Silesian University of Technology in Gliwice, Poland
Automatic Control, Electronics, and Computer Science Department

Title: Classification of white blood cells based on single-cell sequencing data for biodosimetry purposes

Author: Katarzyna Sieradzka

Supervisor: prof. dr hab. inż. Joanna Polańska
Advisor: dr Christophe Badie

Acknowledgments: This work has been supported by European Union under the European Social Fund grant AIDA – POWR.03.02.00-00-I029.

Poszerzony abstrakt

Prezentowana rozprawa doktorska opiera się na dwóch fundamentalnych hipotezach:

1. **Połączenie metod inżynierii cech i zaawansowanych technik redukcji wymiarowości z nienadzorowanymi algorytmami grupowania pozwala na skuteczną identyfikację podtypów białych krwinek w danych pochodzących z eksperymentów sekwencjonowania RNA pojedynczych komórek.**

2. Zaproponowany inteligentny i warstwowy algorytm konstrukcji zbioru treningowego wspomaga system klasyfikacji, zwłaszcza w przypadku zbiorów heterogenicznych.

Niniejsza rozprawa doktorska zakłada dwa główne cele analizy danych z pochodzących eksperymentów sekwencjonowania pojedynczej komórki (scRNA-seq): stworzenie przepływu pracy analizy niezbędnej do rozpoznania sygnatury genowej białych krwinek napromienionych dawką 1 Gy oraz wykrycie i wydzielenie subpopulacji białych krwinek w danych sekwencjonowania pojedynczej komórki. Dodatkowo zakłada się jeden cel poboczny: porównanie dwóch metod opartych na uczeniu maszynowym pod kątem klasyfikacji komórek kontrolnych i napromienionych oraz uzyskanych profili genetycznych komórek napromienionych. Wszystkie te cele wiążą się z wdrożeniem wielu metod bioinformatycznych, w szczególności schematu przebiegu analizy, który automatyzuje poszczególne etapy pracy. Najistotniejsze pod względem treści i dające największe możliwości późniejszej manipulacji procesami zastosowanych kroków analizy jest zbudowanie odpowiedniego schematu pracy związanego z kluczowym jej etapem, jakim jest selekcja cech.

Rozpoznanie sygnatury genowej komórek napromienionych dawką 1 Gy wymusza wdrożenie metody opartej na uczeniu maszynowym, która będzie przemyślana i dostosowana do złożoności zadania i analizowanych danych. Po wprowadzeniu danych scRNA-seq białych krwinek stworzony algorytm powinien nauczyć się specyficznych struktur genowych, które pozwolą na najskuteczniejsze rozdzielenie komórek kontrolnych i napromienionych. Ze względu na znaczną złożoność problemu badawczego konieczne jest zastosowanie kombinacji i współdziałania wielu narzędzi i metod określających wpływ poszczególnych genów struktur genowych, w tym współdziałanie genów i łączny wpływ na różnicowanie komórek. Jedną z istotnych części proponowanego przepływu pracy jest procedura selekcji cech na szczegółowych i iteracyjnie realizowanych metodach modelowania. W wyniku procedury selekcji cech zakłada się, że zostanie utworzona lista wybranych genów odpowiadająca danemu problemowi a więc, że czynnik promieniowania jonizującego wpływa na zróżnicowanie ekspresji genów. Następnie, wykryta sygnatura genowa może być wykorzystana do biologicznego wnioskowania o strukturze i sile zmian w krwinkach białych. Niezbędne będzie również wykorzystanie modelu genetycznego wykrytych napromienionych komórek w szerszy i bardziej ogólny sposób, tj. rozróżnienie komórek kontrolnych i napromienionych pochodzących z eksperymentów sekwencjonowania pojedynczych komórek do celów klasyfikacji komórek.

Porównanie aplikacji i wyników uzyskanych z dwóch technik uczenia maszynowego, tj. modelowania w oparciu o regresję logistyczną i sieci neuronowe, musi dotyczyć dwóch najważniejszych aspektów pracy. Pierwszy to oczywiście etap selekcji cech, w wyniku którego zostanie określona sygnatura genowa komórek napromienionych w środowisku ex vivo. Zakłada się, że sygnatura dla obu podejść powinna być podobna. Jednak bezwzględna liczba cech zawartych w tej sygnaturze nie jest brana pod uwagę, ale ich specyficzne nazwy i funkcje.

Rozpoznanie subpopulacji komórkowych to kolejny aspekt bezpośrednio związany z różnymi podtypami komórek występującymi w analizowanych danych dotyczących białych krwinek. Taka analiza heterogeniczności komórek może otworzyć dalsze możliwości porównawcze i analityczne dla podejmowanych badań. Różnice w odpowiedziach podtypów krwinek białych mogą być na tyle znaczące, że niekorzystnie wpłyną na samouczące się algorytmy klasyfikacji.

Problem identyfikacji subpopulacji komórkowych nie jest jeszcze szeroko dyskutowany w literaturze, zwłaszcza jeśli chodzi o dane krwinek białych. Do niedawna problem identyfikacji subpopulacji komórkowych w zbiorze danych biologicznych opierał się jedynie na manualnej interpretacji i przypisywaniu komórek ze względu na ich fenotyp lub morfologię [1]. W dzisiejszych czasach, w związku z postępującą automatyzacją procesów pozyskiwania danych biologicznych oraz wprowadzaniem coraz dokładniejszych metod otrzymywania tych danych, takich jak technologia sekwencjonowania pojedynczej komórki, generująca dane szczególnie wysokowymiarowe, bardzo pożądana jest również automatyzacja procesów rozpoznawania określonych struktur komórkowych czy ogólna analiza zbiorów danych biologicznych.

Białe krwinki dzielą się na dwie główne grupy: granulocyty i agranulocyty. Wymienione, dzielą się również na wiele mniejszych podtypów. Granulocyty obejmują neutrofile, eozynofile i bazofile. Agranulocyty zawierają podtypy takie jak limfocyty (w tym komórki T, komórki B i komórki NK) i monocyty. Jednak najbardziej znaczący odsetek stanowią neutrofile, które pokrywają około 50-70% frakcji białych krwinek. W dalszej kolejności, ze względu na procentowy udział we frakcji białych krwinek, wyróżnia się limfocyty (25-35%), monocyty (4-6%), eozynofile (1-3%) i bazofile (do 1%) [2]. Komórki T stanowią najliczniejszą frakcję wśród limfocytów, stanowiąc od 80 do 90% ich frakcji, na drugim miejscu plasują się komórki B, a trzecie pod względem udziału procentowego komórki NK. Fascynujące badanie opublikowane przez R. C. Wilkinsa i in. [3] w 2002 roku wykazało, że krwinki białe mają istotnie różne odpowiedzi komórkowe na czynnik promieniowania. W tej analizie uwzględniono subpopulacje granulocytów, komórek B, komórek NK i komórek T. Wykorzystując zmodyfikowany test kometowy, analizowano frakcję apoptotyczną próbki kontrolnej i próbki poddanej działaniu promieni rentgenowskich. W wyniku badania frakcji apoptotycznej komórek kontrolnych wykazano, że największą frakcją spontanicznej apoptozy charakteryzowały się granulocyty. Subpopulacja komórek B i komórek NK również wykazywała wysoki wskaźnik, podczas gdy najniższy wskaźnik opisywał frakcję komórek T. Ponadto, w wyniku naświetlania komórek, granulocyty ponownie wykazywały najwyższą wartość frakcji apoptotycznej, podczas gdy najbardziej znaczący wzrost wystąpił wśród subpopulacji komórek T. Wskazuje to na bardzo dużą wrażliwość subpopulacji komórek T na działanie promieniowania.

