wymienionych prac i osiągniętych wyników

4.3.1 Wprowadzenie, motywacja i opis tematyki badawczej

Modelowanie procesów chorobowych z danych historycznych używając wiedzy dziedzinowej wymaga zastosowania wielu różnych metod sztucznej inteligencji i opracowania nowych algorytmów adekwatnych do przedmiotu modelowania.

Można zauważyć, że w przebiegu choroby mamy do czynienia z powtarzającymi się zdarzeniami w czasie (temporal data) np. przyjmowanie leków, wykonywanie badań. Zdarzenia te mogą trwać i wchodzić w relacje (np. jedno zdarzenie jest konsekwencją innego). Takie dane czasowe wymagają w zależności od celu analizy, różnej reprezentacji i poziomu abstrakcji, a podczas analizy pojawia się wiele pytań. Jaka jest relacja między sekwencją zdarzeń w czasie i czy możemy oszacować wartość interesującej nas zmiennej? Czy istnieją jakieś powtarzające się zachowania w przebiegu choroby? Czy możemy wyróżnić grupy pacjentów z podobnym przebiegiem choroby, czy potrafimy je scharakteryzować, aby ich podobnie leczyć? Czy możemy zidentyfikować znaczniki czasu, w których najczęściej zachodzą niepożądane zdarzenia? Odpowiedzi na te pytania wspomagają lekarza w podejmowaniu najlepszych decyzji terapeutycznych, a zastosowane algorytmy są podstawą do tworzenia komputerowych systemów wspomagania decyzji (DSS).

Wraz z rozwojem metod sztucznej inteligencji już w latach 80 zaczęły powstawać systemy ekspertowe. Podjęto próby stworzenia takich systemów dla różnych dziedzin. Do najbardziej znanych należą: system MYCIN [1] dla medycyny, którego silnik wnioskowania opierał się na ~600 regułach decyzyjnych czy budowany przez 15 osobolat system Dendral służący do ustalanie struktury molekularnej nieznanych chemicznych związków organicznych na podstawie analizy widm spektroskopowych [2]. Współczesnym przykładem jest (ostatecznie porzucony) niedawny projekt IBM, aby wykorzystać superwydajny system komputerowy Watson do poprawy leczenia pacjentów onkologicznych [3].

Intensywne prace zmierzające w istocie do całkowitego zastąpienia ekspertów przez systemy automatyczne nie zakończyły się jak dotąd sukcesem. Najważniejsze problemy związane są z pozyskaniem od eksperta i wyczerpującym opisaniem wiedzy dziedzinowej („kruchość systemów" (brittleness)) [4][5][6]. Można powiedzieć, że podstawową przyczyną niepowodzeń jest w dalszym ciągu brak zrozumienia rozumowań bazujących na doświadczeniu:

"The quest for machines that can make abstractions and analogies is as old as the AI field itself, but the problem remains almost completely open" [7].

Opinia Davida Husserla, twórcy fenomenologii wskazuje, że pominiecie doświadczenia eksperckiego stawia projektantów systemów wspomagania decyzji w niezwykle trudnej sytuacji, zwłaszcza że nasza wiedza o rozumowaniach eksperckich dotyczących ich doświadczenia, jak zauważyliśmy, jest nikła.

„Husserl was frustrated by the idea that science and mathematics were increasingly conducted on an abstract plane [treating nature itself as a mathematical manifold] that was disconnected from human experience and human understanding, independently of questions of truth and applicability. He felt that the sciences increasingly dealt with idealized entities and internal abstractions a world apart from the concrete phenomena of daily life" [8].

W związku z tym, że dotychczasowe rozwiązania nie były wystarczająco skuteczne, w przedstawianym osiągnieciu naukowym prezentuję inne (skromniejsze) podejście. Jego celem jest ułatwianie podejmowania decyzji przez eksplorację i dostarczenie ekspertom medycznym odpowiedniej analizy danych historycznych. Tak pozyskana wiedza w każdym momencie umożliwia odwołania się do przypadków wspierających przedstawiane wnioski (case-based reasoning [9]). Następnie system poprzez właściwą prezentację (wizualizację) jak i dialog z ekspertem umożliwia podjęcie najlepszej decyzji. Aby więc w pełni wykorzystać doświadczenie eksperta i uniknąć wad takich jak „kruchość systemu" musimy również ograniczyć system wspomagania decyzji do węższej dziedziny reprezentowanej przez eksperta. Moje badania skupiają się na modelowaniu leczenia cukrzycy typu 1 w sytuacji świeżego ujawnienia cukrzycy. Doświadczenie eksperta dziedzinowego jest wykorzystane na różnych etapach projektowania systemów DSS w tym przy tworzeniu algorytmów zaproponowanych w osiągnięciu naukowym.

W prowadzonych badaniach wspomaganie decyzji opiera się na nowych opracowanych przeze mnie metodach i algorytmach; eksploracji czasowych danych dziedzinowych [A4-A8] (mining of temporal data [10]) oraz ich wizualizacji w postaci specjalnie przygotowanego grafu - prace [A2, A3, A4]. W pozostałych pracach [A1,A9,A10] proponowane narzędzia DSS opierają się na predykcji szeregów czasowych [11] z wykorzystaniem sieci neuronowych i zastosowaniu teorii zbiorów przybliżonych [12].

4.3.2 Podstawowe informacje dziedzinowe

W moich badaniach nad systemami wspomagania decyzji jako wiedzę dziedzinową wykorzystuję wiedzę z zakresu medycyny, w szczególności diabetologii. Przyjęcie tej dziedziny wiąże się z możliwością konsultacji z lekarzem diabetologiem, ale głównie wynika to też ze szczególnych cech wiedzy w tym zakresie.:

- duża dynamika zmienności danych i różnorodność sytuacji klinicznych,
- zależności czasowe podejmowanych decyzji terapeutycznych: glikemia/dawka insuliny,
- złożoność problemu medycznego: wpływ różnorodnych czynników ostatecznie kształtujących glikemię,
- ważność i aktualność badań w tym zakresie.

Cukrzyca jest jedną z ważniejszych chorób cywilizacyjnych. W ostatnich latach liczba przypadków zachorowań gwałtownie rośnie [13]. W wielu krajach zaczęło to być poważnym problemem społecznym i gospodarczym. Dlatego prowadzonych jest wiele badań dotyczących różnych aspektów tej choroby.

Podstawowym celem terapii cukrzycy jest doprowadzenie do stabilizacji poziomu glukozy we krwi pacjenta, czyli tzw. normoglikemii przez cały dzień przy użyciu insuliny podawanej pacjentowi w ściśle dobranych dawkach [14]. W celu kontroli poziomu glukozy przeprowadza się w ciągu dnia wielokrotne pomiary. W razie znaczącego przekroczenia zalecanego poziomu (hipo/hiper glikemii) dokonuje się bieżącej interwencji (dodatkowa dawka insuliny, wstrzymanie się od jedzenia bądź spożycie posiłku) oraz analizując dotychczasowy przebieg choroby wprowadza się korekty w zastosowanym leczeniu.

Generalnie leczenie w cukrzycy typu 1 polega na podawaniu insuliny w określony sposób. Dawka dobowa dzielona jest na insulinę bazową i doposiłkową. Insulina doposiłkowa zapewnia regulację poziomu glukozy po spożyciu węglowodanów (wraz z posiłkiem) a insulina bazowa utrzymuje właściwy poziom glukozy poza posiłkami. Powstaje więc wiele prac, które w różny sposób, stosując różne metody wspomagają decyzję lekarza oraz pacjenta w doborze dawki insuliny [15][16][17][18][19][20]. Bardzo pomocny jest również szybki rozwój technologii. Coraz powszechniej stosowane są pompy insulinowe oraz systemy ciągłego monitorowania glikemii (CGM). Jednym z obiecujących rozwiązań jest zastosowanie tzw. pętli zamkniętej, czyli systemu, który na podstawie ciągłego odczytu glikemii na bieżąco koryguje wlew insuliny w pompie insulinowej [21].

4.3.3 Zaproponowane rozwiązania oraz uzyskane wyniki

Poniżej przedstawiam wprowadzenie teoretyczne dotyczące stosowanych metod oraz prezentuje poszczególne etapy przeprowadzonych badań. Na każdym etapie badań opisuję podjęty problem wraz z jego motywacjami, przedstawiam proponowane rozwiązanie (nowość teoretyczną) oraz wybrane efekty praktyczne wprowadzonych nowości.

Problem 1

W moich badaniach podjąłem problem wspomagania pracy lekarza diabetologa w ustaleniu terapii dla świeżych zachorowań na cukrzycę. W momencie zachorowania stan pacjenta jest niestabilny; reakcja organizmu różni się u różnych pacjentów i zależy od wielu czynników [14]. Poza wagą pacjenta czy szerzej zapotrzebowaniem energetycznym do najważniejszych czynników należy stan kliniczny pacjenta; częste jest występowanie stanu zapalnego czy kwasicy ketonowej oraz wciąż

występujące własne wydzielanie insuliny. Dlatego ustalenie właściwej dawki insuliny w czasie krótkiego okresu pobytu pacjenta w szpitalu jest wyzwaniem [20]. Opiera się ono głównie na doświadczeniu lekarza i obserwacji pacjenta przez kilka dni. Niewielką pomocą są tutaj nowe technologie takie jak pompy insulinowe, które też wymagają ustawienia właściwego dawkowania insuliny. System zamkniętej pętli natomiast potrzebuje czasu, aby nauczyć się właściwej reakcji dla danego pacjenta [21]. Ponadto nie wszyscy pacjenci mogą skorzystać z tych nowoczesnych technologii, m.in. z powodu wysokich kosztów takiej terapii lub ze względów medycznych [22].

Zadaniem systemu wspomagania decyzji jest transparentne przedstawienie zaleceń medycznych z możliwością łatwego zweryfikowania przesłanek na których te zalecenia się opierają (celem nie jest jedynie predykcja dawki insuliny). System powinien też pokazywać lekarzowi konsekwencje wyboru danego sposobu leczenia jak i konsekwencje wprowadzenia zmian w leczeniu.

W tym celu niezbędne jest opracowanie modelu przebiegu leczenia umożliwiającego wydobycie wskazówek z danych historycznych i stworzenie zaleceń dla kolejnych pacjentów.

Wprowadzenie teoretyczne

Zauważmy, że w przypadku danych z przebiegu leczenia dla świeżych zachorowań na cukrzycę mamy do czynienia z szeregami czasowymi. Generalnie, można wyróżnić dwa standardowe podejścia do odkrywania wzorców czasowych z sekwencji [25]. Po pierwsze, aby odkryć powtarzające się podciągi ze zbiorów sekwencji możemy zastosować metodę wydobywania wzorców sekwencyjnych [23]. Drugie podejście koncentruje się na danych strumieniowych, które składają się z sekwencji tzw. zdarzeń. W szczególności techniką częstych epizodów można wykryć powtarzające się podsekwencje (nazywane epizodami) występujące w jednej sekwencji [24].

Ponieważ kolejne prace były inspirowane obydwoma wyżej wymienionymi podejściami, poniżej przedstawię podstawowe pojęcia dotyczące wzorców sekwencyjnych i częstych epizodów.

Niech $\mathrm{E} = \{\mathrm{e}_1, \mathrm{e}_2, \ldots, \mathrm{e}_n\}$ oznacza zbiór elementów (będących w rzeczywistości dowolnymi termami symbolicznymi). Dowolny podzbiór elementów $a_i \subseteq \mathrm{E}$ jest nazywany zestawem elementów. Sekwencja $a = <a_1, a_2, \ldots, a_k>$ jest zdefiniowana jako uporządkowana lista zestawów elementów a_i. Termin k-sekwencja oznacza dowolny ciąg o liczności k = $\mathrm{card}(a)$, gdzie k jest liczbą elementów, które zawiera. Pojęcie inkluzji odgrywa kluczową rolę podczas analizy sekwencji. Przykładowy ciąg a = < a_1, a_2, ..., a_n> zawarty jest w drugim ciągu *b* = <b_1, b_2, ..., b_m>, czyli $a \subseteq b$ jeśli istnieją liczby całkowite i_1, i_2, ..., i_n takie, że $a_1 \subseteq b_{i1}$, $a_2 \subseteq b_{i2}$, ..., $a_n \subseteq b_{in}$ [2]. Sekwencja a nazywana jest podciągiem *b*, a sekwencja *b* jest nazywana superciągiem a.

Zbiór sekwencji oznaczono jako $\mathrm{S} = \{\mathrm{s}^1, \mathrm{s}^2, \ldots, \mathrm{s}^m\}$, gdzie s^j jest j-tą sekwencją. Do oceny poziomu uogólnienia wykazanego pewnym podciągiem $a \subseteq \mathrm{s}^j$ stosuje się miarę wsparcia:

$$\sup(a) = \frac{\mathrm{card}(\mathrm{s}^j | a \subseteq \mathrm{s}^j)}{\mathrm{card}(\mathrm{S})}$$

Celem eksploracji danych jest znalezienie podsekwencji ze wsparciem powyżej określonego progu, tj. z $\sup(a) \geq \sup_{\min}$. Takie podsekwencje nazywane są wzorcami sekwencyjnymi. Najbardziej znanymi algorytmami do wydobycia wzorców sekwencyjnych są AprioriAll i AprioriSome [23]. Kilka innych algorytmów zostało zaproponowanych głównie w celu zmniejszenie złożoności obliczeniowej [27]. Przegląd algorytmów eksploracji wzorców sekwencyjnych można znaleźć w [26].

Jak wspomniano wcześniej, innym podejściem stosowanym do odkrywania wzorców w sekwencjach czasowych jest eksploracja częstych epizodów [24]. Aby uniknąć przedefiniowania symboli już

zdefiniowanych dla wzorców sekwencyjnych, oryginalna notacja używana dla częstych epizodów została nieznacznie zmodyfikowana. Pojęcie zdarzenia definiuje się jako $e = <et, \tau>$, gdzie $et \in ET$ jest etykietą (nazywaną typem zdarzenia) ze skończonego zbioru predefiniowanych etykiet, τ jest znacznikiem czasu zdarzenia [24]. Sekwencja zdarzeń jest zdefiniowana jako Trójka $s = <s, \tau_s, \tau_e>$, gdzie: $s = <e_1, e_2, ..., e_n>$. Kolejność zdarzeń w s zależy od czasu $\tau_s \leq \tau_i \leq \tau_e$, gdzie τ_s, τ_e oznaczają odpowiednio czas rozpoczęcia i zakończenia sekwencji.

