Postęp nauk biologicznych w dobie big data będzie zależał od tego, w jaki sposób zajmiemy się następującym pytaniem: "Jak połączyć wiele różnych typów danych w celu uzyskania pełnego zrozumienia funkcji biologicznych organizmu?”

Technologie „omiczne” są stosunkowo nową dziedziną badań. Należy do nich m.in. genomika (ocena genomu), transkryptomika (ocena mRNA), proteomika (ocena białek) i metabolomika (ocena metabolitów). Dzięki połączeniu danych z różnych technologii omicznych można uzyskać pełniejszy obraz złożonych zdarzeń molekularnych a przede wszystkim zwiększyć zrozumienie jak funkcjonują zdrowe i zmienione chorobowo komórki. Analiza danych na podstawie zintegrowanych danych omicznych otwiera wiele nowych możliwości w leczeniu i diagnostyce, jednak stwarza również nowe wyzwania dla algorytmów uczenia maszynowego.

Celem pracy jest weryfikacja hipotezy, że analiza zintegrowanych danych omicznych wykorzystująca relacje i zależności zachodzące pomiędzy cechami indywidualnego pacjenta może być z powodzeniem wykorzystana do prognozowania i wspomagania diagnostyki. W tym celu zaproponowany zostanie nowy zintegrowany system oparty na uczeniu maszynowym, który wspierać będzie poszukiwanie prostych w interpretacji reguł decyzyjnych. Ogólna idea systemu nazwanego Relative Dependencies Analysis (RDA) polega na poszukiwaniu relacji pomiędzy cechami w obrębie jednej osoby a następnie skonfrontowane ich z resztą danych pacjentów w celu znalezienia wzorców, które rozróżniają zdefiniowane klasy w danych. Innowacyjne podejście polega na zastąpieniu danych rzeczywistych informacją o relacjach porządkowych pomiędzy cechami. Ten krok oczywiście powoduje pewną utratę potencjalnie ważnych informacji, jednak otwiera on również daleko idące nowe możliwości. Przede wszystkim system RDA będzie odporny na czynniki metodologiczne i techniczne, specyficzne dla badań uprzedzenia, jak również procedury normalizacji i standaryzacji. Dodatkowo, wykorzystanie wyłącznie relacji w obrębie każdej próbki pozwoli na integrację danych z różnych omik, platform i eksperymentów.

Główne powody podjęcia się danej tematyki badawczej wynikają z naszych dotychczasowych doświadczeń związanych z projektowaniem algorytmów, które łatwo zrozumieć i zinterpretować. Dodatkowo, zdecydowana większość obecnie opracowywanych metod klasyfikacji danych biologicznych koncentruje się wyłącznie na jak najwyższej dokładności. Powoduje to wysoką złożoność wygenerowanych przez nie reguł decyzyjnych co znacznie utrudnia interpretację, biologiczne zrozumienie czy wykrywanie potencjalnych biomarkerów.

Istnieją dwa główne problemy naukowe, do których odnosi się ten projekt. Pierwszym z nich jest analiza integracyjna i jednoczesne badanie wielu danych omicznych, które są kluczem do nowych odkryć biomedycznych i postępów w medycynie spersonalizowanej. W tym celu zaprojektowany i zaimplementowany zostanie system RDA, który składać się będzie z kilku połączonych ze sobą algorytmów i koncepcji, w tym drzew decyzyjnych, algorytmów ewolucyjnych, wielo-testowym podziale a także relacyjnej analizie ekspresji (RXA). Dodatkowo, dzięki zrównolegleniu RDA przeprowadzonym na akceleratorach graficznych (GPU) możliwa będzie wydajna analiza dużych zbiorów danych. Drugi z nich koncentruje się na samym odkryciu wiedzy w celu wydobycia znaczących molekularnych sygnatur procesów biologicznych. Pionierski charakter projektu będzie polegać na rozwiązywaniu obu tych problemów za pomocą analizy zależności względnych.

Są trzy konkretne zadania, które mają na celu zweryfikowanie hipotezy badawczej projektu. W pierwszym przeprowadzona zostanie szczegółowa analiza względnych zależności pomiędzy cechami i ich zastosowaniem w klasyfikacji danych ekspresji genów. Zadanie polegać będzie na analizie (i) wpływu różnych reprezentacji relacji; (ii) hierarchicznym i horyzontalnym rankingu opartym o RXA; (iii) złożoności, podobieństwa i siły dyskryminacyjnej zależności. Dodatkowo, w ramach tego zadania zaprojektowany i zaimplementowany zostanie system RDA. W drugim zadaniu system RDA dostosowany zostanie do analizy różnych pojedynczych danych omicznych. Ostatnie zadanie dotyczy eksploracji zintegrowanych wielu danych omicznych i odkrywaniu w nich wiedzy przy wykorzystaniu systemu RDA. Przeprowadzona zostanie również analiza wygenerowanych reguł decyzyjnych w kontekście potencjalnego zastosowania w badaniach podstawowych biologii molekularnej, w tym odkrywaniu i klasyfikacji biomarkerów.