Problem selekcji cech ma fundamentalne znaczenie w budowaniu wydajnego i dokładnego klasyfikatora. W kontekście dostępnych cech, wymiarowość zbioru danych niewątpliwie wpływa na koszty obliczeniowe i czasowe niezbędne do osiągnięcia tego celu. Właśnie dlatego redukcja wymiarowości jest pożądanym krokiem przed podejściem do budowania modelu. Pozwala ona ograniczyć liczbę cech do tych, które są istotne z punktu widzenia rozpatrywanego problemu badawczego. Zjawisko selekcji cech zostało szczegółowo opisane w opublikowanym manuskrypcie zatytułowanym *Feature selection methods for classification purposes* [4]. Uogólniając, metody selekcji cech obejmują podejście nienadzorowane i nadzorowane. Nienadzorowana selekcja cech odnosi się do procesu, który nie potrzebuje etykiety klasy wyjściowej do wyboru cech, a ten typ podejścia może być z powodzeniem stosowany w przypadku danych nieoznaczonych. Z drugiej strony, nadzorowana selekcja cech odnosi się do metody wykorzystującej etykiety klasy wyjściowej. Metody nadzorowane dzielą się na opakowujące, filtrujące i hybrydowe [5]. Filtry nie są bardzo czasochłonne, ponieważ nie wykorzystują skomplikowanych algorytmów uczenia maszynowego [5]. Bez ekstremalnych zasobów obliczeniowych i znacznej ilości czasu można z powodzeniem zmniejszyć wymiarowość danych. Większość metod z dziedziny filtrów jest łatwa do implementacji, ponieważ opierają się one głównie na statystykach, które umożliwiają tworzenie rang cech [6]. Cechy są selekcjonowane na podstawie ich

relacji do danych wyjściowych lub tego, jak są skorelowane z tymi danymi. Filtry są dość ogólnym podejściem, dlatego są szczególnie polecane do danych wysokowymiarowych [7]. Metody filtrujące mają jedną zasadniczą wadę — analizują każdą cechę z osobna, oceniając jej stosunek do celu. To czyni je skłonnymi do odrzucania cennych cech, które są słabymi predyktorami celu, ale w połączeniu z innymi cechami dodają wiele wartości do modelu [6]. Kolejnym typem nadzorowanej metody jest podejście opakowujące. Ta metoda wykorzystuje algorytm uczenia się do wyboru najbardziej krytycznych, różnicujących cech. Jednak zastosowanie złożonej metodyki wymaga znacznych zasobów obliczeniowych, zwłaszcza w przypadku danych wysokowymiarowych [7]. Dlatego nie jest to całkowicie zalecana metoda dla złożonych zbiorów danych [8]. Algorytm opakowujący dzieli dane wejściowe na podzbiory i trenuje model [5], który jest następnie używany do oceniania różnych podzbiorów cech w kontekście wybrania najlepszego z nich. Jak wspomniano wcześniej, podejście hybrydowe jest trzecim rodzajem nadzorowanej metody selekcji cech. Łączy zarówno metody opakowujące, jak i filtry w jeden system. Daje największe pole do manipulacji i selekcji pod kątem zamierzonych celów. Niewłaściwe i nieprzemyślane zastosowanie losowej kolejności występowania technik z kategorii filtrów czy metod opakowujących może skutkować niepoprawnie interpretowanymi wynikami analizy [9].

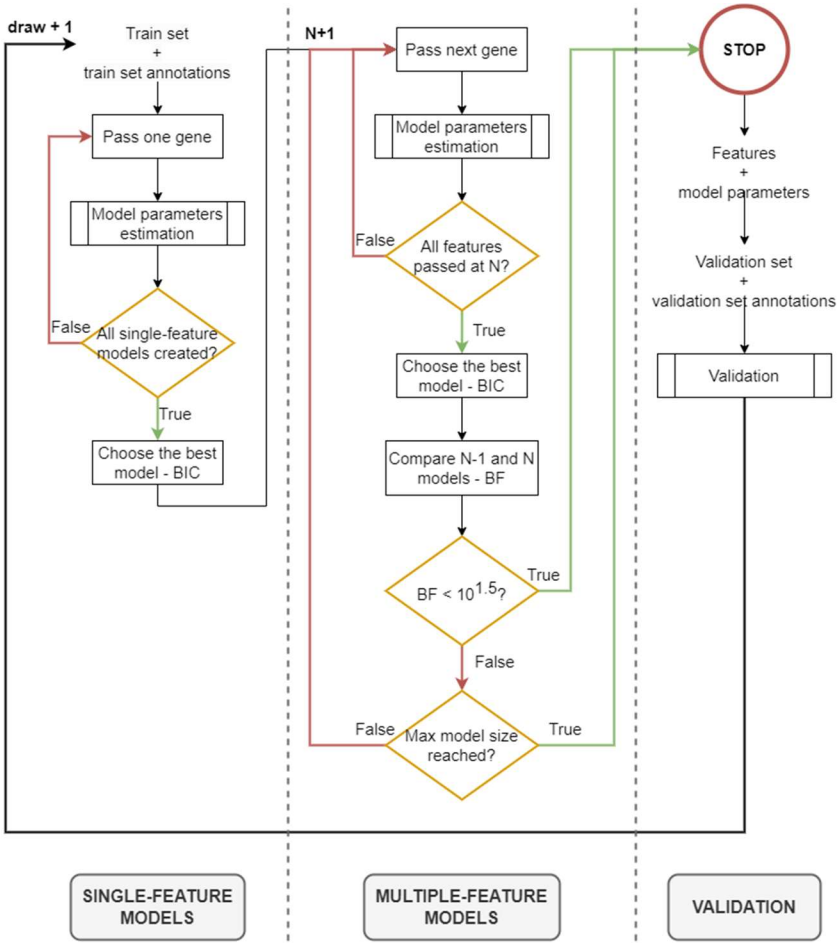
W rozprawie analizowane są dwa wysokowymiarowe zbiory danych. Oba zestawy pochodzą z eksperymentów sekwencjonowania RNA pojedynczych komórek opartych na białych krwinkach i są technicznymi powtórzeniami tego samego eksperymentu. Aby zachować przejrzystość i czytelność odniesień do poszczególnych zestawów danych, zostały one przedstawione i opisane jako zestaw A oraz B. Oba zestawy zawierają dwie próbki komórek: normalną (kontrolną) i napromienioną dawką 1 Gy w środowisku ex vivo. Przebieg analizy danych w tej rozprawie doktorskiej został dostosowany do postaci macierzy zliczeniowych stworzonych przez system BD Rhapsody™. Wiersze w wygenerowanych macierzach zliczeń są reprezentowane przez listę indeksów komórek wprowadzonych do analizy. Kolumny zawierają listę genów zastosowanego panelu odpowiedzi immunologicznej. Ze względu na taki projekt macierzy, każdy element w jej strukturze reprezentuje liczbę cząsteczek wykrytych w komórce na określony gen. Pierwsze dwie kolumny, opisane w Tabeli 1, pokazują liczbę komórek kontrolnych i napromienionych dawką 1 Gy. Trzecia kolumna reprezentuje komórki zidentyfikowane przez filtry, zastosowane przez platformę BD Rhapsody jako te, które nie przeszły pomyślnie wstępnych etapów kontroli jakości.

Tabela 1. Liczba komórek oraz genów w kolejnych zbiorach danych ex vivo..

Zbiór danych	Komórki kontrolne	Komórki napromienione	Inne komórki	Komórki całkowicie	Geny całkowicie
Zbiór A	1584	1139	1516	4239	452
Zbiór B	2301	1988	2633	6922	452

W rozprawie doktorskiej wykorzystano publicznie dostępne bezpłatne narzędzia i opracowane przez siebie przepływy pracy z wykorzystaniem metod statystycznych i metod uczenia maszynowego. Publicznie dostępne narzędzia, takie jak UMAP [10] i HDBSCAN [11], zostały wykorzystane do wizualizacji zbiorów danych metodami uczenia nienadzorowanego i rozpoznawania poszczególnych subpopulacji krwinek białych. Zastosowano opracowane przez siebie przepływy pracy w szerokim zakresie, w tym główne aspekty, takie jak selekcja cech, budowanie modeli w oparciu o metody regresji logistycznej i sieci neuronowych, klasyfikacja komórek kontrolnych i napromienionych oraz budowanie profilu genetycznego komórek napromienionych dawką 1 Gy. Przyczyną opracowania nowego przepływu pracy związanego z analizą danych z eksperymentów sekwencjonowania pojedynczych komórek był przede wszystkim brak opracowanej i publicznie dostępnej metody selekcji cech i budowy modelu opartego na profilu genetycznym komórek napromienionych. Dostępne metody przetwarzania wymagają wielu narzędzi, w których niemożliwe jest zapewnienie pełnej kontroli nad przepływem i analizą wprowadzanych danych. Brak kontroli na pewnym etapie analizy doprowadzi do niedokładności i zniekształconych wyników w postaci profilu genetycznego komórek napromienionych. W konsekwencji proces klasyfikacji komórek również zostanie poważnie zaburzony.