Okno przesuwne w ciągu s definiuje się jako $w = (\boldsymbol{w}, \tau_{ws}, \tau_{we})$, gdzie $\boldsymbol{w} \subseteq s$ oznacza podciąg zdarzeń zachodzących od $\tau_{ws} > \tau_s$ do $\tau_{we} < \tau_e$ [24]. Przedział czasowy okna przesuwnego jest obliczany jako $\tau_{win} = \tau_{we} - \tau_{ws}$. Zestaw wszystkich okien przesuwnych w obrębie s jest oznaczony jako $W(s)$.

Epizod α jest trójką $<V, \leq, g>$, gdzie: V oznacza zbiór węzłów, ≤ jest częściowym porządkiem na V, a g: $V \rightarrow ET$ jest odwzorowaniem wiążącym każdy węzeł z typem zdarzenia. W szczególności epizod α nazywa się seryjnym, gdy relacja ≤ staje się porządkiem całkowitym <. Epizod α występuje w ciągu s (oznaczonym tutaj jako $\alpha \preccurlyeq s$), gdy istnieje odwzorowanie h: $V \rightarrow [1, ..., n]$ z węzłów na zdarzenia s takie, że $\forall_{x \in V}\, g(x) = e_{h(x)}$ oraz $\forall_{x,y \in V, x<y}\, \tau_{h(x)} < \tau_{h(y)}$. Zdolność do generalizacji wykazana przez epizod jest obliczana jako stosunek liczebności okien, w których wystąpił epizod do liczby wszystkich okien.

$$\sup(\alpha) = \frac{\mathrm{card}(\{w \in W(s) | \alpha \preccurlyeq w)}{\mathrm{card}(W(s))}$$

Rozważanym zadaniem eksploracji danych jest wykrycie wszystkich epizodów ze wsparciem powyżej zadanego progu, takie epizody są nazywane jako częste [24]. W najprostszym przypadku częste epizody można wydobywać, przekształcając źródłową sekwencję zdarzeń s na zbiór podciągów s^j, z których każdy będzie przyjmował rolę okna przesuwnego w. Następnie można wykorzystać znane algorytmy do wydobywania wzorców sekwencyjnych. Jednak ze względu na duży koszt obliczeń zaproponowano kilka innych algorytmów w celu poprawy wydajności obliczeniowej; ich przegląd można znaleźć w pracy [28].

Podsumowując, wzorzec sekwencyjny a różni się od epizodu α sposobem, w jaki jest oceniany w odniesieniu do danych źródłowych. Wsparcie dla wzorca sekwencyjnego nie uwzględnia faktu, że wzorzec może powtarzać się wielokrotnie w sekwencji pojedynczego pacjenta. Z drugiej strony, wsparcie epizodu liczy wyłącznie wystąpienia α w ramach jednej sekwencji.

Reprezentacja danych klinicznych

Problem eksploracji wzorców terapeutycznych w cukrzycy wymaga dostosowania istniejących ogólnych podejść do eksploracji wzorców sekwencyjnych i częstych epizodów. Zacznijmy od reprezentacji danych klinicznych. Zaproponowany w pracach sposób wydobycia danych klinicznych z elektronicznych rejestrów przebiegu leczenia (EHR – electronic healths records) i ich właściwa reprezentacja jest jednym z osiągnięć badawczych. Podejście to było rozwijane w kolejnych pracach [A8, A4, A3, A2] - notacje przedstawione poniżej nieznacznie różnią się od tych zastosowanych w pracach.

Zdefiniujmy zdarzenie medyczne $u \in U$ jako parę: $u = <V_i = v, \tau>$, gdzie $V_i \in V$ jest zmienną, a v oznacza wartość, którą V_i przyjmuje w czasie τ . Innymi słowy, mówimy, że zdarzenie u występuje w czasie τ , gdy zmienna V_i otrzymuje określoną wartość v ze swojej dziedziny dom(V_i) w określonym czasie τ . Zbiór U to uniwersum wszystkich możliwych zdarzeń.

W naszych badaniach przyjmujemy $\mathrm{V} = \{\mathrm{G}, \mathrm{I}\}$, czyli bierzemy pod uwagę tylko te zmienne, które dotyczą pomiarów glikemii (zmienna G) i dawek insuliny (zmienna I), gdzie z kolei $\mathrm{I}=\{I_B, I_P\}$ w zależności od obszaru analizy oznacza podanie insuliny bazowej (I_B) lub doposiłkowej (I_P).

Sekwencję kliniczną definiujemy jako $s = <u_{\tau 1}, u_{\tau 2}, ..., u_{\tau n}>$, gdzie τ_i jest czasem rzeczywistym, w którym zachodzi zdarzenie. Długość s zależy od okresu, w którym pacjent przebywa w szpitalu. Przez S oznaczamy zbiór wszystkich sekwencji.

Kolejny krok naszego podejścia do modelowania przebiegu leczenia jest związany z upływem czasu. Czasy posiłków mogą się różnić dla różnych pacjentów, podobnie z konkretnym terminem pomiaru glikemii. Z tego powodu, zdecydowaliśmy się na sekwencjonowanie czasu [25]. Jednak w przypadku naszego badania nie sekwencjonujemy całego okresu terapii, lecz w obrębie jednego terapeutycznego dnia pacjenta. Jest to zgodne z dyskretną skalą czasową stosowaną przez lekarzy do planowania codziennej insulinoterapii.

Aby sekwencjonować czas, zdefiniujemy zbiór etykiet $L = \{t_1, t_2, ..., t_{nw}\}$ i przyporządkujemy je do kolejnych przedziałów czasu podanych przez lekarzy, n_w jest liczbą etykiet rozpatrywanych w analizowanym problemie. Ponadto, aby odwzorować zdarzenia medyczne na dyskretną skalę czasu, definiujemy funkcję $t: RT \rightarrow L$, gdzie RT oznacza dziedzinę czasu rzeczywistego. Każde zdarzenie u_τ występujące w czasie rzeczywistym jest mapowane na nową, dyskretną skalę czasu jako $u_{t(\tau)}$.

Czas	Opis	Wartość
7:55	*G* - pomiar glikemii	139 mg/dl
8:00	I_P - podanie insuliny	3.5 units
8:05	Pierwszy posiłek	240 kcal
10:55	*G* - pomiar glikemii	189 mg/dl
11:00	I_P - podanie insuliny	2 units
11:05	Drugi posiłek	170 kcal
13:55	*G* - pomiar glikemii	65 mg/dl
14:00	I_P - podanie insuliny	4 units
14:05	Trzeci posiłek	380 kcal
16:55	*G* - pomiar glikemii	71 mg/dl
17:00	I_P - podanie insuliny	4.5 units
17:05	Czwarty posiłek	480 kcal

Tabela 1. Fragment oryginalnych danych medycznych

Zobaczmy, jak wygląda sekwencjonowanie czasu na przykładzie. W Tabeli 1 przedstawiono fragment danych klinicznych z przebiegu leczenia wybranego pacjenta, natomiast w Tabeli 2 prezentujemy definicję zbioru etykiet.

L	Opis	Czas	Zdarzenie
t_1	before breakfast	[6:00 - 10:00]	*G*
t_2	breakfast	[6:00 - 10:00]	I_P
t_3	after breakfast	[9:00 - 12:00]	*G*
t_4	second breakfast	[9:00 - 12:00]	I_P
t_5	after second breakfast	[11:00 - 15:00]	*G*
t_6	lunch	[11:00 - 15:00]	I_P
t_7	after lunch	[14:00 - 17:00]	*G*
t_8	dinner	[14:00 - 17:00]	I_P
t_9	after dinner	[16:00 - 20:00]	*G*
t_{10}	supper	[16:00 - 20:00]	I_P
t_{11}	after supper	[19:00 - 22:00]	*G*

Tabela 2. Przedziały czasu dla dobowej terapii

Jak można zauważyć na przykładzie podanym w Tabeli 2, doba terapeutyczna pacjenta jest podzielony według określonych przedziałów czasowych, które są związane m.in. ze spożywanymi przez pacjenta posiłkami. Należy zauważyć, że przedziały czasowe mogą się pokrywać, co jest zgodne z praktyką kliniczną. Dzięki wprowadzeniu etykiet, zamiast zajmować się ciągłym upływem czasu, możemy odwzorować planowanie terapii według określonej, dyskretnej skali czasowej. Co więcej, wprowadzając dyskretną skalę czasu, nie tylko abstrahujemy od ciągłego czasu, ale także od konkretnego dnia, w którym ma miejsce zdarzenie medyczne – doba terapeutyczna np. może zaczynać się od wieczornego podania insuliny bazowej co stosujemy w pracach [A2, A6].

Z teoretycznego punktu widzenia skonstruowana reprezentacja danych źródłowych jest wielowymiarowym szeregiem czasowym. Jednak z medycznego punktu widzenia tak szczegółowa, interpretacja często nie jest wymagana. Na przykład dla większości pacjentów istotną informacją jest to, czy poziom cukru we krwi po posiłku zinterpretowano jako normoglikemię. Mniej istotne są konkretne rzeczywiste wartości i dokładny czas wielokrotnych pomiarów wykonywanych w nocy. Z tego powodu wymagana jest wprowadzenie pewnych abstrakcji medycznie istotnych wartości zmiennych. Po pierwsze więc wprowadzono dyskretyzację wartości glikemii zgodnie z przyjętymi normami medycznymi – Tabela 3 oraz Tabela 4.

Glycemia [*mg/dl*]	**Znaczenie medyczne**		
	przed śniadaniem	**przed innym posiłkiem**	**po posiłku**
< 70	hypoglycemia	hypoglycemia	hypoglycemia
[70, 90]	normoglycemia	normoglycemia	normoglycemia
(90, 100]	mild-hyperglycemia	normoglycemia	normoglycemia
(100, 140)	mild-hyperglycemia	mild-hyperglycemia	normoglycemia
[140, 200]	mild-hyperglycemia	mild-hyperglycemia	mild-hyperglycemia
> 200	hyperglycemia	hyperglycemia	hyperglycemia

Tabela 3. Zakresy glikemii i znaczenie medyczne

Poziom glikemii	**Wartość dyskretna** $\mathrm{dom(G)}$
hypoglycemia	1
normoglycemia	2
mild-hyperglycemia	3
hyperglycemia	4

Tabela 4. Dyskretyzacja poziomu glikemii

Wartości dawek insuliny dyskretyzujemy w różny sposób dla różnych rozwiązań. Wydaje się, że najlepsza metoda została zaproponowana w pracach [A1, A2, A3, A4]. Zaproponowano tam posługiwanie się współczynnikiem insuliny. Tzn. wartość podanej dawki jest odniesiona do wagi pacjenta oraz w przypadku insuliny doposiłkowej do wielkości posiłku. Współczynnik ten jest zrozumiały dla lekarza i po zaokrągleniu daje dobry poziom abstrakcji.

Rozwiązanie 1

W celu wspomagania decyzji lekarza dotyczących ustalania wymaganych dawek insuliny zaproponowano nową metodę opartą na modelowaniu czasowej zależności poziomu glukozy od podanej insuliny [A8]. Proponowana metoda opiera się na odkrywaniu wyspecjalizowanych wzorców sekwencji z danych historycznych. W pierwszej kolejności zaproponowano realizację abstrakcji funkcyjnej dotyczącej poziomu glikemii w okresie nocnym.

Abstrakcja funkcyjna

Aby uzyskać ogólne rozwiązanie dla różnych wymagań generalizacji, w pracy [A8] zaproponowano podejście oparte na abstrakcji funkcyjnej. Taką abstrakcją może być obliczanie nocnej glikemii za pomocą wyrażeń logicznych, np. za pomocą poniższych równań:

$$\begin{cases} t(night) = normoglycemia \mid \bigwedge\limits_{l\epsilon\{t_2,t_3,t_4,t_5\}} u_l = 2\ (normoglycemia) \\ t(night) = hypoglycemia \mid \bigvee\limits_{l\epsilon\{t_2,t_3,t_4,t_5\}} u_l = 1\ (hypoglycemia) \\ t(night) = \max\limits_{l\epsilon\{t_2,t_3,t_4,t_5\}} u_l \quad \text{w przeciwnym razie} \end{cases}$$

Zbór etykiet czasu definiowany jest w następujący sposób L ={t_1, t_2, t_3, t_4, t_5} a zastosowane mapowanie do tych etykiet przedstawiono w Tabeli 5.

L	**Opis**	**Czas**	**Zdarzenie**
t_1	basal insulin injection	[21:30 - 22:30]	I_B
t_2	night0 (pomiar glikemii)	[23:30 - 0:30]	*G*
t_3	night3 (pomiar glikemii)	[2:30 - 3:30]	*G*
t_4	night5 (pomiar glikemii)	[4:30 - 5:30]	*G*
t_5	night7 (pomiar glikemii)	[6:30 - 7:30]	*G*

Tabela 5. Sekwencjonowanie czasu - znaczenie etykiet

Dyskretyzacja dawek insuliny została przeprowadzona zgodnie z zakresami podanymi w Tabeli 6.

a) Insulina bazowa

Poziom (I_B)	**Zakres [mg/dl]**
b_1	[0, 3.7)
b_2	[3.7, 6.0)
b_3	[6.0, 9.0)
b_4	[9.0, 14.5)
b_5	≥ 14.5

b) Insulina doposiłkowa

Poziom (I_P)	**Zakres [mg/dl]**
p_1	[0, 0.75)
p_2	[0.75, 1.08)
p_3	[1.08, 1.48)
p_4	[1.48, 2.05)
p_5	≥ 2.05

Tabela 6. Dyskretyzacja dawek insuliny

Dyskretyzacja i obliczanie nocnej glikemii są w rzeczywistości funkcjami z argumentami ze zbioru U zdarzeń medycznych. W celu uogólnienia podejścia proponuje się zdefiniowanie pojęcia zdarzenia funkcyjnego jako pary $<F_i = f, t(\tau)>$, gdzie F_i jest dowolną funkcją obliczoną przy użyciu zdarzeń medycznych, $f \in \mathrm{dom}(F_i)$ a $l = t(\tau)$ jest symbolem (etykietą) okresu czasu. Dla uproszczenia zdarzenia funkcyjne są zapisywane w następujący sposób $F_i(l) = f$.

Aby zilustrować tę ideę, przykładowa sekwencja w pojedynczym dniu modalnym pacjenta została podana w Tabeli 7. Funkcje G i I_P odpowiadają odpowiednio dyskretyzacji zmiennych G (pomiar glikemii) i I_P oraz I_B (podanie insuliny).