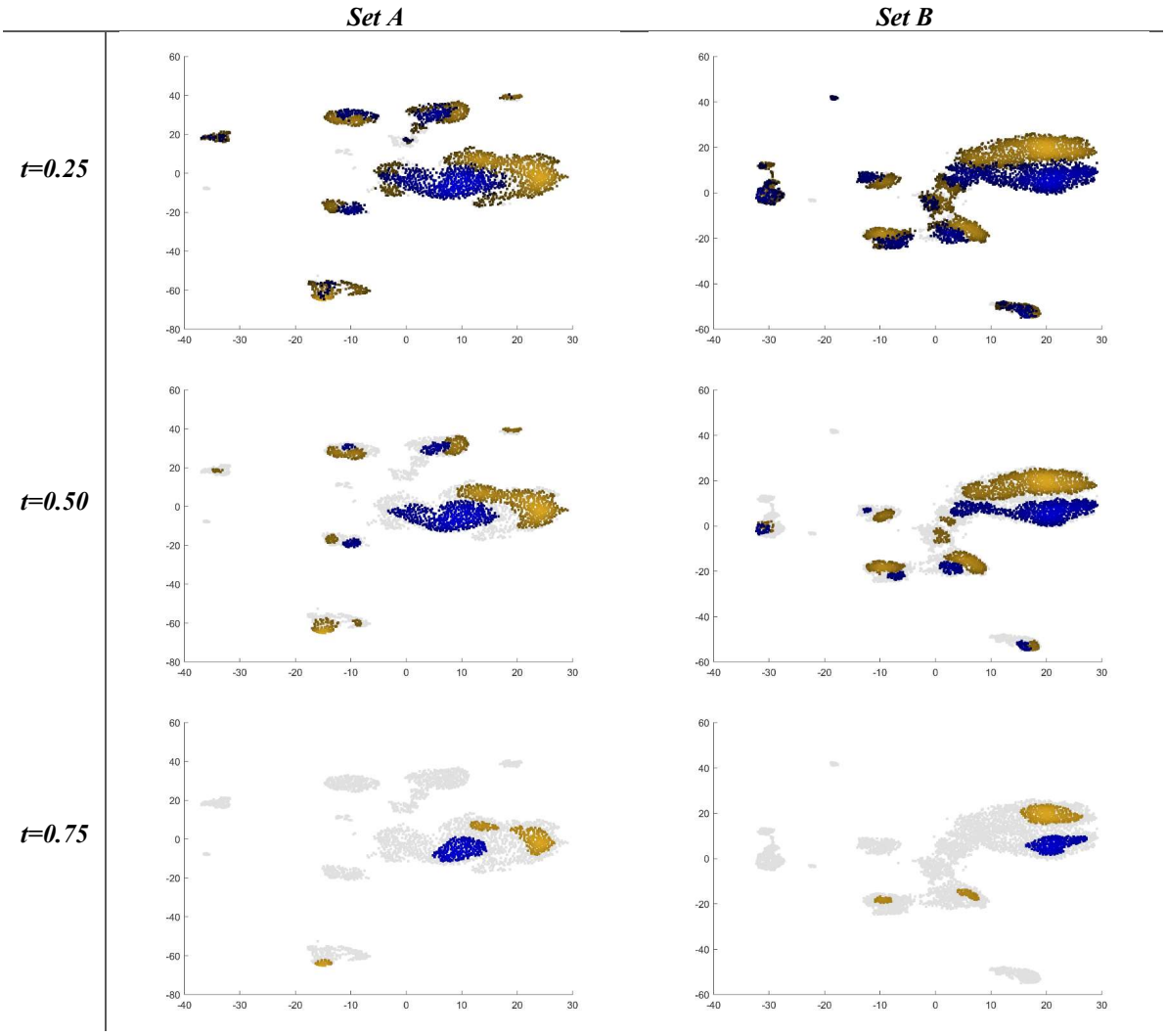
Algorytm oparty na metodologii uczenia maszynowego, realizowany w ramach pracy doktorskiej, ma na celu wygenerowanie modelu opisującego sygnaturę genową komórek poddanych działaniu promieniowania jonizującego. Zastosowano binarną regresję logistyczną, której zadaniem było nauczenie klasyfikatora przypisania obserwacji do jednej z dwóch klas: pozytywnej (napromienionej) lub negatywnej (kontrolnej). Regresja logistyczna szacuje wektor wag przypisanych do składowych modelu. Wagi mogą przyjmować zarówno wartości dodatnie, pokazujące wpływ na klasyfikację frakcji pozytywnej, jak i ujemne, pokazujące wpływ na klasyfikację frakcji negatywnej [12]. Aby obliczyć wartość prawdopodobieństwa danej obserwacji należącej do klasy pozytywnej, konieczne jest użycie funkcji sigmoidalnej. Decyzja o przynależności do klasy pozytywnej jest podejmowana za pomocą prostej procedury decyzyjnej – obserwacja jest zaliczana do klasy pozytywnej, jeśli jej prawdopodobieństwo przynależności wynosi co najmniej 0,50. Estymacja wag, czyli parametrów modelu, odbywa się w sposób nadzorowany. Oznacza to, że znana jest prawdziwa klasa dla konkretnej obserwacji. Model ma na celu wykonanie przewidywań klas dla obserwacji, które są jak najbardziej zbliżone do rzeczywistej klasy.



Rysunek 1. Schemat zaimplementowanego algorytmu do celów selekcji cech i klasyfikacji.

Przed właściwą analizą konieczne jest jednak zapoznanie się z charakterystyką danych. Wizualizacja tak złożonych, wielowymiarowych danych sekwencjonowania pojedynczej komórki wymaga wcześniejszej redukcji wymiarowości. Wymiarowość zbiorów danych została zmniejszona przy użyciu procedur analizy głównych składowych (PCA). Narzędzie UMAP zostało użyte w oparciu o wybrane komponenty PCA na zbiorach danych o zredukowanej wymiarowości. Wizualizacja została przeprowadzona metodą nienadzorowaną, aby umożliwić wykrycie ewentualnych struktur w danych, niezależnie od przypisania poszczególnych komórek do próbki kontrolnej lub napromienionej. Tabela 2 przedstawia wyniki analizy skupisk komórek dla trzech różnych progów do testowania układu gęstości komórek.

Tabela 2. Wyniki analizy gęstości komórek wewnątrz wyznaczonych klastrów dla różnych wartości gęstości t.



Na podstawie dwuwymiarowych projekcji UMAP można łatwo wytyczyć oddzielne klastry obserwacji dla analizowanych zbiorów danych. Każda kolekcja składa się zarówno z komórek kontrolnych, jak i napromieniowanych. Separacja poszczególnych klastrów nie jest więc spowodowana obecnością dwóch próbek komórek, ale efektem znacznie silniejszym niż czynnik promieniowania. Analiza ta pozwoliła stwierdzić, że ukryta struktura danych jest związana z wysoką heterogenicznością badanych zbiorów danych. Konieczne jest zatem wykrycie przyczyny zaobserwowanego zjawiska, aby dokonać wymaganych manipulacji w celu monitorowania tego efektu.

Etap identyfikacji profilu genetycznego komórek napromienionych metodami regresji logistycznej oparty jest na zbiorze danych B. Opracowany przebieg pracy polega na selekcji cech za pomocą modelowania w oparciu o metody regresji logistycznej, określeniu istotności cech na podstawie wygenerowanych list cech z wykorzystaniem odpowiedniej metryki, zbudowania modelu końcowego i niezależnego testowania modelu. Algorytm ten został zaprojektowany do zbudowania 50 kompletnych modeli w oparciu o dostarczony zestaw modelowej struktury danych zbioru B. Gdy algorytm zostanie uruchomiony raz, możliwe jest wygenerowanie tylko jednego zestawu wybranych cech; w związku z tym, aby otrzymać 50 założonych modeli, algorytm został uruchomiony 50 razy. System generowania zbiorów został zaprogramowany na początku implementacji. Zadaniem tego systemu jest wylosowanie walidacyjnej struktury danych składającej się z 30% komórek wejściowych dostępnych w modelowej strukturze danych (pozostałych po wydzieleniu zbioru testowego). Ponadto zbiór danych walidujących

składa się z tej samej liczby komórek kontrolnych i napromienionych, aby zapewnić zrównoważony skład komórek. Pozostała część modelowej struktury danych utworzyła strukturę treningową. W wyniku procesu selekcji cech z wykorzystaniem samouczącego się algorytmu opartego na RL uzyskano 50 zestawów cech. Dla każdego modelu oszacowano ważoną dokładność klasyfikacji dla zestawu walidacyjnego. Po wygenerowaniu list wybranych cech, kolejno występujące geny zebrano w jedną zbiorczą listę wraz z ważonymi wartościami dokładności dla zbioru walidacyjnego wśród 50 wygenerowanych modeli. Do przypisania odpowiednich rang cech zastosowano metrykę GeneRank. Wartości te znormalizowano do zakresu 0-1, aby wszystkie dostępne cechy były porównywalne. Wartość progową dla wybranej liczby cech wyznaczono na podstawie metryki GeneRank, gdzie istotna różnica między kolejnymi wartościami była niezauważalna. Pełna lista posortowanych cech obejmowała 159 genów, natomiast po ustaleniu odpowiedniej wartości granicznej pozostawiono 29 genów do dalszej analizy rozpoznawania profilu genetycznego komórek napromienionych. Uznano je za iistotne pod względem zdolności rozpoznawania tych komórek. Po selekcji cech i określeniu istotności genów w problemie rozróżnienia komórek kontrolnych i napromienionych, oszacowano wartości parametrów modelu z wykorzystaniem zaimplementowanego algorytmu. Obliczenia oparto na strukturze modelowej zbioru danych B, a oszacowane wartości parametrów przedstawiono w Tabeli 3.

Tabela 3. Obliczone wartości parametrów dla modelu końcowego..

Intercept	**BAX**	**RPS19P1**	**RPL23AP42**	**RPS27L**	**DDB2**	**TNFSF8**	**CCNG1**
-2,47	0,79	0,25	-0,21	0,74	1,27	0,71	0,48
STAT5A	**LCK**	**TNFRSF10B**	**AQP9**	**CD3D**	**PHPT1**	**AEN**	**LAMP3**
-0,63	-0,29	0,85	0,39	-0,21	0,34	0,90	-0,17
TYMS	**CD40**	**TMEM97**	**RUNX3**	**GZMH**	**MYC**	**CXCL9**	**IL15**
-0,69	-0,33	-0,47	-0,27	-0,20	-0,19	-0,06	-0,40
FYB	**MCM2**	**FLT3**	**LAT**	**TRIB2**	**GAPDH**		
-0,31	0,29	-0,26	-0,25	-0,47	-0,14		

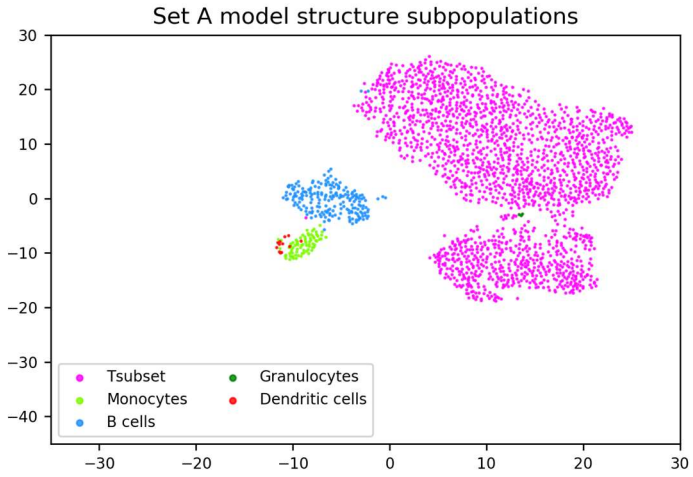
Ostatnim krokiem niezbędnym do zbudowania w pełni funkcjonalnego, dostrojonego modelu było wybranie odpowiedniej wartości progowej prawdopodobieństwa klasyfikacji napromienionych komórek. W tym celu przeprowadzono krok kontroli progu klasyfikacyjnego w oparciu o strukturę testową zestawu danych B, wykorzystując indeks Youdena. Przeprowadzono tę procedurę w celu sprawdzenia, czy po wykorzystaniu nowej wartości progowej prawdopodobieństwa klasyfikacji nastąpiła znacząca zmiana. Aby porównać wyniki, wprowadzono kilka wskaźników jakościowych: prawdziwie pozytywne (TP), prawdziwie negatywne (TN), fałszywie pozytywne (FP), fałszywie negatywne (FN), precyzję, czułość, specyficzność i ważoną jakość klasyfikacji. Aby zweryfikować skuteczność nowej wartości prawdopodobieństwa klasyfikacji komórek napromienionych, wyniki dla domyślnych i nowych progów prawdopodobieństwa klasyfikacji przedstawiono w Tabeli 4.

Tabela 4. Porównanie metryk jakości klasyfikacji dla dwóch wartości progowych prawdopodobieństwa.

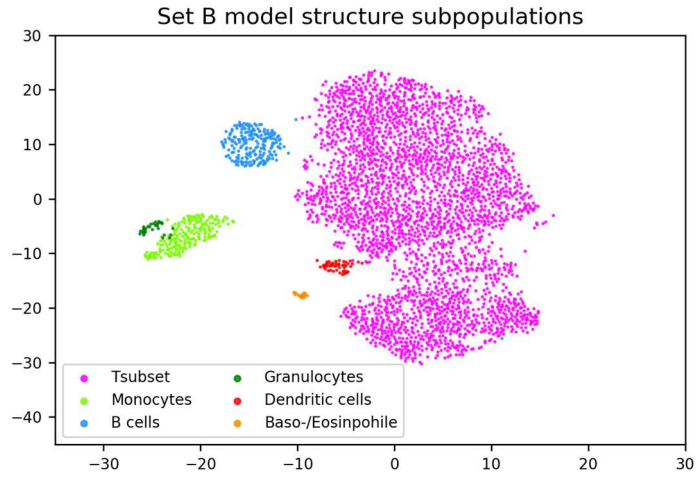
	Próg prawdopodobieństwa	
Metryka jakości	**Ustalony 0,5000**	**Youden 0,7047**
TP	361	351
TN	404	429
FP	48	23
FN	30	40
Precyzja	0,8826	0,9385
Czułość	0,9233	0,8977
Specyficzność	0,8938	0,9491
Ważona jakość	0,9075	0,9253
$F1_{score}$	0,9025	0,9176
Liczba komórek	843	843
Liczba poprawnie zaklasyfikowanych komórek	765	780
Liczba niepoprawnie zaklasyfikowanych komórek	78	63
Niepoprawnie zaklasyfikowane komórki [%]	9,25	7,47

Analiza ta została przeprowadzona w oparciu o strukturę testową zestawu danych B. Oszacowane wartości metryki $F1_{score}$ wskazują na przewagę jakości klasyfikacji dla zmienionego progu klasyfikacji prawdopodobieństwa. Operacje związane z selekcją cech i dopracowaniem ostatecznego modelu pozwoliły uzyskać ważoną jakość klasyfikacji na bardzo wysokim i zadowalającym poziomie, wynoszącym ponad 92%.

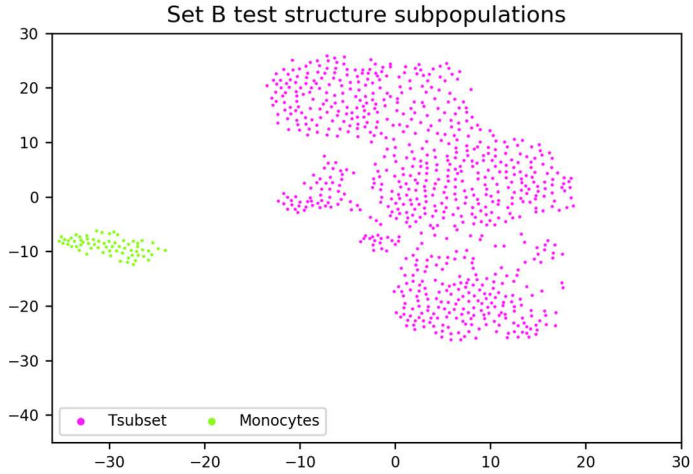
Po szczegółowej analizie genów wchodzących w skład rozpoznanego profilu genetycznego komórek napromienionych okazało się, że nie wszystkie cechy służą do odróżniania komórek napromienionych od kontrolnych. Stwierdzono, że aż 9 z 29 genów profilu genetycznego tych komórek jest odpowiedzialnych za rozróżnianie poszczególnych subpopulacji komórkowych. Stanowią one więc dodatkowe obciążenie dla zbudowanego profilu genetycznego. Ta analiza pozwoliła odkryć krytyczny czynnik korygujący problem rozpoznawania profilu genetycznego komórek napromienionych - wewnętrzną heterogeniczność zbioru danych białych krwinek wpływającą na ostateczną sygnaturę genetyczną komórek napromienionych. Dalszym celem jest wyeliminowanie wewnętrznych różnic między komórkami, które nie wynikają z wpływu czynnika promieniowania, ale wpływają na zachowanie profili genowych tych komórek. Aby określić dokładną przyczynę zmienności, przeprowadzono procedury izolacji poszczególnych subpopulacji krwinek białych. Problem rozpoznawania subpopulacji został podzielony na następujące etapy: selekcja cech, rozpoznawanie klastrów komórkowych za pomocą narzędzia HDBSCAN oraz rozpoznawanie subpopulacji krwinek białych za pomocą genów markerowych charakterystycznych dla oczekiwanych subpopulacji komórek (BD Rhapsody Immune Response Panel Hs). Celem selekcji cech było wybranie zestawu genów, który wykazuje najistotniejszą zmienność w wartościach zliczeń komórek kontrolnych i napromienionych. W tym celu zastosowano miarę współczynnika zmienności (CV). Wartość progowa została oszacowana na podstawie podejścia CV i Gaussian Mixture Modeling. Narzędzie HDBSCAN wykorzystano do podziału komórek na klastry odpowiadające ich zmienności w przestrzennie zredukowanych zbiorach danych. Do określenia zestawu parametrów, który najlepiej oddziela analizowane komórki kontrolne pod względem ich zróżnicowania, wykorzystano miarę efektu omega-kwadrat. Następnie komórki kontrolne i napromienione poddano procedurze grupowania. W wyniku analizy danych ex vivo narzędziem HDBSCAN, komórki podzielono na klastry. Zakłada się, że klastry te odpowiadają wewnętrznej heterogeniczności komórek, którą wykryto w procesie wizualizacji zbiorów danych ex vivo. Etap pozwalający na wykrycie określonych subpopulacji krwinek białych przeprowadzono z wykorzystaniem informacji o genach markerowych charakterystycznych dla poszczególnych subpopulacji białych krwinek. Aby zidentyfikować odpowiednie struktury tych subpopulacji, w tym nawet bardzo słabo reprezentowane typy komórek, wygenerowano wykresy pudełkowe dla każdego klastra komórek i każdego genu markerowego subpopulacji, pokazując szczegółowo rozkład zliczeń. Na podstawie wygenerowanych wykresów pudełkowych, oddzielnie dla komórek kontrolnych i napromienionych, zdecydowano o poszczególnych klastrach należących do odpowiedniej subpopulacji komórek. Klastry oznaczone jako ta sama subpopulacja zostały połączone, tworząc większe struktury odpowiadające wewnętrznej heterogeniczności danych. Wykryte subpopulacje zwizualizowano za pomocą narzędzia UMAP. Końcowe wykresy UMAP z zaznaczonymi zidentyfikowanymi subpopulacjami komórkowymi są pokazane na Rysunku 2, Rysunku 3 i Rysunku 4, odpowiednio dla struktur modelowych zestawu A i zestawu B oraz dla struktury testowej zestawu B.