Patient ID	**Date**	L	**Functional events**
584	2009-03-10	t_1	I_B(basal insulin injection) = b_3
584	2009-03-11	t_2	G(night0) = *normoglycemia*
584	2009-03-11	t_3	G(night3) = *normoglycemia*
584	2009-03-11	t_4	G(night5) = *normoglycemia*
584	2009-03-11	t_5	G(before breakfast) = *normoglycemia*, t(night) = *normoglycemia*
584	2009-03-11	t_7	I_P(*breakfast*) = p_3
584	2009-03-11	t_8	G(*after breakfast*) = *hyperglycemia*
584	2009-03-11	t_9	I_P(second *breakfast*) = p_3

Patient ID	Date	L	Functional events
584	2009-03-11	t_{10}	$\mathrm{G}(before\ lunch) = hyperglycemia$
584	2009-03-11	t_{11}	$\mathrm{I_P}(lunch) = p_3$
584	2009-03-11	t_{12}	$\mathrm{G}(after\ lunch) = hyperglycemia$

Tabela 7. Przykładowa sekwencja zdarzeń funkcyjnych w obrębie jednego dnia terapeutycznego

Wzorce oparte na szablonach sekwencji

Kolejnym krokiem budowy modelu leczenia było zaproponowanie wzorców opartych na szablonach definiowanych przez lekarzy i wykorzystywanych w celu odkrywania wzorców sekwencji ukierunkowanych na osiągnięcie celu terapeutycznego [A8]. Ideą szablonów jest wprowadzenie ogólnego rozwiązania problemu reprezentacji ograniczeń medycznych nałożonych na sekwencje, które chcemy odkrywać. Na przykład może być wymagane, aby sekwencja zaczynała się od pomiaru glukozy w określonym czasie, następnie rozważana jest wartość dawki insuliny, po której następuje weryfikacja jej wpływu na organizm pacjenta.

Szablon T definiujemy jako sekwencja termów T = <T_1, T_2, ..., T_n>, gdzie: T_i to symboliczny term opisujący odpowiednie podzbiory zdarzeń funkcyjnych, $n = \mathrm{card}(T)$ to długość szablonu. Każdy term T_i jest interpretowany przez funkcję semantyczną δ jako podzbiór zdarzeń, tj. $\delta(T_i) \subseteq Z$, gdzie Z jest zbiorem wszystkich możliwych zdarzeń funkcyjnych. Jedyną wzajemną zależnością, jaka jest zakładana między T_1, T_2, ..., T_n, jest całkowity porządek odzwierciedlający czasową hierarchię zdarzeń w odpowiednich zbiorach $\delta(T_1)$, $\delta(T_2)$, ..., $\delta(T_n)$. Zakłada się, że ciąg s = <e_1, e_2, ..., e_n> pasuje do T, gdy $e_1 \in \delta(T_1)$, $e_2 \in \delta(T_2)$, ..., $e_n \in \delta(T_n)$, takie dopasowanie jest oznaczane jako $s \preccurlyeq T$. Na przykład zakładając wymaganą funkcję jako $F_i = \mathrm{I_P}$ przykładowy element szablonu można opisać jako T = ($\mathrm{I_P}$(*breakfast*) =?). Interpretacja jest taka, że w poszukiwanej sekwencji powinna się pojawić dawka insuliny przed śniadaniem. Odpowiednio, wsparcie dla sekwencji α w odniesieniu do szablonu T oraz sekwencji zdarzeń j-tego pacjenta jest liczone w następujący sposób:

$$\mathrm{tsup}(\alpha, s^j) = \frac{\mathrm{card}(\{w \in W(s^j) | \alpha \preccurlyeq w \ \wedge \ \alpha \preccurlyeq T)}{\mathrm{card}(W(s^j))}$$

Następnie w [A8] zaproponowaliśmy algorytm do wydobywanie wzorców opartych na szablonach i przetestowano zaproponowane podejścia na rzeczywistych danych medycznych. W eksperymentach użyto 2 i 3 elementowe szablony. Np. w celu weryfikacji dawki insuliny doposiłkowej skonstruowano następujący szablon: T_{GPG} =<$\mathrm{G}(t_{x-1})$ =g, $\mathrm{I_P}(t_x)$ =p, $\mathrm{G}(t_{x+1})$ =?>, gdzie $t_x \in$ {*breakfast, second breakfast, lunch, dinner, super*} $\subset L$ oznacza etykietę opisujący rozważany posiłek.

Po przeprowadzeniu licznych prób okazało się, że minimalna długość terapii, dla której skuteczność predykcji była rozsądna (ze wskaźnikiem powodzenia powyżej 0.3), wynosi co najmniej 6 dni modalnych. Predykcję poziomu glikemii po podaniu insuliny doposiłkowej przyjęto jako skuteczną w przypadku, gdy g' = g, dla g' i g oznaczających odpowiednio wartości wyznaczone i rzeczywiste. Wskaźnik jakości predykcji obliczono jako liczbę skutecznych predykcji podzielone przez liczbę wszystkich wykonanych predykcji, np. dla etykiety *breakfest* otrzymano 15/27 = 0.56 co oznacza, że dla 27 wykonanych predykcji 15 było skutecznych. Jak można zauważyć, zadowalający poziom skuteczności uzyskano tylko dla okresów opisujących pierwsze i drugie śniadanie (*breakfast* oraz *second breakfast*). Szczegóły wyników dla różnych eksperymentów zostały zaprezentowane w pracy [A8].

Rozwiązanie 2

Jak wynika z poprzedniej pracy wsparcie dla dłuższych wzorców sekwencyjnych jest stosunkowo niskie. Prawdopodobnie wynika to ze zmienności odpowiedzi organizmu i wpływie innych czynników na przebieg leczenia dla świeżych zachorowań. Również wpływ na wyniki może mieć sposób dyskretyzacji dawek insuliny. Dlatego w kolejnej pracy [A7] zaproponowaliśmy różnicowe wzorce sekwencyjne (DSP) jako podstawa do systemu wspomagania decyzji lekarza w ustaleniu terapii dla świeżych zachorowań na cukrzycę typu I. Główną ideą wprowadzenia DSP jest wspomaganie lekarza w podjęciu decyzji o zmianie dotychczasowego leczenia (zwiększeniu lub zmniejszeniu dawki insuliny). Uwzględniono też jeden z ważniejszych czynników wpływających na dawkę insuliny to jest wagę pacjenta. Bazowa dawka insuliny jest przeliczona na 10 kg masy ciała pacjenta i zaokrąglona do pełnych wartości w celu generalizacji wartości tego współczynnika. W szczególności w pracy [A7] analizowano przebieg leczenia pacjentów w nocy i zdefiniowano wartość glikemii nocnej w następujący sposób:

$$\begin{cases} L(night) = normoglycemia \mid \bigwedge\limits_{l\epsilon\{l_2,l_3,l_4,l_5\}} u_l = 2\ (normoglycemia) \\ L(night) = hypoglycemia \mid \bigvee\limits_{l\epsilon\{l_2,l_3,l_4,l_5\}} u_l = 1\ (hypoglycemia) \\ L(night) = hyperglycemia \mid \bigvee\limits_{l\epsilon\{l_2,l_3,l_4,l_5\}} u_l = 4 \wedge \bigwedge\limits_{l\epsilon\{l_2,l_3,l_4,l_5\}} u_l \neq 1 \\ w\ przeciwnym\ razie \quad L(night) = mild\text{-}hyperglycemia \end{cases}$$

Można zauważyć, że po wprowadzeniu abstrakcji dla glikemii nocnej i insuliny bazowej przebieg leczenia dla nocy będzie składał się z sekwencji dwóch zdarzeń; podanie insuliny, wartość glikemii nocnej.

Zdefiniujmy uporządkowaną parę $s_i = <z_i, c_i>$, gdzie z_i oznacza wartość podstawowej dawki insuliny, a c_i wartość nocnej glikemii. W ten sposób każda sekwencja $s = <s_1, s_2, \ldots, s_n>$ jest związana z terapią bazową insuliną u danego pacjenta. Każdy zestaw elementów w takiej sekwencji jest uporządkowaną parą zdarzeń związanych z nocną terapią. Celem jest odkrycie częstych wzorców $p \subset s$, które z dużym prawdopodobieństwem powtórzą się u nowo przyjętego pacjenta.

Jak wyżej wspomniano terapia jest często rozważana przez lekarzy jako zmiana w stosunku do poprzednio ustalonej i obserwowanej w poprzednich dniach. Aby odzwierciedlić tę procedurę, załóżmy, że dawka insuliny i poziom glikemii są oznaczone odpowiednio jako *a* i *b* w dniu przyjęcia pacjenta; w ten sposób $s_1 = <a, b>$, gdzie *a* i *b* są stałymi. Kolejna dawka insuliny (następnego dnia) może zostać zwiększona (oznaczona jako „+"), zmniejszona (oznaczona jako „-") lub może zostać utrzymana na tym samym poziomie (oznaczona jako „0"). Podobną operację wykonuje się przy poziomie glikemii. Dla $i \geq 2$ mamy:

$$d_i = \begin{cases} -, \text{dla } z_i < z_{i-1} \\ 0, \text{dla } z_i = z_{i-1} \\ +, \text{dla } z_i > z_{i-1} \end{cases} \quad g_i = \begin{cases} -, \text{dla } c_i < c_{i-1} \\ 0, \text{dla } c_i = c_{i-1} \\ +, \text{dla } c_i > c_{i-1} \end{cases}$$

Sekwencje reprezentujące przebieg leczenia są więc redefiniowane do sekwencji różnicowych dla każdego pacjenta. Na przykład sekwencja S = <<3, 3> , <3 , 2> , <3, 1> , <2, 2>> jest konwertowana na sekwencję różnicową D = <<a, b> , <0, −> , <0, −> , <−, +>> lub prościej przy pominięciu nawiasów D = <*a*, *b*, d0, g−, d0, g−, d−, g+> - gdzie d oznacza dawkę insuliny, a g oznacza glikemię.

W pracy [A7] zaproponowano algorytm wydobywający różnicowe wzorce sekwencyjne i przedstawiono sposób ich użycia w praktyce lekarskiej. W części eksperymentalnej przedstawiono

wyniki wyznaczenia DSP dla rzeczywistych danych klinicznych. Jednym z eksperymentów była ocena przydatności wydobytych DSP dla nowo przyjętych pacjentów przez 5-krotną walidację krzyżową z losowym próbkowaniem. Zbiór sekwencji pacjentów podzielono na pięć części. W każdej próbie, wybrana część służyła jako zbiór testowy, a pozostałe sekwencje stanowiły zbiór uczący. Ze zbioru uczącego wydobyto wzorce różnicowe, a następnie obliczono ich wsparcie w zbiorze testowym. Dla 89% przykładów testowych w każdej walidacji były dostępne wzorce z minimalnym poziomem wsparcia większym lub równym 0,2 i o długości 2,5 dnia lub więcej. Podsumowując poprzez zastosowanie wzorców różnicowych ułatwiono lekarzowi podejmowanie decyzji w zakresie zmian w prowadzonej terapii. Dodatkowo uzyskano bardzo interesujące z punktu widzenia lekarza informacje dotyczące przebiegu leczenia – przykłady częstych wzorców różnicowych podano w Tabeli 8.

DSP	**Support**
<a, b, d0, g0, d0>	0,39
<a, b, d0, g−, d−>	0,24
<a, b, d0, g+, d0>	0,23
<a, b, d0, g0, d0, g0>	0,16
<a, b, d0, g0, d0, g−>	0,15

Tabela 8. Przykładowe DSP wydobyte z danych historycznych

Rozwiązanie 3

W kolejnej pracy [A6] zaproponowaliśmy nowe podejście do automatycznego generowania zaleceń medycznych (ang. medical gudelines). Przede wszystkim zostały uwzględnione dodatkowe dane, które mogą mieć wpływ na dawkowanie insuliny i są brane pod uwagę przez lekarza przy ustalaniu inulino-terapii. Jak opisano we wprowadzeniu (punkt 4.1) poza wagą pacjenta brany jest pod uwagę stan kliniczny pacjenta; występowanie stanu zapalnego (CRP) czy kwasicy ketonowej (PH) oraz wciąż własne wydzielanie insuliny (C-peptyde). Dane te są czasami nazywane „statyczne” jako że się znacząco wolniej zmieniają niż dane opisujące przebieg leczenia – nazywane w pracach „dynamicznymi”. Zestaw danych statycznych został przedstawiony w Tabeli 9.

Atrybut	**Znaczenie medyczne**
Wiek	The patient age at onset
Płeć	Male (1) or female (0)
Waga	The weight at onset
C-peptyde	Insulin secretion
CRP	Certificate of infection, 1 (high) or 0 (in norm range)
PH	ACID based balance

Tabela 9. Statyczne dane pacjenta

W pracy [A6] i kolejnych [A4, A3, A2] zaproponowano hybrydowe podejście do wygenerowania zaleceń medycznych. W pierwszej kolejności, statyczne dane dotyczące pacjentów zostały poddane grupowaniu. W ten sposób uzyskano grupy pacjentów o podobnej charakterystyce medycznej. W dalszej kolejności, dla każdego pacjenta określonej grupy zaproponowano model przebiegu leczenia.

Do grupowania pacjentów wykorzystano algorytm fuzzy C-means [31], a liczba klastrów została określona eksperymentalnie [A4] oraz w pracach [A3, A2] na podstawie kryterium Calinskiego-Harabasza (CH) [29] i indeksu Xie-Bieni [30]. Stabilność grupowania została zweryfikowana poprzez zmianę losowej inicjalizacji centroidów w kilku cyklach (Rand indeks wyniósł 0.97). Podział na grupy pacjentów na podstawie danych statycznych okazał się bardzo dobrym pomysłem – można zauważyć,

że pacjenci wewnątrz grup są podobnie leczeni. Ponadto grupy te można wygodnie scharakteryzować medycznie na podstawie centroidów [A3].

W pracy [A6] analizowano leczenie nocne, w pierwszym eksperymencie szukano wzorców dotyczących dawek insuliny bazowej a w kolejnym uwzględniono również wszystkie nocne pomiary glikemii. Krótsze wzorce zwłaszcza w ramach pierwszego eksperymentu mają bardzo obiecujące poziomy wsparcia i mogą służyć jako wskazówki do przebiegu leczenia dla nowych pacjentów co zostało potwierdzone na danych rzeczywistych z pomocą walidacji krzyżowej.

Rozwiązanie 4

Następnie w pracy [A4] zaproponowaliśmy modelowanie przebiegu leczenia insuliną doposiłkową dla danej grupy pacjentów. Zauważono, że lekarz często ocenia skuteczność leczenia dla każdego posiłku oddzielnie. Zdefiniowano więc trzy elementową sekwencję zdarzeń <<G=v1; τ_1><I_p = v2; τ_2><G= v3, τ_3>> gdzie $\tau_1 < \tau_2 < \tau_3$, która opisuje zdarzenia: pomiar glikemii, podanie dawki insuliny i powtórny pomiar glikemii po posiłku (zazwyczaj po ok 2h). Zakładamy, że τ_1, τ_2, τ_3 należą do tego samego przedziału czasu $t \in \mathrm{L}$ (Tabela 10).