Rysunek 2. Wizualizacja UMAP dla końcowych subpopulacji białych krwinek dla struktury modelowej zbioru danych A.



Rysunek 3. Wizualizacja UMAP dla końcowych subpopulacji białych krwinek dla struktury modelowej zbioru danych B.



Rysunek 4. Wizualizacja UMAP dla końcowych subpopulacji białych krwinek dla struktury testowej zbioru danych B.

Zestawienie wykrytych subpopulacji krwinek białych przedstawiono w Tabeli 5. Zawiera ona liczbowy i procentowy udział komórek określonej subpopulacji w odniesieniu do analizowanych zbiorów danych.

Tabela 5. Podsumowanie procedury rozpoznawania subpopulacji białych krwinek.

Subpopulacja	Typ komórki	Struktura modelowa zbioru A	Struktura modelowa zbioru B	Struktura testowa zbioru B
Komórki T	Kontrolna	83,88% (1051)	85,56% (1540)	94,91% (429)
	Napromieniona	89,70% (801)	84,66% (1330)	89,26% (349)
Monocyty	Kontrolna	3,67% (46)	6,28% (113)	5,09% (23)
	Napromieniona	3,47% (31)	5,60% (88)	10,74% (42)
Komórki B	Kontrolna	11,25% (141)	5,56% (100)	nie wykryto
	Napromieniona	6,83% (61)	6,36% (100)	nie wykryto
Granulocyty	Kontrolna	0,24% (3)	1,44% (26)	nie wykryto
	Napromieniona	nie wykryto	1,59% (25)	nie wykryto
Komórki dendrytyczne	Kontrolna	0,96% (12)	0,61% (11)	nie wykryto
	Napromieniona	nie wykryto	1,34% (21)	nie wykryto
Bazofile/Eozynofile	Kontrolna	nie wykryto	0,55% (10)	nie wykryto
	Napromieniona	nie wykryto	0,45% (7)	nie wykryto

Rozpoznane subpopulacje białych krwinek bardzo dobrze pokrywają się ze strukturami wygenerowanymi przez nienadzorowane podejście do grupowania danych za pomocą narzędzia UMAP. Udziały poszczególnych subpopulacji w danym zbiorze danych wskazują na istotną przewagę zawartości komórek T we wszystkich analizowanych zbiorach. Możliwe było również wykrycie małolicznych subpopulacji, takich jak granulocyty, komórki dendrytyczne oraz bazofile/eozynofile. Wskazuje to na dużą czułość przeprowadzonej analizy mającej na celu rozpoznanie subpopulacji komórek białych krwinek. Wykrywanie subpopulacji komórkowych ujawniło podstawy heterogeniczności widocznej na dwuwymiarowych wykresach UMAP przy użyciu technik nienadzorowanych. Wykorzystując wcześniej określony profil genetyczny komórek napromienionych i zbudowany na jego podstawie model, można również zdecydować o skuteczności modelu w klasyfikacji obserwacji należących do wybranych subpopulacji krwinek białych. Podsumowanie wyników jakości klasyfikacji dla poszczególnych subpopulacji komórek, na podstawie niezależnej struktury modelowej zbioru danych A, jest przedstawione w Tabeli 6. Przedstawia ona trzy subpopulacje komórek wykryte dla próbek kontrolnych i napromienionych.

Tabela 6. Wartości metryk jakości klasyfikacji dla niezależnego zbioru danych dla trzech rozpoznanych subpopulacji komórkowych.

Metryka jakości	Komórki T	Monocyty	Komórki B
TP	749	20	54
TN	995	33	106
FP	56	13	35
FN	52	11	7
Precyzja	0,9304	0,6061	0,6067
Czułość	0,9351	0,6452	0,8852
Specyficzność	0,9467	0,7174	0,7518
Ważona jakość	0,9417	0,6883	0,7921
$F1_{score}$	0,9328	0,6250	0,7200
Liczba komórek	1852	77	202
Liczba poprawnie zaklasyfikowanych	1744	53	160
Liczba niepoprawnie zaklasyfikowanych	108	24	42
Niepoprawnie zaklasyfikowane [%]	5,83	31,17	20,79

Na podstawie analizy jakości klasyfikacji poszczególnych subpopulacji stwierdzono, że tylko subpopulacja komórek T jest prawidłowo procedowana z zachowaniem zadowalających wyników.

Wszystkie przedstawione wartości wskaźników jakości są wyższe dla subpopulacji komórek T, porównując ją z pozostałymi podtypami komórek białych krwinek. Ważona jakość klasyfikacji dla monocytów i komórek B wynosi odpowiednio 69% i 79%, podczas gdy dla komórek T jest to aż 94%. Znacząco niższe wartości tej metryki dla mniejszościowych subpopulacji, przy jednoczesnym uwzględnieniu miary $F1_{score}$ oznaczają, że model nauczył się specyficznych wzorców adekwatnych dla stanowiących większość komórek T, przywiązując mniejszą wagę do mniej licznych subpopulacji.

Biorąc pod uwagę powyższe etapy analizy, wyodrębniono zestaw najliczniejszej subpopulacji komórek T w celu określenia profilu genetycznego komórek napromienionych. Umożliwiło to wykluczenie stałego czynnika zakłócającego, jakim jest heterogeniczność zbioru danych. Zbudowano 50 modeli z wykorzystaniem zaimplementowanego algorytmu, opartego na znormalizowanych danych dotyczących komórek T. Wśród wygenerowanych modeli znalazły się 54 cechy unikalne. Każdemu genowi, który wystąpił co najmniej jeden raz w 50 modelach, przypisano odpowiednią wartość metryki GeneRank. Dla liczby istotnych cech wyznaczono wartość progową. Oszacowane wartości parametrów modelu dla 21 wybranych cech przedstawiono w Tabeli 7. Parametry modelu obliczono na podstawie struktury modelowej zestawu danych B, składającej się z 2870 znormalizowanych komórek T.

Tabela 7. Obliczone wartości parametrów modelu końcowego dla znormalizowanych danych komórek T.

Intercept	**RPS19P1**	**RPL23AP42**	**BAX**	**DDB2**	**HLA-A**	**RPS27L**	**PHPT1**
-4,47	1,76	-1,44	0,60	0,22	-0,75	0,40	0,32
MYC	**CCNG1**	**FYB**	**CD52**	**CD74**	**TNFSF8**	**THBS1**	**STAT1**
-0,54	0,32	-0,39	-0,39	-0,22	0,10	-0,30	-0,31
TRIB2	**LCK**	**AEN**	**CDKN1A**	**STAT5A**	**CHI3L1**		
-0,32	-0,33	0,06	0,05	-0,29	0,11		

Ostatnim etapem budowy kompletnego modelu było wyznaczenie nowej wartości progowej prawdopodobieństwa klasyfikacji komórek napromienionych z wykorzystaniem indeksu Youdena. Wartość ta została wyznaczona na podstawie struktury testowej zestawu danych B. Skuteczność wykorzystania nowej wartości progowej prawdopodobieństwa do celów klasyfikacji komórek kontrolnych i napromienionych, porównano z wcześniej ustaloną wartością prawdopodobieństwa klasyfikacji 0,50. Wyniki te przedstawiono w Tabeli 8.