Oznaczenie przedziału L	**Opis**	**Przedział czasu**
t_1	breakfast	[7:00 - 11:00]
t_2	second breakfast	[10:00 -14:00]
t_3	lunch	[13:00 - 17:00]
t_4	dinner	[16:00 - 20:00]
t_5	supper	[19:00 - 23:00]

Tabela 10. Symboliczna skala czasu

Na tej podstawie definiujemy zdarzenie złożone jako <tv1v2v3; k>, gdzie ciąg połączonych symboli tv1v2v3 jest etykietą zdarzenia (zgodnie z definicją zdarzenia), a k jest znacznikiem czasu, w którym zdarzenie ma miejsce, k mapujemy do symbolicznej skali czasu L. Dla uproszczenia, w przypadku, gdy zdarzenia złożone występują w sekwencji, jeden po drugim, wartości zmiennej związanej z symboliczną skalą czasu zostają pominięte. Warto podkreślić, że dane są dyskretyzowane; glikemia zgodnie z Tabelą 3, a dawka insuliny jest odniesiona do wagi pacjenta (100kg) i wielkości posiłku (100kcal) – otrzymany współczynnik jest zaokrąglany do pełnej wartości.

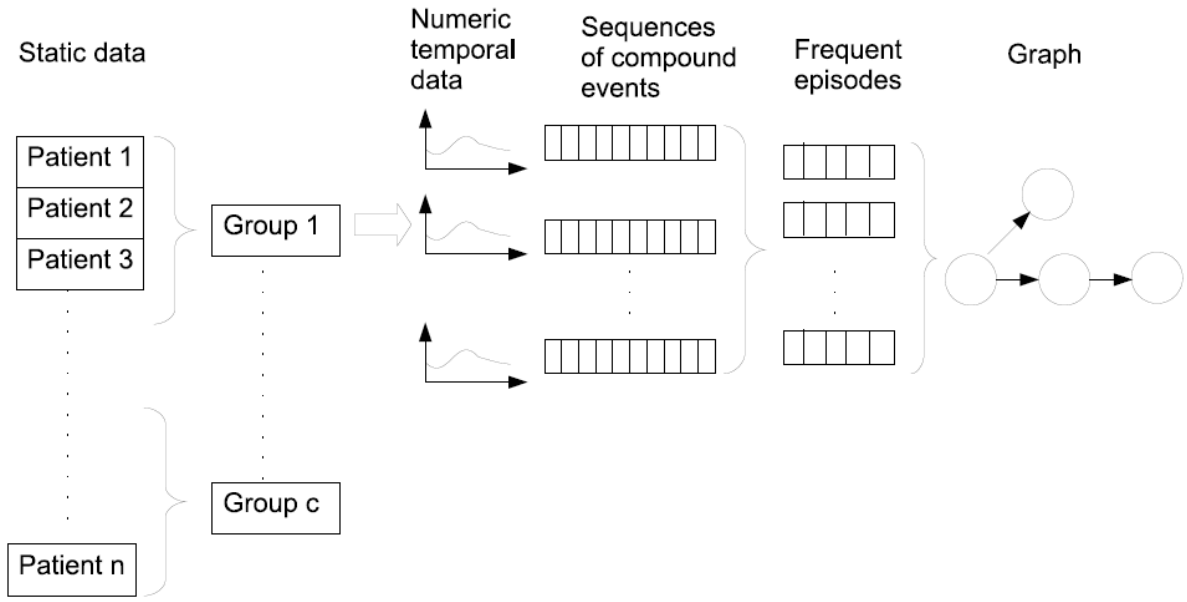
Ponadto dla wygody lekarza wprowadziliśmy miary epizodu α; confidence $conf(\alpha)$ – która ocenia stosunek liczby normoglycemii do wszystkich wyników glikemii w epizodzie oraz $MGCR(\alpha)$ która ocenia średni wskaźnik wyrównania glikemii.

$$conf(\propto) = \frac{1}{2 \cdot card(\propto)} \left(\sum_{k=1}^{card(\propto)} \left(v_{1,k} \middle| v_{1,k} = 2\right) + \sum_{k=1}^{card(\propto)} \left(v_{3,k} \middle| v_{3,k} = 2\right)\right)$$

$$MGCR(\propto) = 1 - \frac{1}{2 \cdot card(\propto)} \sum_{k=1}^{card(\propto)} \left(\frac{\left|v_{1,k} - 2\right|}{maxDev\left(v_{1,k}\right)} + \frac{\left|v_{3,k} - 2\right|}{maxDev(v_{3,k})}\right)$$

Funkcje $maxDev(v_{1,k})$, $maxDev(v_{3,k})$ reprezentują wartości oznaczające maksymalne możliwe odchylenia glikemii od wartości normalnej dla danego pacjenta.

Zaproponowane podejście [A4] do wygenerowania wskazówek dla lekarza można przedstawić z pomocą Rysunku 1.



Rysunek 1. Proponowane podejście

W pierwszym kroku dane statyczne pacjentów są grupowane. W ten sposób uzyskujemy grupy pacjentów o podobnych cechach. Dla każdej grupy pacjentów bierzemy pod uwagę dane z przebiegu leczenia. Dane te są dyskretyzowane i reprezentowane w symbolicznej skali czasu. Wynikiem tej operacji są sekwencje zdarzeń elementarnych, a sekwencje te są wykorzystywane do konstruowania sekwencji zdarzeń złożonych. W drugim kroku naszego podejścia identyfikujemy częste epizody z sekwencji złożonych zdarzeń. W tym celu wykorzystano nowo zaproponowany algorytm. Uzyskane epizody są wyświetlane graficznie, a graf jest prezentowany lekarzom wraz z listą częstych epizodów. Zestaw częstych epizodów jest oceniany za pomocą wcześniej zaproponowanych miar $MGCR(\alpha)$ oraz $conf(\alpha)$.

Zaproponowane podejście zostało zaimplementowane i opisane algorytmicznie przez Algorytm 1 oraz Algorytm 2.

Algorithm 1: Generation of medical guideline

Input:
$P = \{p_1, p_2, ..., p_n\}$ - the set of vectors describing n patients by features F,
S - the set of sequences of compound events for all aptients,
c - number of clusters.
Output:
A - the set of frequent episodes.

Function GenerateGuideline(P, S, c)
$A \leftarrow \emptyset$; /* set of frequent episodes */
$C^F(P) =$ FuzzyCmeans(P, c) ; /* clustering static data */
FOR EACH$(cluster \in C^F(P))$ {
  $A_{cluster} =$ FrequentEpisodes$(S_{cluster})$; /* discovery of frequent episodes */
  $A \leftarrow A + A_{cluster}$; /* accumulation of frequent episodes */
}
return A ;

Dane wejściowe do algorytmu to: zbiór wektorów opisujących pacjentów, zbiór sekwencji zdarzeń reprezentujących dane dotyczące leczenia oraz liczba skupień.

Po wykonaniu grupowania metodą fuzzy c-means otrzymujemy zbiór skupień $C^F(P)$ w zbiorze pacjentów P. Dla każdego klastra w obrębie $C^F(P)$ rozważamy zestaw sekwencji $S_{cluster}$ z przebiegu leczenia. W zbiorze sekwencji $S_{cluster}$ identyfikowane są częste epizody.

Wynikiem działania funkcji $\mathrm{FrequentEpisodes}(S_{cluster})$ jest zestaw częstych epizodów specyficznych dla klastra. Ostatecznym wynikiem Algorytmu 1 jest zbiór A wszystkich zidentyfikowanych częstych epizodów. Funkcja $\mathrm{FrequentEpisodes}(S_{cluster})$ zwraca częste epizody z zestawu sekwencji w każdym klastrze. Poniżej zostaną przedstawione szczegóły tej funkcji, podane jako Algorytm 2.

Algorithm 2: Mining frequent episodes from treatment data

Global parameters:
$wsize$ - the size of the window,
$step$ - the step according to which the events are collected,
m_{start} - starting event,
sup_{min} - minimum acceptable support of the episode

Input: $S_{cluster}$ - set of sequences corresponding to the cluster c,
Output: $A_{cluster}$ - set of frequent episodes

Function FrequentEpisodes($S_{cluster}$)
$A_{cluster} = \emptyset$; /*no events in set A_c */
$episodes[] = \emptyset$; /*a map that is storing the number of episodes occurrences */
FOR EACH($s \in S_{cluster}$)
  FOR EACH($w \in \mathcal{W}(s)$) { /* for each window in the sequence */
    $\alpha = \emptyset$;
    $m = m_{start}$;
    WHILE ($m \leq wsize$) {
      $k = m$;
      WHILE ($k \leq wsize$) { /* look for s sub-sequences */
        $\alpha = \alpha + e(k)$; /* add next event */
        $episodes[\alpha] = episodes[\alpha] + 1$; /* count the no. of occurrences */
        $k = k + step$;
      }
      $m = m + step$; /* move to next event */
    }
  }
FOR EACH(α)
  IF ($episodes[\alpha]/card(\mathcal{W}(S_{cluster})) > sup_{min}$) {
    $A_{cluster} = A_{cluster} + \alpha$;
    Calculate $conf$ and $MGCR$ for each frequent episode
  }
return $A_{cluster}$;

Na początku definiujemy zestaw parametrów globalnych. Parametry te są podawane przez lekarzy przed zastosowaniem algorytmu. Ustawiając $wsize$ i $step$, można ustawić różne strategie zarządzania rozmiarem okna i selekcją zdarzeń w oknie. Ustawiając m_{start}, lekarz ustala zdarzenie początkowe (posiłek), które jest brane pod uwagę przy wykrywaniu częstych epizodów. Ostatni parametr, sup_{min}, określa minimalne wsparcie wymagane dla wzorców.

Zbiór sekwencji $S_{cluster}$ stanowi dane wejściowe funkcji $\mathrm{FrequentEpisodes}$.

Dla każdego okna w sekwencji $s \in S_{cluster}$, epizod α jest konstruowany poprzez rozpoczęcie od zdarzenia wskazanego przez m_{start} i dodanie kolejnych zdarzeń z danym krokiem. Następnie algorytm powtarza konstrukcję epizodu α zaczynając od kolejnych zdarzeń. Dla każdego epizodu α, liczba jego

wystąpień jest przechowywana na liście *episodes*[]. Liczba wystąpień na liście jest zwiększana, jeśli epizod α zostanie znaleziony w innym oknie lub w innej kolejności.

Na koniec algorytm oblicza wsparcie epizodów na podstawie informacji z listy i łącznej liczby okien, tj.: $episodes[\alpha]/\mathrm{card}(W(s))$. Jeśli poziom wsparcia jest większy niż sup_{min}, epizod α jest dodawany do zestawu częstych epizodów dla klastra. Ponieważ każdy epizod zawiera wartość $m \in L$, wsparcie można również obliczyć dla epizodów rozpoczynających się zdarzeniem związanym z konkretnym posiłkiem.

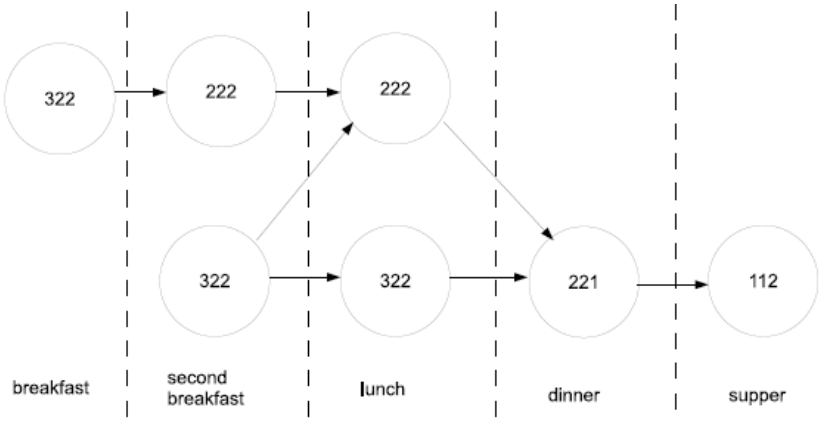
Złożoność obliczeniowa algorytmu 2 jest nie większa niż $O(n \cdot h^2 \cdot \mathrm{card}(W(s)))$, gdzie n to liczba sekwencji, h to rozmiar okna, a $\mathrm{card}(W(s))$ to liczba okien.

W pracy [A4] przedstawiono przykład liczbowy, który szczegółowo wyjaśnia działanie algorytmu.

Zastosowanie modelu w praktyce klinicznej wygląda w następujący sposób:

a) Nowo przyjęty pacjent do szpitala zostaje przydzielony do jednego z odkrytych skupień. Wykorzystywany jest w tym celu leniwy klasyfikator oparty na odległości do centroidów. Ponadto lekarzom przedstawiane są wartości funkcji przynależności wektora cech statycznych do każdej z grup. Dzięki temu lekarze mogą ocenić, w jakim stopniu każda grupa jest reprezentatywna dla danego pacjenta.
b) Następnie lekarzom przedstawiane są częste epizody i związany z nimi grafy uzyskane dla wybranej grupy jako ścieżki leczenia dla danego pacjenta.
c) Po pewnym czasie, możliwy przebieg terapii przedstawiany lekarzowi może zostać ograniczony. Jako zalecenia medyczne służą jedynie te częste epizody, które pasują do ustalonego już początkowego przebiegu terapii.

W pracy [A4] przedstawiono wyniki eksperymentów przeprowadzonych w oparciu o rzeczywiste dane kliniczne. Na podstawie danych statycznych pacjenci zostali podzieleni na 6 grup. W pierwszym eksperymencie okno obejmowało cały przebieg leczenia. Przyjmując to założenie, każdy epizod może wspierać tylko jedną sekwencję przypisaną pacjentowi. Natomiast w drugim eksperymencie ustawiliśmy wielkość okna na jeden dzień terapeutyczny. W obu przypadkach częste epizody składały się z 3-4 zdarzeń złożonych, a najwyższe poziomy wsparcia wynosiły od 0,11 do 0,5 w zależności od grupy pacjentów. Przedstawiono też wartości oceny wzorców oraz przykładowe grafy wizualizujący przebieg leczenia (Rysunek 2).



Rysunek 2. Przykładowy graf przedstawiający możliwe ścieżki leczenia

Podejście przedstawione powyżej zostało bardzo dobrze ocenione przez lekarzy. Dzięki grupowaniu pacjentów wskazówki leczenia są dobrane do odpowiedniej grupy – chorych którzy są podobnie

leczeni. Częste epizody obrazują bardzo prawdopodobny przebieg leczenia pacjenta oraz występujące alternatywne przebiegi. Lekarz może być przygotowany na różne scenariusze i rozważać konsekwencje zastosowania danej terapii. Szczególnie pomocna okazała się wizualizacja potencjalnych ścieżek leczenia pacjenta w postaci grafu.

Rozwiązanie 5

Możliwość analizy ścieżek leczenia z pomocą grafu okazała się bardzo przydatna dla lekarzy przy planowaniu terapii, dlatego postanowiliśmy rozwinąć ten pomysł w kolejnej pracy [A3]. Główna idea jest taka, że rezygnujemy z wyznaczania częstych epizodów, natomiast generalizujemy ścieżki leczenia grupując je w postaci ścieżek grafu i przedstawiamy dodatkowe informacje pozwalające lekarzowi ocenić prawdopodobieństwo wystąpienia. W systemie wspomagania decyzji wprowadzamy też możliwość filtrowania, tak aby pokazać te ścieżki, które są najbardziej reprezentatywne dla danego pacjenta albo np. te które prowadzą do niepożądanych zdarzeń (hypo/hiper glikemii) [A2]. Patrząc na problem szerzej badania dotyczą zagadnienia znanego w literaturze jako wydobywanie ścieżek leczenia (mining clinical pathways) z surowych danych klinicznych (electronic health records) [32][40][41].

Zgodnie z definicją Australian Queensland Health Board, ścieżka leczenia opisuje „standaryzowany, oparty na dowodach plan medyczny, który określa odpowiednią sekwencję interwencji klinicznych, ramy czasowe, kamienie milowe i oczekiwane wyniki dla jednorodnej grupy pacjentów". Według tej samej organizacji, głównym celem ścieżki klinicznej jest „wspieranie praktyki opartej na dowodach, ulepszanie procesów klinicznych poprzez zmniejszanie ryzyka i wreszcie zmniejszenie zmienności w dostarczaniu usług zdrowotnych".

Medical Treatment Graph (MTG)

Najważniejszym elementem rozwiązania jest graf modelujący przebieg leczenia. MTG definiujemy jako skierowany, acykliczny graf: $MTG = (N, E, \sigma, \omega)$, gdzie N jest zbiorem węzłów, $E \subseteq N \times N$ jest zbiorem krawędzi reprezentujących zależność między dwoma węzłami. Funkcje $\sigma : N \rightarrow [0, 1]$ and $\omega: E \rightarrow [0, 1]$ przypisują wagę do każdego węzła i każdej krawędzi w postaci liczby rzeczywistej.

Poniżej wyjaśniamy semantykę MTG.

Rozważmy $U_t \subset U$ jako podzbiór zdarzeń występujących w czasie $t \in L$ (w dyskretnej skali czasu związanej z dniem terapeutycznym pacjenta). W ramach U_t wyróżniamy podzbiór tych zdarzeń $N_{tj} \subset U_t$ wyznaczonych przez konkretną zmienną i jej wartość. Oznacza to, że wszystkie $u \in N_{tj}$ odnoszą się do tej samej zmiennej G lub I zakładając pewną stałą wartość glikemii lub insuliny. Zakładamy, że $N_{tj} \in N$ jest węzłem MTG, gdzie indeks *t* odnosi się do okresu dziennej terapii, a indeks *j* do unikalnej pary rozpatrywanej zmiennej i jej wartości. Oznacza to, że zbiór N_{tj} zawiera podobne zdarzenia, czyli takie, które występują w tym samym okresie doby terapeutycznej i dodatkowo odnoszą się do tej samej zmiennej i wartości.

Niech $\sigma(N_{tj}) = \frac{card(N_{tj})}{card(U_t)}$ wyznacza prawdopodobieństwo wystąpienia zdarzenia z N_{tj} w grupie zdarzeń z U_t. Wartość $\sigma(N_{tj})$ pełni rolę wagi węzła N_{tj} w ramach MTG.

Rozważmy teraz wzajemne zależności między węzłami. Zdefiniujmy krawędź MTG jako parę uporządkowaną $E_{tjk} = (N_{tj}, N_{(t+1)k})$, gdzie N_{tj}, $N_{(t+1)k}$ to węzły związane ze zbiorami zdarzeń występujących w czasie *t* i *t* + 1, odpowiednio.

Niech $S_t \subset S$ będzie zbiorem najkrótszych możliwych podciągów składających się tylko z dwóch kolejnych zdarzeń u_t, u_{t+1}. Odróżnijmy od S_t te ciągi S'_t, które pasują do danej pary sąsiednich węzłów z MTG. Definiujemy zbiór $S'_t = \{u_t, u_{t+1}\} | u_t \in N_{tj}, u_{t+1} \in N_{(t+1)k}$. Wówczas $\omega(E_{tjk}) = \frac{card(S'_t)}{card(S_t)}$ wyznacza prawdopodobieństwo wystąpienia kolejnych zdarzeń w sekwencji zdarzeń klinicznych. Funkcja ω przypisuje wagę do krawędzi E_{tjk} w MTG.

Aby rozszerzyć możliwość interpretacji alternatywnych ścieżek leczenia w ramach MTG, wprowadzamy współczynnik pewności. Dla krawędzi MTG definiujemy $cer(E_{tjk}) = \frac{\omega(E_{tjk})}{\sigma(N_{tj})}$. Należy zwrócić uwagę, że współczynnik pewności opisuje rozkład prawdopodobieństwa zdarzeń wzdłuż krawędzi rozpoczynających się w danym węźle.

Załóżmy teraz, że p = [p_1, p_2, ..., p_n] jest dowolną ścieżką w obrębie MTG, gdzie p_i jest wybranym węzłem MTG, a $p_i \in N$, $1 < n \leq \mathrm{card(L)}$. W tym przypadku indeks i wskazuje miejsce węzła w ścieżce. Należy zauważyć, że p jest ścieżką kliniczną zgodną z definicjami podanymi na początku Rozwiązania 5.

Rozszerzając tak aby dokonać oceny dowolnej ścieżki w ramach MTG, otrzymujemy $\omega(p) = \sigma(p_1) \cdot \prod_{i=1}^{n-1} \frac{\omega(p_i, p_{i+1})}{\sigma(p_i)}$, a wskaźnik pewności dla ścieżki obliczamy jako $cer(p) = \prod_{i=1}^{n-1} cer(p_i, p_{i+1})$.

Korzystając z funkcji $\omega(p)$ i cer(p), lekarze mogą ocenić pewność dowolnej ścieżki w obrębie MTG. Mogą również odfiltrować z MTG te ścieżki, które są mniej prawdopodobne, tj. te związane np. z rzadkimi przypadkami medycznymi. Zakładając, że ω_{min} jest wartością progową podaną przez lekarzy, można stworzyć podgraf: MTG' = (N', E', ω, σ) dla którego $\omega(p) > \omega_{min}$ dla dowolnego p.

Algorytm

Ważnym elementem rozwiązania jest zaproponowany algorytm opisujący tworzenie MTG z sekwencji danych klinicznych.

Algorytm przeszukuje listę sekwencji zdarzeń medycznych. Każde zdarzenie w sekwencji jest kandydatem na węzeł w grafie, a każda para zdarzeń jest kandydatem na krawędź grafu. Zostaną dodane do kolekcji węzłów i krawędzi, jeśli nie są jeszcze zarejestrowane w grafie.

Najpierw w wierszach 2 i 3 algorytm inicjuje kolekcje N i E, które służą do przechowywania odpowiednio węzłów i krawędzi MTG.

Później algorytm iteruje przez sekwencje kliniczne (pętla zaczyna się w wierszu 4) i zdarzenia w nich zawarte (pętla zaczyna się w wierszu 5). Sekwencje kliniczne podawane są jako dane wejściowe do algorytmu w postaci tablicy S. Pierwszy indeks tej tablicy, oznaczony przez j, odnosi się do rozważanej sekwencji, a drugi, oznaczony jako i, wskazuje i-te zdarzenie w j-tej sekwencji i odnosi się do czasu τ. Sekwencjonowanie czasu następuje w linii 8.

Następnie algorytm przeszukuje kolekcje N i E, sprawdzając, czy zawierają one określone zdarzenie (linia 12) i krawędź (linia 16) wykryte w j-tej sekwencji. Ta pętla zaczyna się w wierszu 11.

Jeśli węzeł lub krawędź zostanie znaleziona w MTG, algorytm zwiększa powiązane liczniki NCount i ECount (linie 14 i 18). W przeciwnym razie węzeł lub krawędź jest dodawany do odpowiednich kolekcji (linie 22 i 26).

Wreszcie, w wierszach 31-36, algorytm iteruje przez skonstruowany MTG, aby obliczyć „cer" i ω.

Algorithm 3 Constructing the MTG.

Require: S—set of clinical sequences, w—number of clinical sequences

Function GraphBuild(S, w)
N = null; NCount $\leftarrow 0$; {a collection of nodes}
E = null; ECount $\leftarrow 0$; {a collection of edges}
$l = 1$;
for $j = 1$ **to** w **do** {for each sequence}
  for $i = 1$ **to** length($S[j]$) **do** {for each event}
    $l = l + (i \mod \text{card}(T))$;{determine the offset}
    $t = t(\tau_i)$;
    node $= S[l][t]$; {create a node}
    edge $= \langle \text{node}, S[l][t+1] \rangle$; {create an edge}
    N_{exists} = false; {lacking node}
    E_{exists} = false; {lacking edge}
    for $k = 1$ **to** l **do** {for the added nodes}
      if $N[k][t] ==$ node **then**
        {Is the node added?}
        NCount$[k][t]$++; N_{exists} = true; {a number of nodes}
      end if
      if $E[k][t] ==$ edge **then**
        {Is the edge added?}
        ECount$[k][t]$++; E_{exists} = true; {a number of edges}
      end if
    end for
    if not N_{exists} **then**
      $N[l][t]$ = node; NCount$[l][t] = 1$;
      {adding the node}
    end if
    if not E_{exists} **then**
      $E[l][t]$ = edge; ECount$[l][t] = 1$;
      {adding the edge}
    end if
  end for
end for
for $j = 1$ **to** l **do**
  for $i = 1$ **to** card$(T) - 1$ **do**
    $\sigma[j][i] = \text{NCount}[j][i] / \sum_{k=1}^{l} (\text{NCount}[k][i])$
    $\omega[j][i] = \text{ECount}[j][i] / \sum_{k=1}^{l} (\text{ECount}[k][i])$
    $\text{cer}[j][i] = \text{ECount}[j][i] / \text{NCount}[j][i]$
  end for

pierwszej został przedstawiony na Rysunku 3.

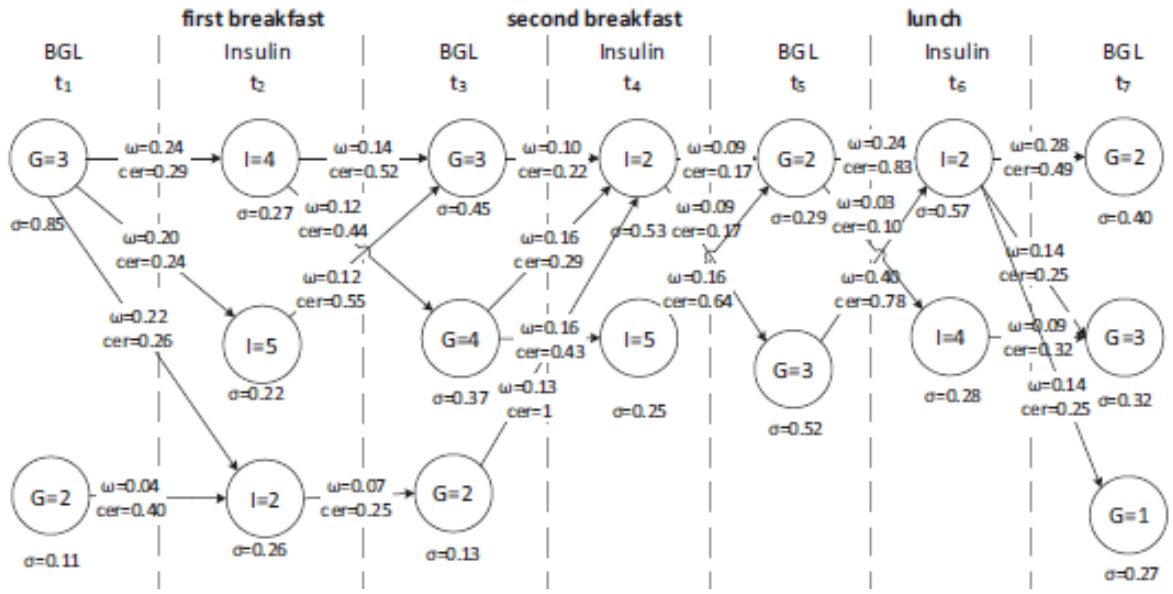
Zauważmy, że algorytm ma liniową złożoność obliczeniową ze względu na liczbę sekwencji pacjenta i liczbę zdarzeń w sekwencji.

Zastosowanie MTG

Zastosowanie naszej metody w praktyce klinicznej można przedstawić w następujących krokach.

a) Podobnie jak poprzednio, po przyjęciu do szpitala pacjent zostaje przydzielony do jednej z rozważanych grup (klastrów). Ostatecznego przypisania do grupy dokonuje lekarz na podstawie przedstawionego stopnia przynależności oraz wiedzy medycznej.
b) Lekarzowi przedstawiany jest MTG odpowiedni dla klastra, do którego pacjent należy. W zależności od aktualnego stanu pacjenta lekarz może odfiltrować z MTG mniej prawdopodobne ścieżki, tworząc w ten sposób podgraf. Filtrowanie jest zwykle wykonywane kilka razy, co pozwala lekarzowi na analizę alternatywnych ścieżek.
c) Lekarz interpretuje otrzymane MTG i na tej podstawie konstruuje plan terapii.
d) W trakcie terapii lekarz konfrontuje aktualny stan pacjenta z powiązaną częścią MTG. Na tej podstawie lekarz dostosowuje leczenie.