Tabela 8. Porównanie metryk jakości klasyfikacji dla dwóch wartości prawdopodobieństwa klasyfikacji.

	Próg prawdopodobieństwa	
Metryka jakości	**Ustalony 0,5000**	**Youden 0,5124**
TP	323	323
TN	405	407
FP	24	22
FN	26	26
Precyzja	0,9308	0,9362
Czułość	0,9255	0,9255
Specyficzność	0,9440	0,9487
Ważona jakość	0,9357	0,9383
$F1_{score}$	0,9282	0,9308
Liczba komórek	778	778
Liczba poprawnie zaklasyfikowanych komórek	728	730
Liczba niepoprawnie zaklasyfikowanych komórek	50	48
Niepoprawnie zaklasyfikowane komórki [%]	6,43	6,17

Drugie podejście do uczenia maszynowego miało również na celu określenie profilu genetycznego komórek napromienionych oraz klasyfikację komórek napromienionych i kontrolnych na podstawie zbudowanego modelu. Ta część rozprawy doktorskiej opiera się na wcześniej

wyselekcjonowanym i znormalizowanym podzbiorze danych subpopulacji komórek T. Przeprowadzona analiza obejmowała kilka głównych aspektów, takich jak dobór struktury klasyfikatora w oparciu o metody sieci neuronowych, selekcję cech metodą filtrów, zbudowanie modelu na podstawie wybranego zbioru genów oraz przetestowanie modelu na strukturze testowej zbioru danych B. Model sieci neuronowych został zbudowany na strukturze modelowej zbioru danych B dla wszystkich 406 dostępnych cech. W tym celu wykorzystano funkcję Sequential() z biblioteki TensorFlow.Keras [13]. Aby zachować możliwość wyboru najlepszego zestawu cech wykorzystano funkcję ModelCheckpoint() z tej samej biblioteki w środowisku Python. Wprowadzona funkcjonalność pozwala na bieżąco monitorować zmieniające się parametry modelu i zapisywać tylko te, które spełniają zadany warunek. W analizie warunkiem tym było uzyskanie najwyższej wartości metryki $F1_{score}$ na podstawie zbioru treningowego. Model został wytrenowany przy użyciu funkcji fit() z informacją o pochodzeniu komórek, liczbą epok ustawioną na zadaną wartość 150 oraz zestawem walidacyjnym z anotacjami o pochodzeniu obserwacji. Po dopasowaniu modelu do zadanego zbioru treningowego i zwróceniu wartości parametrów modelu sieci neuronowej można było przejść bezpośrednio do procedury selekcji cech. W tym celu wykorzystano popularną i szeroko stosowaną bibliotekę shap [14] do wyjaśnienia problemu „czarnej skrzynki" w sieciach neuronowych. Podstawowym zadaniem narzędzia Shapley Additive exPlanations (SHAP) jest interpretacja modelu uczenia maszynowego w kontekście uczenia modeli i wyników predykcji [15]. W wyniku zastosowania opisanych procedur uzyskano miarę istotności poszczególnych cech nazwaną ShapScore. Na podstawie dziesięciu wybranych genów zbudowano ostateczny model z oszacowanymi wartościami parametrów. Stworzony model oparty na sieci neuronowej został poddany procedurze testowej. Wartości metryk jakości klasyfikacji przedstawiono w Tabeli 9.

Tabela 9. Wartości metryk jakości klasyfikacji dla struktury testowej zbioru danych B.

Metryka jakości	**Wartość metryki jakości klasyfikacji**
TP	308
TN	402
FP	27
FN	41
Precyzja	0,9194
Czułość	0,8825
Specyficzność	0,9371
Ważona jakość	0,9098
$F1_{score}$	0,9016
Liczba komórek	778
Liczba poprawnie zaklasyfikowanych komórek	710
Liczba niepoprawnie zaklasyfikowanych komórek	68
Niepoprawnie zaklasyfikowane komórki [%]	8,74

Osiągnięte przez klasyfikator metryki jakości klasyfikacji są zadowalające, o czym świadczą wysokie wartości precyzji i specyficzności powyżej 0,90 oraz czułości powyżej 0,88. Klasyfikator bardzo dobrze radzi sobie z rozpoznawaniem komórek kontrolnych i nieco gorzej z identyfikacją komórek napromienionych, o czym świadczy wyższy poziom przypadków FN, co niekorzystnie wpływa na miarę czułości. Ważona jakość klasyfikacji została określona na prawie 91%, przy wysokiej wartości miary $F1_{score}$ powyżej 0,90. Ogólny odsetek błędnie zaklasyfikowanych komórek wyniósł ponad 8,50%.

Rozpoznawanie subpopulacji białych krwinek

W związku z przeprowadzoną serią analiz **potwierdzono hipotezę 1. mówiącą o tym, że połączenie metod inżynierii cech i zaawansowanych technik redukcji wymiarowości z nienadzorowanymi algorytmami grupowania pozwala na skuteczną identyfikację podtypów krwinek białych w danych sekwencjonowania RNA pojedynczej komórki.** Przebieg analizy i zawarte w niej procedury umożliwiły wykrycie subpopulacji krwinek białych i określenie przyczyny obserwowanej zmienności komórkowej niezwiązanej z czynnikiem promieniowania. Jako źródło tej zmienności wskazano dużą wewnętrzną heterogeniczność analizowanych danych. Wnioski zostały poparte zastosowaniem wykrytych subpopulacji do predefiniowanych klastrów wyodrębnionych przy użyciu nienadzorowanych technik uczenia się UMAP. W analizowanych zbiorach danych ex vivo osiągnięto niemal idealne dopasowanie. W przypadku tego typu danych wizualizacja przestrzennie rozmieszczonych komórek i kodowanie kolorami odpowiednich subpopulacji w pełni pokrywały się z klastrami. Ścieżka analizy umożliwiła zatem wykrycie kilku subpopulacji krwinek białych, z których zdecydowaną większość, bo aż 85-90% zbiorów danych, stanowiły komórki T. Subpopulacja monocytów została również wykryta dla wszystkich podzbiorów danych. W przypadku struktur modelowych zestawów danych A i B wykryto również mniej liczne subpopulacje, takie jak komórki B, granulocyty i komórki dendrytyczne. W przypadku modelowej struktury zbioru danych B możliwe było również wykrycie małolicznej subpopulacji bazofilów/eozynofilów, stanowiącej około 0,5% analizowanych danych. Wyniki te wskazują na wysoką wydajność i czułość proponowanego podejścia do wykrywania subpopulacji dla danych z eksperymentów sekwencjonowania pojedynczej komórki.

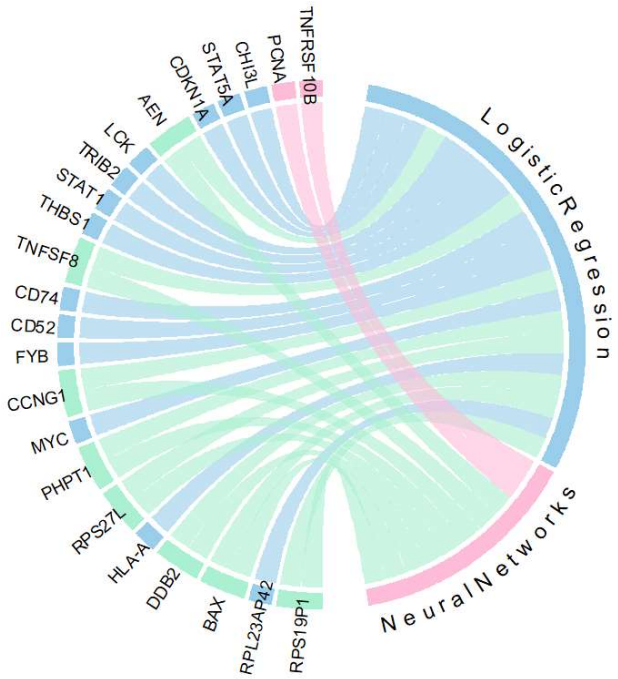
Profil genetyczny komórek napromienionych w środowisku ex vivo