Zaproponowane rozwiązanie zostało przeanalizowane i sprawdzone na rzeczywistych danych klinicznych pacjentów chorych na cukrzycę typu 1. Pacjenci zostali podzielenie na grupy na podstawie danych statycznych i dla każdego klastra został wygenerowany graf MTG. Fragment przykładowego grafu dla grupy



Rysunek 3. Przykład MTG (fragment dla klastra pierwszego)

Notacja na rysunku została uproszczona, aby graf był bardziej czytelny. Ścieżki w grafie zostały przefiltrowane usuwając te dla których $\omega(p) > 0.00015$. Podążając pierwszą ścieżką na wykresie, interpretujemy ją w następujący sposób; 85% pacjentów z pierwszej grupy ma łagodną hiperglikemia rano <G = 3, t1>, 29% z nich podczas pierwszego śniadania otrzymywało około 4 jednostki insuliny na 100 kcal na 100 kg masy ciała. Z kolei 40% pacjentów z normoglikemią rano <G = 2, t1> otrzymywało 2 jednostki insuliny doposiłkowej na pierwsze śniadanie.

Około 3 godziny po pierwszym śniadaniu <G = 4, t3> stwierdzono hipergligemię u około 37% pacjentów, a 0,12/0,37 = 32% z nich otrzymało wcześniej 4 jednostki insuliny <I = 4, t2> na 100 kcal i na 100 kg masy ciała.

Filtrując MTG z różnymi wartościami wagi ω można również wyznaczyć najbardziej oczekiwane ścieżki w grafie. Jest to bardzo przydatna informacja dla lekarzy, gdyż wskazuje prawdopodobną ścieżkę leczenia dla nowego pacjenta w ramach danej grupy. Gdy mamy już pierwsze informacje na temat przebiegu leczenia pacjenta, MTG może z kolei zostać ograniczone (skrócony). Dokładniej rozważmy węzeł N_{tj} oraz zbiór węzłów $N_{(t+1)*}$ połączony z nim zbiorem krawędzi E_{tj*}. Zakładając, że zdarzenie reprezentowane przez węzeł N_{tj} już wystąpiło, graf można skrócić tak, aby przedstawiał tylko kolejne ścieżki, czyli wychodzące z N_{tj}. Po skróceniu wykresu współczynniki σ dla kolejnych węzłów zostały przeliczone jako $\sigma(N_{(t+1)k}) = \omega(E_{tjk})$. W konsekwencji ω każdej krawędzi $E_{(t+1)jk}$ zostaje dostosowane proporcjonalnie do rozkładu σ. Wartości współczynnika cer oczywiście pozostają niezmienione.

Walidacja

W celu weryfikacji zaproponowanego podejścia zaproponowaliśmy procedurę oceny opartą na technice walidacji krzyżowej. Losowo podzieliliśmy wszystkie dostępne sekwencje kliniczne na zbiór treningowy zawierający 80% z nich i zbiór testowy zawierający pozostałe z nich. Dla zestawu treningowego wyznaczyliśmy 7 klastrów i wygenerowaliśmy odpowiednie MTG. Aby wyeliminować wyjątkowe sytuacje medyczne związany z danymi, przefiltrowaliśmy otrzymane MTG przy użyciu ω_{min} = 0,000006.

Na potrzeby walidacji definiujemy współczynnik dopasowania terapii $\kappa = avg_{s' \in S^c} length(s')$, gdzie S^c jest zbiorem sekwencji pacjentów z grupy testowej przypisanej do klastra c. Wyższa wartość κ wskazuje, że dłuższa sekwencja kliniczna pasuje do ścieżki leczenia w obrębie MTG.

Skupienie	1	2	3	4	5	6	7
κ	5,1	3,8	4,1	3,3	3,3	4,3	3,0

Tabela 11. Średnia wartość κ dla 5-cio krotnej walidacji krzyżowej

Dla pacjentów ze zbioru testowego obliczyliśmy κ. Wyniki eksperymentów przeprowadzonych dla każdego ze skupień przedstawiono w Tabeli 11. Podkreślmy, że otrzymane wyniki uwzględniają tylko najdłuższe ciągłe sekwencje kliniczne. Z tej perspektywy prezentowanie 3-5 kolejnych kroków terapii medycznej przez MTG można interpretować jako dobry wynik.

Ostatecznie zaproponowana metoda została też zweryfikowana przez lekarza diabetologa. Szczegóły obserwacji dla wybranych pacjentów są zaprezentowane w pracy. Główne wnioski płynące z przeprowadzonej walidacji są następujące:

- Proponowane podejście wspiera wstępną klasyfikację pacjentów do odpowiednich grup. Informacje te pomagają porównać stan pacjenta z innymi pacjentami, a tym samym znacznie ułatwiają planowanie początkowej terapii pacjenta.
- MTG pozwala lekarzom śledzić i dostosowywać decyzje medyczne dla każdego podania insuliny.
- Za bardzo przydatną lekarze uznali możliwość wizualizacji skutków zmian w terapii, np. zmniejszenia dawki insuliny do danego posiłku. Również rozkład dawek insuliny w ciągu dnia terapeutycznego można łatwiej dostosować przy użyciu MTG. Zaproponowana metoda nie pozwala jednak na dokładne oszacowanie wymaganych dawek doposiłkowych dla danego pacjenta.

Rozwiązanie 6

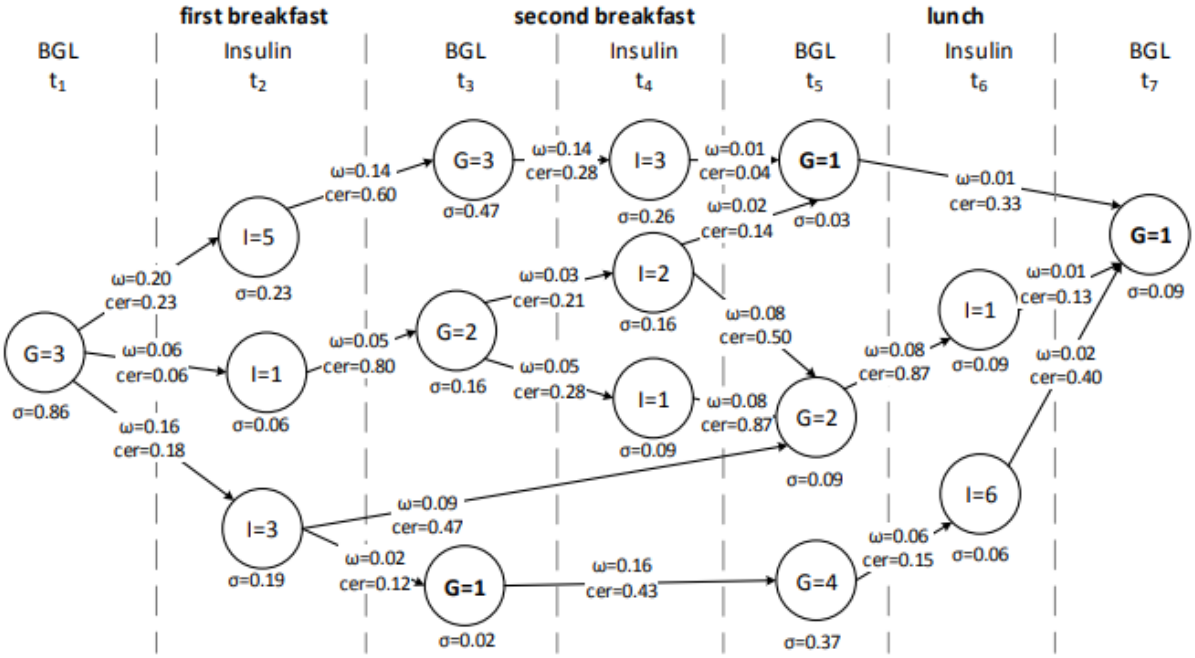
Jednym z ważniejszych problemów podczas leczenia cukrzycy, jest takie prowadzenie pacjenta, aby unikać bardzo wysokich i bardzo niskich poziomów cukru. Zwłaszcza hypoglikemia może być bezpośrednio niebezpieczna dla pacjenta – niewykryta np. w nocy może prowadzić do utraty przytomności. Dlatego szczególnie interesujące dla lekarza są te ścieżki leczenia, które prowadzą do incydentów hypo/hiper glikemii. Ścieżki takie cechują się raczej małym prawdopodobieństwem, dlatego prezentowanie ich poprzez filtrowanie MTG nie jest wygodne.

W pracy [A2] zaproponowaliśmy Rule Based Medical Treatment Graph ($RB\text{-}MTG$), który rozwiązuje ten problem. W badaniu zajmujemy się przebiegiem leczenia w obrębie jednego dnia. Przypomnijmy, że zdarzenia w obrębie dnia sekwencjonujemy wprowadzając dyskretną skalę czasu $\mathrm{L} = \{\mathrm{t}_1, \mathrm{t}_2, \ldots, \mathrm{t}_\mathrm{w}\}$, w jest liczbą rozpatrywanych okresów w ciągu dnia. Ostatecznie dla każdego pacjenta otrzymujemy zbiór ścieżek leczenia $\mathrm{S}^\mathrm{P} = \{\mathrm{s}_1, \ldots, \mathrm{s}_\mathrm{nw}\}$, gdzie $s_i = \langle u_{t_1}, u_{t_2}, \ldots, u_{t_w} \rangle$ jest sekwencją zdarzeń w jednym dniu terapeutycznym a n_w jest liczbą dni.

W celu wydobycia reguł decyzyjnych ze ścieżek leczenia zastosowaliśmy teorię zbiorów przybliżonych [12]. Rozważane przez nas dane medyczne tworzą system informacyjny $\mathrm{I} = (\mathrm{D}, \mathrm{U})$, gdzie D to zbiór dni terapeutycznych każdego pacjenta, a U to zbiór atrybutów, które są tutaj zdarzeniami medycznymi, gdzie $\mathrm{u} \in \mathrm{U}$ jest odwzorowaniem $\mathrm{u}: \mathrm{U} \rightarrow \mathrm{V}_\mathrm{u}$ i zbiór V_u jest zbiorem wartości u. Zakładamy, że zdarzenia medyczne są uporządkowane zgodnie ze skalą czasu L.

Na podstawie ścieżek leczenia tworzymy zbiór tablic decyzyjnych T_k takich, że $T_k = (D, U_k \cup \{d\})$, gdzie $d = <G = v, t_{k+1}>$ nazywamy atrybutem decyzyjnym, a $t_k \in L$ jest etykietą czasu zdarzenia i $k \in \{2, 4, 6, 8, 10\}$, U_k jest ograniczone do wszystkich zdarzeń poprzedzających zdarzenie d, tj. $t(u \in U_k) < t(d)$. Sekwencjonowanie czasu zostało przeprowadzone zgodnie z mapowaniem przedstawionym w Tabeli 2. Jak już wspomniano, interesują nas tylko reguły prowadzące do hiper lub hipoglikemii, więc v = {1, 4}.

Na podstawie tablic decyzyjnych minimalne reguły decyzyjne ze względu na liczbę atrybutów warunkowych zostały wygenerowane. Zauważmy, że każda reguła jest podzbiorem ścieżki leczenia przedstawionej w formie sekwencji zdarzeń. Niech reguła będzie w postaci $w_1 \wedge \ldots \wedge w_l \rightarrow d_k$ wtedy $w_1, .., w_l, d_k \subseteq s \in S$ może być postrzegana jako podciąg jednej z sekwencji pacjenta. Oczywiście w_l i d_k opisują zdarzenia, które zachodzą w określonym przedziale czasu zdefiniowanym przez zbiór L. Przez W oznaczamy zbiór wszystkich podciągów zdefiniowanych przez zbiór reguł, a przez $U_w \subseteq U$ zbiór zdarzeń istniejących w zbiorze reguł. Na podstawie zbioru zdarzeń U_w oraz zbioru sekwencji W budujemy graf *RB-MTG (Rule Based Medical Treatment Graph)*. Szczegółowy sposób budowy *RB-MTG* został opisany przez algorytm dostępny w pracy [A2]. Przykład takiego grafu przedstawiono na Rysunku 4.



Rysunek 4. Przykładowy RB-MTG dla hypoglikemii (fragment)

Interpretacja medyczna przykładowego grafu *RB-MTG* jest następująca: wszystkie ścieżki prowadzące do hipoglikemii rozpoczynają się łagodną hiperglikemią przed pierwszym śniadaniem $<G = 3, t_1>$. 16% takich pacjentów otrzymywało 3 jednostki insuliny na 100 kcal na 100 kg masy ciała do pierwszego śniadania (first breakfast). Następnie 2% z nich miało hipoglikemię po śniadaniu. Hipoglikemia po drugim śniadaniu może wystąpić, gdy zaobserwujemy łagodną hiperglikemię lub normoglikemię w poprzednim pomiarze poziomu cukru. Wystąpienie hipoglikemii jest jednak bardziej prawdopodobne w pierwszym przypadku i gdy dawka insuliny jest wyższa (równa 5 w przypadku pierwszego śniadania i 3 dla drugiego śniadania). Prawdopodobne jest również (cer = 0,33) wystąpienie hipoglikemii po obiedzie, jeśli hipoglikemię zaobserwowano po drugim śniadaniu. Zauważmy również, że niektóre ścieżki leczenia zostały uogólnione np. podanie insuliny doposiłkowej $<I_p = 3, t_2>$ determinuje normoglykemię w okresie t_5.

Zastosowanie grafu *RB-MTG* niesie ze sobą wiele korzyści. Przede wszystkim zapewnia przejrzystą wizualizację ścieżek leczenia prowadzących do incydentów glikemicznych. Zastosowanie reguł decyzyjnych i teorii zbiorów przybliżonych spowodowało, ograniczenie zdarzeń medycznych, które trzeba analizować w ścieżkach leczenia. Dzięki współczynnikom cer i ω przypisanych do każdej krawędzi *RB-MTG*, możliwa jest szybka analiza różnych potencjalnie ryzykownych zmian terapii. Przedstawienie graficzne alternatywnych ścieżek jest intuicyjne dla lekarza diabetologa i ułatwia podejmowanie decyzji terapeutycznych.

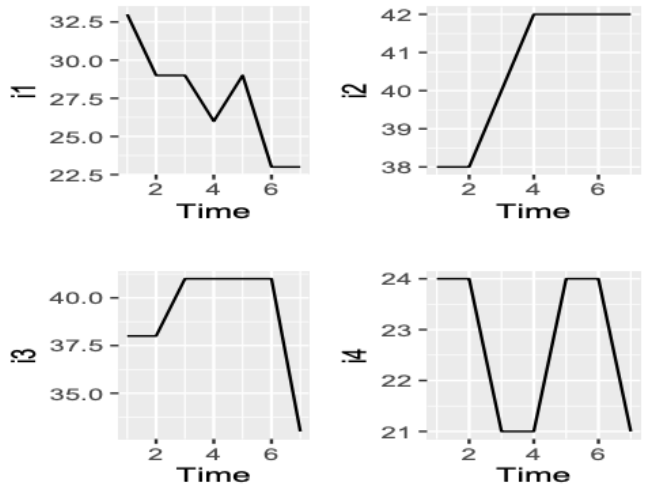
Problem 2

Niezależnie od problemu wydobywania ścieżek leczenia i ich prezentacji autor również podjął się próby wspomagania decyzji diabetologa w określeniu dawki insuliny. Jest to szczególnie istotne przy określeniu dobowej dawki insuliny w pierwszych dniach leczenia. Lekarz dysponuje wtedy jedynie podstawowymi danymi pacjenta oraz jego stanem klinicznym – Tabela 8. Poniższe rozwiązanie może stanowić uzupełnienie do wcześniejszych rozważań; wskazówki związane z przebiegiem leczenia można powiązać z zaproponowaniem konkretnej dawki insuliny.