Jednym z najcenniejszych, pod względem technicznym, aspektów tej rozprawy jest opracowanie przebiegu procedur selekcji cech. Co więcej, komponenty tego rozwiązania mogą nie tylko wykrywać istotne zmiany w profilu genetycznym komórek napromienionych. Może być ono również z powodzeniem wykorzystywane do celów klasyfikacji, w oparciu o utworzony wcześniej pełny model, dla zbioru komórek kontrolnych i napromienionych. Zaimplementowany schemat pracy umożliwia stabilne prowadzenie analizy, z gwarantowanym wyborem najbardziej krytycznych parametrów, które powinny być użyte w przypadku konieczności ich manipulacji ze względu na różne cele i specyficzne założenia przeprowadzanej analizy. Za pomocą tej implementacji możliwe jest przeprowadzenie wszystkich niezbędnych etapów budowy modelu, testowania i wykorzystania skonstruowanego modelu do klasyfikacji komórek zewnętrznych zbiorów danych. Rozpoznanie profilu genetycznego dla całej dostępnej puli genowej pozwoliło na wykrycie czynnika wartego odnotowania, a mianowicie konieczności przeprowadzenia dokładnej i dopracowanej procedury selekcji cech. Wykryty profil genetyczny komórek napromienionych z takiego zestawu danych umożliwił rozpoznanie genów odpowiedzi radiacyjnej. Profil ten był jednak zanieczyszczony cechami nieodpowiadającymi za rozróżnianie komórek napromienionych. Wykryte geny, takie jak AQP9, CD3D, FYB, LAT, LAMP3, LCK i TRIB2, wybrane na etapie selekcji cech, miały przede wszystkim wykrywać różnice między subpopulacjami białych krwinek obecnymi, a wcześniej niekontrolowanymi w zbiorze danych. Aby przeprowadzić odpowiednią ścieżkę analizy danych pod kątem rozpoznania prawidłowego profilu genetycznego komórek napromienionych w środowisku ex vivo, usunięto zmienność związaną z występowaniem subpopulacji komórek, odfiltrowując jedynie subpopulację komórek T, która stanowiła zdecydowaną większość wszystkich analizowanych komórek. Na podstawie zastosowanego przepływu pracy udowodniono **hipotezę 2. mówiącą o tym, że zaproponowany inteligentny i warstwowy algorytm konstrukcji zbioru treningowego wspomaga system klasyfikacji, zwłaszcza w przypadku zbiorów heterogenicznych.** Takie podejście pozwoliło zbadać czym różnią się komórki kontrolne od napromienionych bez dodatkowych czynników zakłócających. W składzie genów modelu zbudowanego na podstawie znormalizowanych zliczeń komórek T znaleziono prawie wyłącznie geny opisane w doniesieniach literaturowych jako geny odpowiedzi na promieniowanie. Aż 18 z 21 takich genów znalazło się w modelu. Pozostałe trzy geny, HLA-A, CD52 i TRIB2, były związane ze szlakami procesów biologicznych odpowiedzialnych głównie za odpowiedź komórkową na szkodliwy czynnik zewnętrzny. Po usunięciu przyczyny heterogeniczności zbioru, tj. obecności różnych subpopulacji i skupieniu pracy klasyfikatora wyłącznie na większościowej klasie subpopulacji komórek T, możliwe było wykrycie jedynej przyczyny różnic między komórkami, tj. obecności dwóch typów komórek – kontrolnej i napromienionej. Wybrany zestaw cech w pełni odpowiada, założonemu w niniejszej

rozprawie doktorskiej, problemowi rozpoznawania profilu genetycznego komórek napromienionych. Zidentyfikowane geny odpowiedzi na promieniowanie wraz z przypisanymi wartościami parametrów przedstawiono w Tabeli 7. Skonstruowany model profilu genetycznego komórek napromienionych pozwolił na uzyskanie wysokiej ważonej jakości klasyfikacji dla zestawu testowego, wynoszącej 93,83%. Osiągnięto również doskonałą precyzję, czułość i specyficzność o wartościach 0,9362, 0,9255 i 0,9487. Odsetek błędnie zaklasyfikowanych komórek wyniósł 6,17% dla zbioru testowego, co oznacza, że tylko 48 z 778 obserwacji nie zostało poprawnie przypisanych.

Schematy pracy oparte o regresję logistyczną i sieci neuronowe

Porównanie zastosowanych metod uczenia maszynowego w zakresie wyboru cech i klasyfikacji napromienionych i kontrolnych komórek jest wyzwaniem. Należy bowiem wziąć pod uwagę wiele czynników, wśród których najważniejsze to stopień złożoności analizy, koszty obliczeniowe, czas potrzebny obliczeń oraz poprawność i interpretowalność uzyskanych wyników. Biorąc pod uwagę złożoność analizy przeprowadzonej z wykorzystaniem metod regresji logistycznej i sieci neuronowych nie można bezpośrednio porównać obu przepływów pracy. Jest to spowodowane tym, że analizy związane z metodami regresji logistycznej zostały w całości zaimplementowane na potrzeby niniejszej rozprawy doktorskiej. Wykorzystanie sieci neuronowych wiązało się wykorzystaniem gotowych i publicznie dostępnych funkcji. Ponadto, dla metod regresji logistycznej przeprowadzono dokładniejsze badanie związane z doborem cech, polegające na 50-krotnym losowaniu zbioru treningowego, co zapewniło większą zmienność i uogólnienie problemu w porównaniu z podejściem wykorzystującym sieci neuronowe. Dodatkowo, w przypadku sieci neuronowych zastosowano narzędzie wspomagające wyjaśnienie struktury modelu co było elementem dodatkowym, nie procedowanym w podejściu RL. Poza różnicami w procesie selekcji cech, czyli wykorzystaniem metod opakowujących, oba podejścia były spójne w dalszym stosowaniu metody z dziedziny filtrów. W obu przypadkach oszacowano liczbę istotnych cech i na ich podstawie zbudowano ostateczny model. Kolejnym krytycznym czynnikiem są koszty obliczeniowe i czas potrzebny do uzyskania wyników. W tym kontekście modelowanie z wykorzystaniem sieci neuronowych przewyższa opracowane podejście oparte na regresji logistycznej. Uzyskanie 50 modeli opartych na RL w celu wybrania cech, wymagało istotnie dużo czasu i intensywnych obliczeń. W przypadku sieci neuronowych zbudowano tylko jeden model, na podstawie którego wnioskowano o istotności poszczególnych cech. Tym samym koszty obliczeniowe były stosunkowo mniejsze, a czas uzyskania wyników nieporównywalnie krótszy. Wymienione aspekty i wnioski nie wynikają z charakteru działania poszczególnych rozwiązań uczenia maszynowego; porównanie to opiera się na rozwiązaniach zastosowanych w niniejszej rozprawie i nie może być wykorzystywane do wyciągania ogólnych wniosków na temat skuteczności zastosowanych metod uczenia maszynowego. Opisane czynniki nie mogą jednak przeważyć nad znaczeniem tego ostatniego, tj. interpretowalności i poprawności wyników. Oba procesy pozwoliły uzyskać profil genetyczny komórek napromienionych w środowisku ex vivo. Co więcej, aż osiem wykrytych genów jest wspólnych dla obu przepływów pracy, co pokazano na Rysunku 5. Wskazuje to na dużą spójność w rozpoznawaniu odpowiednich struktur zmian komórkowych pod wpływem czynnika promieniowania jonizującego. Biorąc pod uwagę złożoność zidentyfikowanego modelu komórek napromienionych, w metodach regresji logistycznej istotnych było 21 cech. Dla porównania, w sieciach neuronowych



Rysunek 5. Porównanie profilu genetycznego komórek napromienionych dla metod regresji logistycznej i sieci neuronowych.

wykorzystano tylko dziesięć wyselekcjonowanych genów. Geny te dla obu zastosowanych podejść przedstawiono w Tabeli 10. Ponadto, podano odpowiednie odniesienia literaturowe dotyczące genów odpowiedzi radiacyjnej. Na podstawie aktualnych źródeł literaturowych aż 18 z 21 genów wykrytych metodami regresji logistycznej określa się jako geny odpowiedzi radiacyjnej. Z kolei w przypadku genów wykrytych za pomocą sieci neuronowych, wszystkie zostały opisane jako geny związane z odpowiedzią na promieniowanie.

Tabela 10. Sygnatura genowa komórek napromienionych, rozpoznana z wykorzystaniem metod regresji logistycznej oraz sieci neuronowych.