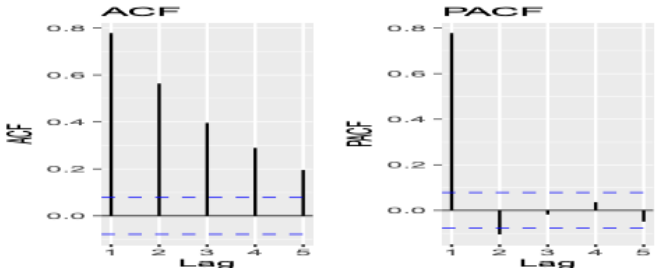
Rozwiązanie

W pracy [A5] przedstawiono konstrukcję narzędzia wspomagania decyzji dla pacjentów ze świeżym zachorowaniem na cukrzycę typu 1. Narzędzie opiera się na rzeczywistych danych medycznych zebranych podczas leczenia pacjentów. Jako wartość decyzyjną przyjęto taką dawkę insuliny, dla której w większości pomiarów poziom cukru był w normie. Podobnie jak inne dane medyczne, zebrane dane są trudne do analizy ze względu na obecność wielu atrybutów o wartościach rzeczywistych, które mogą być niedokładne. Dlatego zaproponowałem zastosowanie teorii zbiorów przybliżonych [12] oraz „roughikacji" [35]. Podejście takie jest skuteczne w sytuacji, gdy mamy do czynienia z atrybutami o wartościach numerycznych, a liczba obiektów nie jest duża. Metoda „roughikacji" została rozszerzona poprzez zastosowanie jej również do atrybutu decyzyjnego i wprowadzenie mechanizmu wyboru reguł decyzyjnych (uzyskane reguły są niedeterministycznych). W wyniku zastosowania w/w klasyfikatora lekarz otrzymuje przedział możliwych wartości dawki dobowej insuliny wraz z informacją o wartości współczynnika wsparcia i pewności podanej decyzji. Warto zauważyć, że proponowany w ten sposób przedział wartości jest wydobywany na podstawie rzeczywistych przykładów, w przeciwieństwie do dyskretyzacji, która opiera się na predefiniowanych, często sztucznych podziałach. Jakość klasyfikatora rozumiana jako odsetek poprawnie sklasyfikowanych nowych/testowych obiektów jest bardzo wysoka.

W kolejnej pracy [A1] zaproponowaliśmy metodę prognozowania (forecasting) bazowej dawki insuliny. Krótki okres analizowanych szeregów czasowych (zebranych podczas leczenia pacjenta w szpitalu) rodzi problem ze stosowaniem metody prognozowania statystycznego, przy rozpatrywaniu indywidualnych szeregów czasowych każdego pacjenta, dlatego zdecydowaliśmy się połączyć dane wszystkich pacjentów i trenować model prognozowania przy użyciu wspólnych danych. Tak przygotowane dane sugerują istnienie trendu i korelacji (Rysunek 6) w szeregach czasowych z oknem 1 (lag 1). Konfrontując autokorelację ze złożonością profili insuliny przedstawionych na Rysunku 5, postanowiliśmy wybrać tak regresory, aby poprawić skuteczności predykcji.



Rysunek 5. Tygodniowy profil insuliny dla przykładowych pacjentów



Rysunek 6. Correlogram dla połączonych szeregów czasowych.

Zaproponowaliśmy dwa modele prognozowania w oparciu o sztuczną sieć neuronową. Pierwszy model prognozuje dawkę insuliny na podstawie stanu zdrowia pacjenta w chwili przyjęcia do szpitala, natomiast drugi umożliwia dostosowanie dawki insuliny bazowej w zależności od przebiegu leczenia. W tym celu drugi model bierze pod uwagę również nocne pomiary glikemii. Przewidywanie dawek insuliny na następny dzień używając zaproponowanych modeli było dość skuteczne, ze średnim błędem bezwzględnym wynoszącym około 3 jednostki insuliny na 100 kg masy ciała pacjenta (Tabela 12).

	Model 1	Model 2
R^2	0.90 (0.87, 0.93)	0.89 (0.85, 0.93)
Mean Absolute Error (MAE)	2.86 (2.45, 3.26)	3.06 (2.65, 3.46)
Mean Squared Error (MSE)	13.49 (10.00, 17.00)	14.61 (10.70, 18.56)

Tabela 12. Dokładność predykcji insuliny

Problem 3

Niejako uzupełnieniem cyklu prac przedstawionego powyżej są moje prace związane ze wspomaganiem pracy lekarza diabetologa na wczesnym etapie potencjalnej choroby (pre-diabetes). Dotyczą one stworzenia systemu prognozowania zachorowania na cukrzycę typu pierwszego.

Powszechnie wiadomo, że cukrzyca typu 1 jest chorobą uwarunkowaną genetycznie. W wielu pracach [33] wykazano, że ryzyko zachorowania na cukrzycę wśród krewnych jest znacznie wyższe niż w normalnej populacji. Niemniej, prócz predyspozycji genetycznej rozwój choroby uzależniony jest również od niesprecyzowanych jednoznacznie innych czynników np. środowiskowych.

Dotychczas nie wykazano dokładnie, które geny i w jakim stopniu odpowiadają za wywołanie choroby – niemniej udowodniono, że wpływ mają geny z układu HLA. Dodatkowo w pracy pod uwagę wzięto genetyczną zdolność do wydzielania cytokin Th1 i Th2, których ilość może mieć znaczący wpływ na rozwój choroby.

Problemem jest zbudowanie systemu wspomagania lekarza w predykcji zachorowania na cukrzycę typu I wśród dzieci obciążonych genetycznie. System taki pozwala na zastosowanie dla takich potencjalnych pacjentów terapii pre-diabetes, która może opóźnić wystąpienie choroby.

Rozwiązanie

W pracy [A10] zaproponowano regułowy system wspomagania decyzji w oparciu o teorię zbiorów przybliżonych [12]. Zebrano dane genetyczne dzieci chorych na cukrzycę i ich zdrowego rodzeństwa ale również nieobciążonej grupy kontrolnej. Najważniejsza idea rozwiązania polegała na najpierw wyznaczeniu klasyfikatora zachorowania na podstawie dzieci chorych na cukrzycę oraz grupy kontrolnej. Następnie klasyfikator ten został poddany ocenie przez grupę zdrowego rodzeństwa i wyznaczono nowy klasyfikator uwzględniający fakt, że nie wszystkie osoby posiadające dany układ genów zachorowały. Wyniki zaproponowanej metody są obiecujące i zostały zweryfikowane przez porównanie różnych metod wyznaczania reguł decyzyjnych w pracy [A9] - dokładność klasyfikatora (accuracy) został wyznaczony metodą walidacji krzyżowej na poziomie 86% używając algorytmu LEM2 [34] do wyznaczenia reguł decyzyjnych.

Dodatkowo wyznaczono również układ genów, który można powiedzieć stanowi ochronę przed zachorowaniem na cukrzycę typu 1.

Literatura

[1] Shortliffe EH, and Buchanan BG (1975). "A model of inexact reasoning in medicine". *Mathematical Biosciences*. 23 (3–4): 351–379

[2] Waterman D.: A Guide to Expert Systems. Addison-Wesley. Reading MA, 1985.

[3] 10 years ago, IBM's Watson threatened to disrupt health care. What happened? (2021) Advisory Board. https://www.advisory.com/daily-briefing/2021/07/21/ibm-watson

[4] Jackson, Peter (1998). *Introduction To Expert Systems* (3 ed.). Addison Wesley. p. 2.

[5] *Kiryanov, Denis Aleksandrovich (2021). "Hybrid categorical expert system for use in content aggregation". Software Systems and Computational Methods (4): 1–22.*

[6] *Kendal, S.L.; Creen, M. (2007). An introduction to knowledge engineering. London: Springer.*

[7] Mitchell, M. (2021). Abstraction and analogy-making in artificial intelligence. *Annals of the New York Academy of Sciences*, *1505*(1), 79-101.

[8] Dourish, P. (2004). *Where the action is: the foundations of embodied interaction*. MIT press.

[9] Hillig, S., & Müller, R. (2021). How do conversational case-based reasoning systems interact with their users: a literature review. *Behaviour & Information Technology*, *40*(14), 1544-1563.

[10] Moskovitch, R. (2022). Multivariate temporal data analysis-a review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *12*(1), e1430.

[11] Masini, R. P., Medeiros, M. C., & Mendes, E. F. (2021). Machine learning advances for time series forecasting. *Journal of Economic Surveys*.

[12] Pawlak, Z., (1981). Information systems – theoretical foundations. Information Systems 6, 205–218.

[13] WHO: Fact sheet no. 312 (2011), http://www.who.int/diabetes/en

[14] ADA (2020). 13. children and adolescents: Standards of medical care in diabetes-2020, Diabetes Care 43: 163–182
[15] Andreassen, S., Benn, J.J., Hovorka, R., Olesen, K.G., Carson, E.R., (1994_. A probabilistic approach to glucose prediction and insulin dose adjustment: description of metabolic model and pilot evaluation study. Computer methods and programs in biomedicine 41, 153–165.
[16] Cappon, G., Vettoretti, M., Marturano, F., Facchinetti, A., Sparacino, G., 2018. A neural-network-based approach to personalize insulin bolus calculation using continuous glucose monitoring. Journal of diabetes science and technology 12, 265–272.
[17] Guzman G´omez, G.E., Burbano Agredo, L.E., Mart´ınez, V., Bedoya Leiva, O.F., 2020. Application of artificial intelligence techniques for the estimation of basal insulin in patients with type I diabetes. International Journal of Endocrinology 2020.
[18] Liu, X., Jankovic, I., Chen, J.H., 2020. Predicting inpatient glucose levels and insulin dosing by machine learning on electronic health records. medRxiv .
[19] Plis, K., Bunescu, R., Marling, C., Shubrook, J., Schwartz, F., 2014. A machine learning approach to predicting blood glucose levels for diabetes management, in: Workshops at the Twenty-Eighth AAAI conference on artificial intelligence.
[20] Torrent-Fontbona, F., 2018. Adaptive basal insulin recommender system based on kalman filter for type 1 diabetes. Expert Systems with Applications 101, 1–7.
[21] Collyns, O. J., Meier, R. A., Betts, Z. L., Chan, D. S., Frampton, C., Frewen, C. M., Hewapathirana, N. M., Jones, S. D., Roy, A., Grosman, B. et al. (2021). Improved glycemic outcomes with medtronic minimed advanced hybrid closed-loop delivery: results from a randomized crossover trial comparing automated insulin delivery with predictive low glucose suspend in people with type 1 diabetes. Diabetes Care, 44 , 969–975.
[22] Demir, G., Er, E., Atik Altınok, Y., Ozen, S., Darcan, S., & Goksen, D. (2021). Local complications of insulin administration sites and effect on diabetes management. Journal of Clinical Nursing.
[23] Agrawal, R., Srikant, R.: Mining Sequential Patterns, *Proceedings of the Eleventh International Conference on Data Engineering, IEEE Computer Society*, 1995.
[24] Mannila, H., Toivonen, H., Verkamo, A. I.: Discovering Frequent Episodes in Sequences, *KDD*, 1995.
[25] Mitsa, T.: *Temporal Data Mining*, CRC Press, Taylor and Francis Group, 2010.
[26] Han, J., Cheng, H., Xin, D., Yan, X.: Frequent pattern mining: current status and future directions, *Data Min. Knowl. Discov.*, **15**(1), 2007, 55–86.
[27] Han, J., Pei, J., Mortazavi-Asl, B., Chen, Q., Dayal, U., Hsu, M.: FreeSpan: frequent pattern-projected sequential pattern mining, *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, 2000.
[28] Huang, K.-Y., Chang, C.-H.: Efficient mining of frequent episodes from complex sequences, *Inf. Syst.*, **33**(1), 2008, 96–114.
[29] Calinski, T. and Harabasz, J. (1974). A dendrite method for cluster analysis, Communications in Statistics - Theory and Methods 3: 1–27
[30] Xie, X. L. and Beni, G. (1991). A validity measure for fuzzy clustering, IEEE Transactions on pattern analysis and machine intelligence 13(8): 841–847.
[31] Dunn, J. C. (1973). A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters, Journal of Cybernetics 3(3): 32–57
[32] Yadav, P., Steinbach, M., Kumar, V. and Simon, G. (2017). Mining electronic health records: a survey, arXiv preprint arXiv:1702.03222

[33] Deja, G., Jarosz-Chobot, P., Polańska, J., Siekiera, U., Małecka-Tendera, E.: Is the association between TNF-alpha-308 A allele and DMT1 independent of HLA-DRB1, DQB1 alleles? Mediators Inflamm. 2006(4), 19724 (2006)
[34] Grzymala-Busse, Jerzy W., and Paolo Werbrouck. "On the best search method in the LEM1 and LEM2 algorithms." *Incomplete Information: Rough Set Analysis*. Physica, Heidelberg, 1998. 75-91.
[35] Slezak, D., Wroblewski, J.: Roughfication of numeric decision tables: The case study of gene expression data. In: International Conference on Rough Sets and Knowledge Technology. pp. 316–323. Springer (2007)
[36] Hripcsak, G., Albers, D. J., & Perotte, A. (2015). Parameterizing time in electronic health record studies. *Journal of the American Medical Informatics Association*, *22*(4), 794-804.
[37] Z. Pawlak, On conflicts, International Journal of Man-Machine Studies, (1984) 127–134.
[38] Z. Pawlak, An inquiry into anatomy of conflicts, Information Sciences, 109 (1-4) (1998) 65–78
[39] Y. Yao, Three-way conflict analysis: Reformulations and extensions of the Pawlak model, Knowledge-Based Systems, 180 (2019) 26-37.
[40] Perer, A., Wang, F., & Hu, J. (2015). Mining and exploring care pathways from electronic medical records with visual analytics. *Journal of biomedical informatics*, *56*, 369-378.
[41] Dagliati, A., Sacchi, L., Zambelli, A., Tibollo, V., Pavesi, L., Holmes, J. H., & Bellazzi, R. (2017). Temporal electronic phenotyping by mining careflows of breast cancer patients. *Journal of biomedical informatics*, *66*, 136-147.
[42] *Feigenbaum, Edward A.; McCorduck, Pamela (1983). The fifth generation(1st ed.). Reading, MA: Addison-Wesley.*

4.4 Pozostałe osiągnięcia naukowo-badawcze

Poniżej przedstawiam omówienie pozostałych osiągnięć naukowo – badawczych z całej kariery zawodowej.

Chciałbym zwrócić uwagę na cykl prac związanych z zagadnieniem analizy konfliktów:

B1. Deja, Rafał. "Rough set approach to conflict analysis." In *Rough set theory and granular computing*, pp. 211-221. Springer, Berlin, Heidelberg, 2003.
B2. Skowron, Andrzej, and R. Deja. "On some conflict models and conflict resolutions." *Romanian Journal of Information Science and Technology* 3, no. 1-2 (2002): 69-82.
B3. Deja, R., Conflict analysis. Int. J. Intell. Syst., 17: 235-253. 2002, https://doi.org/10.1002/int.10019
B4. Deja, Rafał, and Dominik Ślęzak. "Rough set theory in conflict analysis." In *Annual Conference of the Japanese Society for Artificial Intelligence*, pp. 349-353. Springer, Berlin, Heidelberg, 2001.
B5. Deja, R. "Using rough set theory in conflicts analysis." *Institute of Computer Science, Polish Academy of Science, Warsaw* (2000).
B6. Deja, R., L. Polkowski, S. Tsumoto, and T. Y. Lin. "Conflict analysis, rough set methods and applications." *Studies in Fuzzyness and Soft Computing* (2000): 491-520.
B7. Deja, R. "Conflict analysis, rough sets; new developments." *Studies in Fussiness and Soft Computer Science, Physica-Verlag* (2000).
B8. Deja, Rafal. "Conflict analysis." In *Rough Set Methods and Applications*, pp. 491-519. Physica, Heidelberg, 2000.

B9. Deja R., "Conflict Model with Negotiation", Bulletin of the Polish Academy of Sciences, Technical Sciences, vol. 44, no. 4, 1996, pp. 475-498.

B10. Deja R., "Conflict Analysis", Proceedings of the Fourth International Workshop on Rough Sets, Fuzzy Sets and Machine Discovery, The University of Tokyo, November 6-8 1996, pp. 118-124.

Analiza konfliktów przedstawiona w pracach koncentruje się na zrozumieniu i rozwiązaniu konfliktu, który wynika z różnych postaw agentów lub organizacji wobec zestawu problemów/zagadnień. Aby formalnie sformułować problem, Pawlak [26, 27] zaproponował trójwartościową tabelę sytuacji, która przedstawia postawy, opinie lub oceny zbioru agentów w zbiorze zagadnień. Trójwartościowa tablica sytuacji to trójka $\mathrm{S} = (\mathrm{A}, \mathrm{I}, \mathrm{r})$, gdzie A jest skończonym niepustym zbiorem agentów, I jest skończonym niepustym zbiorem zagadnień, oraz $\mathrm{r} : \mathrm{A} \times \mathrm{I} \rightarrow \{+1, -1, 0\}$ to funkcja oceny. Wartość $\mathrm{r}(\mathrm{x}, \mathrm{i})$ nazywana jest oceną agenta $\mathrm{x} \in \mathrm{A}$ w sprawie $\mathrm{i} \in \mathrm{I}$. Jeśli $\mathrm{r}(\mathrm{x}, \mathrm{i}) = +1$, to x jest popiera i; jeśli $\mathrm{r}(\mathrm{x}, \mathrm{i}) = -1$, to x jest przeciwne względem i; jeśli $\mathrm{r}(\mathrm{x}, \mathrm{i}) = 0$, to x jest neutralne względem i.

Te trzy postawy natychmiast dzielą zbiór agentów na trzy grupy w następujący sposób [28]. Dla podzbioru agentów $\mathrm{X} \subseteq \mathrm{A}$ oraz zagadnienia $\mathrm{i} \in \mathrm{I}$, trysekcja X względem i jest zdefiniowana jako:
$X_i^+ = \{x \in X | r(x,i) = +1\},\ X_i^- = \{x \in X | r(x,i) = -1\}, X_i^0 = \{x \in X | r(x,i) = 0\}$

Innym podstawowym i ważnym tematem analizy konfliktów jest badanie relacji między agentami. Zasadniczo istnieją trzy relacje między agentami, a mianowicie sojusz, konflikt i relacje neutralności. Relacje te są definiowane na podstawie ocen agentów dotyczących pojedynczego zagadnienia a dalej podzbioru zagadnień. Formalnie definiujemy funkcję $\Phi_i: A \times A \rightarrow \{+1, -1, 0\}$, jeśli $\Phi_i(x, y) = +1$, to x i y są sojusznikami w kwestii i, jeśli $\Phi_i(x, y) = -1$, to są w konflikcie w sprawie i, a jeżeli $\Phi_i(x, y) = 0$ to agenci są neutralni wobec zagadnienia i.

W początkowych pracach rozszerzyłem model Pawlaka wprowadzając funkcję pozwalającą ocenić odległość agentów na podstawie ich oceny poszczególnych zagadnień. W najprostszej formie możemy zastosować średnią (w dalszej części rozważam również średnią ważoną):
$\Phi_J(x, y) = \frac{\sum_{i \in J} \Phi_i(x,y)}{|J|}$. Dzięki tej funkcji zaproponowałem poszerzenie pojęcia koalicji. Dokładniej [27] można zdefiniować następujące relacje dla wartości progowych l i h, gdzie $-1 \leq l < 0 < h \leq +1$; relacja koalicji $R_{\Phi_J}^{=} = \{(x, y) \in A \times A | \Phi_J(x, y) \geq h\}$, relacja konfliktu $R_{\Phi_J}^{\asymp} = \{(x, y) \in A \times A | \Phi_J(x, y) \leq l\}$ oraz relacja neutralności $R_{\Phi_J}^{\approx} = \{(x, y) \in A \times A | l < \Phi_J(x, y) < h\}$ w zbiorze $\mathrm{J} \subseteq \mathrm{I}$. Następnie zaproponowałem grupowanie agentów będących w koalicji, jak i również redukcję „mało konfliktowych" atrybutów. Prace uzupełnia algorytm znajdujący takie atrybuty, które po usunięciu zmniejszą konflikt omawianej sytuacji.

Analiza konfliktów opisywanych przez model Pawlaka jest ograniczona do podstawowych konkluzji, takich jak znalezienie najbardziej konfliktowych atrybutów lub koalicji agentów. W modelu Pawlaka przyczyny konfliktu nie są widoczne dlatego znajdywania rozwiązania jest utrudnione. Dlatego w dalszych pracach zaproponowałem nowy model sytuacji konfliktowej. Nowy model składa się z następujących elementów:

- stanów lokalnych agentów wraz z subiektywną ich oceną – są one zdefiniowane jako tablice informacyjne, które na podstawie oceny przekształcana są w tablice decyzyjne [21]
- sytuacji, które są niejako złożeniem stanów lokalnych agentów, które również podlegają ocenie np. przez eksperta
- ograniczeń opisujących powiązania lokalnych stanów agentów.

Dalej definiowane są konflikty lokalne, które niejaką definiują stan do którego agent dąży oraz konflikt globalny. Konflikt globalny może powstać dla wybranych obiektów tablicy sytuacji. Najważniejszym problemem w analizie konfliktów jest znalezienie konsensusu. W tak zdefiniowanym modelu rozwiązanie polega na weryfikacji sytuacji globalnych z lokalnym zbiorem celów agentów i ograniczeń. Problem możemy zapisać z pomocą formuły boolowskiej a pierwsze implikanty formuły definiują rozwiązanie problemu (proponowany konsensus) $f = \bigwedge_{x \in A} t_x \wedge f_C \wedge f_\varphi$, gdzie t_x opisuje cele agenta x, f_C opisuje sytuacje globalne a f_φ ograniczenia. Zaproponowany model stał się też podstawą do rozważań współdziałania niezależnych agentów w sytuacji, w której mają zrealizować postawione im zadanie.

Prace w zakresie analizy konfliktów cieszą się dużym zainteresowanie patrząc na liczbę cytowań.

5. Informacja o wykazywaniu się istotną aktywnością naukową, w szczególności zagraniczną

Aktywność naukowa zagraniczna była realizowana przez aktywny udział w międzynarodowych konferencjach/zjazdach, na których przedstawiałem najnowsze wyniki moich badań. Spis konferencji z ostatnich 10 lat):

26th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems, zorganizowana przez KES International and University of Verona, sesja: “Classification, forecasting and decision support”, 2022
(referat: “Selection a group of features based on machine learning algorithms to simplify psycho-technical examination”)

25th International Conference on Knowledge Based and Intelligent information and Engineering Systems, zorganizowana przez KES International and University of Szczecin, sesja: “Decision support systems, forecasting, classification” 2021
(referat: "Rule-based Medical Treatment Graph for the Modeling of Hypo-and Hyperglycemia at Onset.")

21st International Conference on Knowledge-Based and Intelligent Information & Engineering Systems, zorganizowana przez KES International and the Laboratoire des Sciences de l'Information et des Systemes, Aix-Marseille University, 2017
(referat: “Applying Roughication to Support Establishing Intensive Insulin Therapy at Onset of T1D”)

8th International Conference on Computational Collective Intelligence Technologies and Applications, zorganizowana przez Aristotle University of Thessaloniki and Democritus University of Thrace in Greece and Wroclaw University of Science and Technology in Poland, in cooperation with IEEE SMC Technical Committee on Computational Collective Intelligence, Halkidiki, Greece, 2016
(referat: "Building Medical Guideline for Intensive Insulin Therapy of Children with T1D at Onset.")

11th Conference Internet in the information society, zorganizowana przez University of Dąbrowa Górnicza, 2016
(referat: „The Method for Supporting Insulin Therapy at the Onset”)

The 9th International Conference on Advanced Technologies & Treatments for Diabetes, Milan, Italy, February 3-6, 2016
(prezentacja: “Decision Support Tool For Intensive Insulin Therapy Of Children With T1dm At Onset”)

10th Conference Internet in the information society, zorganizowana przez University of Dąbrowa Górnicza, 2013
(referat: “Decision support system for treatment of children with diabetes type 1”)

9th Conference Internet in the information society, zorganizowana przez University of Dąbrowa Górnicza, 2012
(referat: "Comparison of Rules Synthesis Methods Accuracy in the System of Type 1 Diabetes Prediction

The 6th International Conference on Rough Set and Knowledge Technology, zorganizowana przez University of Regina, Banff, Canada, 2011
(referat: Accuracy evaluation of the system of Type 1 diabetes prediction)

Projekty badawcze:

„OPTIMIS - innowacyjny system diagnostyki psychologicznej dostosowanej do potrzeb polskiego orzecznictwa" współfinansowanego przez Unię Europejską ze środków Europejskiego Funduszu Rozwoju Regionalnego w ramach Programu Operacyjnego Inteligentny Rozwój. Projekt realizowany w ramach konkursu Narodowego Centrum Badań i Rozwoju: Szybka Ścieżka 1_2020. Udział w projekcie w roli Data Scientist.

Realizacja grantu ESPIT-CRIT2 No. 20288 współfinansowanego przez Polish National Committee for Scientific Research No. 8T11C00512

6. Informacja o osiągnięciach dydaktycznych, organizacyjnych oraz popularyzujących naukę lub sztukę

Prowadzę zajęcia dydaktyczne od 2008 roku w Akademii WSB. Przygotowałem (również w wersji online) cykle wykładów, ćwiczeń i laboratoriów z następujących przedmiotów

- Inżynieria oprogramowania (wykład i ćwiczenia) również w języku angielskim
- Metody sztucznej inteligencji (ćwiczenia)
- Programowanie obiektowe (wykład i laboratorium) również w języku angielskim
- Technologie internetowe (wykład i laboratorium)
- Podstawy programowania komputerów (laboratorium) również w języku angielskim
- Zarządzanie projektami informatycznymi (studia podyplomowe)

„Rozproszone systemy komputerowe – rozwiązania mobilne”; wykład na V Festiwal Nauki Wyższa Szkoła Biznesu w Dąbrowie Górniczej, 2009

Zorganizowałem sesję naukową p.t.: „Classification, forecasting and decision support”, podczas konferencji 26th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems

Byłem członkiem komitetu naukowego na kilku konferencjach m.in. International Conference on Knowledge-Based and Intelligent Information & Engineering Systems, International Conference on Rough Set and Knowledge Technology, International Conference on Computational Collective Intelligence Technologies and Applications,

Byłem autorem wielu recenzji w czasopismach takich jak Information science, International Journal of Applied Mathematics and Computer Science, Fundamenta Informaticae

Rozdziały w książkach:

- Deja R. „System doboru pracowników", Innowacyjne metody i narzędzia wspomagające podejmowanie decyzji w zarządzaniu, praca zbiorowa pod redakcją Adriana Kapczyńskiego, Wyższa Szkoła Biznesu w Dąbrowie Górniczej, 2010 pp. 145-155
- R. Deja, "Decision support system for treatment of children with diabetes type 1" in Computer Systems Architecture and Security ed. P. Pikiewicz, M. Rostański, Academy of Business in Dabrowa Gornicza, 2013; 69-78
- Deja Rafał, Dyskretyzacja atrybutów o wartościach rzeczywistych na przykładzie systemu predykcji terapii w cukrzycy typu 1, "Systemy wspomagania decyzji" Instytut Informatyki UŚ, 2013

7. Dodatkowe informacje dotyczące kariery zawodowej

Poza pracą naukową i dydaktyczną od wielu lat związany jestem z praktycznymi zastosowaniami informatyki w różnych działach przemysłu i usług. Od 2014 roku pracuje w firmie IBM (obecnie Kyndryl) prowadząc zespół implementujący rozwiązania dla jednego z największych europejskich banków.

Wśród prowadzonych projektów w ramach różnych przedsiębiorstw z którymi współpracowałem należy wspomnieć następujące, w których wdrożeniu miałem znaczący udział:

- Opracowanie komputerowego systemu testów psychologicznych wykorzystywanych m.in. przy rekrutacji w Policji czy w Poczcie Polskiej
- Implementacja komputerowego systemu wspomagania decyzji i procesów zarządzania zasobami ludzkimi w organizacjach.
- Opracowanie modułów systemu magazynowego na urządzenia mobilne dla hurtowni farmaceutycznych.
- Realizacja projektu analizy jakości danych w systemach pomocy społecznej w celu poprawy dystrybucji środków finansowych.