Oparte o metody regresji logistycznej		**Oparte o metody sieci neuronowych**	
Gen	**Odpowiedź radiacyjna**	**Gen**	**Odpowiedź radiacyjna**
RPS19P1	Radiation response [18]	RPS19P1	Radiation response [18]
RPL23AP42	Radiation response [18]	BAX	Radiation response [17]
BAX	Radiation response [17]	DDB2	Radiation response [20]
DDB2	Radiation response [20]	PHPT1	Radiation response [22]
HLA-A	-	RPS27L	Radiation response [19]
RPS27L	Radiation response [19]	CCNG1	Radiation response [26]
PHPT1	Radiation response [22]	AEN	Radiation response [19]
MYC	Radiation response [23]	PCNA	Radiation response [19]
CCNG1	Radiation response [26]	TNFRSF10B	Radiation response [21]
FYB	Radiation response [24]	TNFSF8	Radiation response [25]
CD52	-		
CD74	Radiation response [29]		
TNFSF8	Radiation response [25]		
THBS1	Radiation response [30]		
STAT1	Radiation response [31]		
TRIB2	-		
LCK	Radiation response [28]		
AEN	Radiation response [19]		
CDKN1A	Radiation response [25]		
STAT5A	Radiation response [27]		
CHI3L	Radiation response [32]		

Przeprowadzono również dodatkową analizę w celu ustalenia, czy jeden z dwóch wykrytych profili genetycznych jest bardziej uniwersalny i czy można je przenieść do innej metody klasyfikacji bez utraty jakości klasyfikacji. Porównano trzy podejścia selekcji cech. Pierwszym był brak selekcji cech, czyli zbudowania modelu na wszystkich 406 genach dostępnych dla struktury modelowej zbioru danych B. W drugim i trzecim wariancie wykorzystano struktury genetyczne rozpoznane w wyniku selekcji cech za pomocą metod regresji logistycznej i sieci neuronowych. Struktury te zawierały odpowiednio 21 i 10 wybranych genów. Modele zostały następnie przetestowane z wykorzystaniem podejścia opartego na regresji logistycznej i sieciach neuronowych, tj. każdy model był testowany dwukrotnie, za każdym razem w oparciu o inną metodologię uczenia maszynowego. W Tabeli 11 opisano trzy podejścia do wyboru modeli porównane między dwiema metodami uczenia maszynowego.

Tabela 11. Porównanie jakości klasyfikacji dla trzech podejść selekcji cech dla obu zastosowanych metod uczenia maszynowego.

Podejście selekcji modelu	**Długość sygnatury**	**Regresja logistyczna**		**Sieci Neuronowe**	
		w. jakość	**$F1_{score}$**	**w. jakość**	**$F1_{score}$**
Brak selekcji	406	0,8856	0,8712	0,8638	0,8453
Selekcja w przód dla regresji logistycznej	21	0,9383	0,9308	0,9369	0,9305
ShapScore dla sieci neuronowych	10	0,9165	0,9057	0,9098	0,9016

Wartości metryk jakości klasyfikacji wyznaczono na podstawie struktury testowej zbioru danych B. Brak selekcji cech dla obu analizowanych metod uczenia maszynowego skutkował najgorszymi wynikami na przestrzeni przedstawionych miar jakości klasyfikacji. Model ten był bardzo skomplikowany i zawierał 406 cech, co wymagało określenia wartości parametrów modelu na całej dostępnej przestrzeni cech. Biorąc pod uwagę modele zbudowane w oparciu o selekcję cech dla obu podejść uczenia maszynowego, najlepsze wyniki osiągnięto dla modelu zbudowanego w oparciu o metody regresji logistycznej, a także przetworzone metodami regresji logistycznej. Ważona jakość klasyfikacji wyniosła prawie 94%, przy wartości miary $F1_{score}$ powyżej 0,93. Dla tak zbudowanego modelu i klasyfikacji komórek za pomocą sieci neuronowych uzyskano porównywalnie zadowalające wyniki, biorąc pod uwagę wartość metryki $F1_{score}$ powyżej 0,93 i ważoną jakość klasyfikacji powyżej 93,5%. Model zbudowany w oparciu o metody sieci neuronowych osiągnął znacznie gorsze wyniki, niż wcześniej opisany model oparty na regresji logistycznej. Zarówno dla klasyfikacji z wykorzystaniem metod regresji logistycznej, jak i sieci neuronowych wartość metryki $F1_{score}$ wyniosła powyżej 0,90, a ważona jakość klasyfikacji około 91%.

Przedstawione w rozprawie doktorskiej tezy nie są trywialnymi aspektami analiz danych wysokowymiarowych i interpretacji powstałych rozwiązań. Co więcej, dana tematyka pozostawia wiele do zaoferowania w kontekście możliwości zastosowania usprawnień i wyeliminowania konieczności ingerencji człowieka na poszczególnych etapach. Rozpoznanie subpopulacji komórkowych jest aspektem, który z pewnością pozytywnie wpływa na identyfikację prawidłowego profilu genetycznego komórek napromienionych. Weryfikacja profilu genetycznego, zbudowanego w oparciu o zbiór heterogeniczny ujawniła obecność cech, które nie są związane z problemem badawczym, a wynikają z braku pełnej kontroli nad zjawiskami i zależnościami zachodzącymi w analizowanym zbiorze danych. Profil genetyczny komórek napromienionych, zbudowany na zbiorze po usunięciu wewnętrznej niejednorodności, tj. po odfiltrowaniu subpopulacji komórek T, umożliwił wykrycie tych genów, które wpływają na rozpoznawanie komórek kontrolnych i napromienionych. W ostatecznym modelu, zbudowanym w oparciu o metody regresji logistycznej, nie wykryto genów odpowiedzialnych za rozpoznawanie zmienności innej niż ta wynikająca wprost z napromieniowania frakcji komórkowej. Analiza porównawcza przeprowadzona dla dwóch metod uczenia maszynowego, regresji logistycznej oraz sieci neuronowych, pozwoliła na ustalenie swoistej klamry zamykającej staranne rozważania nad tezami. Ta analiza wyjaśniła, że pomimo zastosowania różnych metod, geny odpowiedzi na promieniowanie są w znacznej większości wspólne dla tych podejść. Ponadto osiem genów wspólnych dla obu badań, tj. AEN, TNFSF8, CCNG1, PHPT1, RPS27L, DDB2, BAX i RPS19P1, jest opisanych w dostępnej i aktualnej literaturze jako dobrze znane geny odpowiedzi na promieniowanie jonizujące.

Istnieje wiele możliwości zastosowania proponowanych podejść związanych z uczeniem maszynowym, opisaną ścieżką rozpoznawania subpopulacji oraz wykorzystaniem ostatecznego modelu profilu genetycznego komórek napromienionych. Zaimplementowany algorytm uczenia maszynowego, oparty na metodach regresji logistycznej, jest prosty w użyciu i pozwala na bardzo zorientowaną na klasy redukcję wymiarowości cech do tych najistotniejszych. Ponadto algorytm ten składa się z panelu parametrów, który można dostosować do potrzeb analizy. Dodatkowo, zaletą zaproponowanego samouczącego się algorytmu jest możliwość zastosowania go do problemów dwuklasowych dla danych biologicznych i dowolnych innych dostępnych danych będących przedmiotem zainteresowania.

Ze względu na bardzo duże pokrycie rozpoznanego profilu genetycznego komórek napromienionych w środowisku ex vivo z aktualnymi doniesieniami literaturowymi, możliwe jest również wykorzystanie go w innych zestawach danych z eksperymentów sekwencjonowania pojedynczej komórki w celu odróżnienia komórek kontrolnych i napromienionych. Warto również podkreślić możliwość wykorzystania pełnego modelu w odniesieniu do wyższych dawek zastosowanego promieniowania jonizującego. Takie podejście wymaga jednak zbadania, jak zaproponowany model radzi sobie z większym zróżnicowaniem klasy kontrolnej i napromienionej. Interesującym podejściem jest porównanie działania zaproponowanego modelu genetycznego napromienionych komórek przy dawkach większych niż 1 Gy, a także przy dawkach poniżej 1 Gy. Analiza porównawcza pozwoliłaby określić uniwersalność modelu w pełnym spektrum dawki pochłoniętej lub uzyskać zadowalające wyniki w określonym zakresie zastosowanych dawek promieniowania jonizującego.

Ze względu na coraz częstsze stosowanie narzędzi generujących dane wysokowymiarowe, prezentowane podejścia są wysoce uniwersalne pod względem różnorodności i składu danych.

Wykorzystanie narzędzi usprawniających pracę analityków, zarówno pod względem czasu, jak i ograniczenia możliwości popełnienia błędu ludzkiego, jest w dzisiejszych czasach coraz bardziej potrzebne i doceniane. Zaprezentowane w rozprawie doktorskiej schematy i rozwiązania mają również bardzo duży potencjał rozwojowy, stanowiąc doskonałą podstawę do przyszłych rozważań.

Bibliografia

[1] J. K. De Kanter, P. Lijnzaad, T. Candelli, T. Margaritis and F. C. Holstege, "CHETAH: a selective, hierarchical cell type identification method for single-cell RNA sequencing," *Nucleic acids research, 47(16),* pp. e95-e95, 2019.

[2] "White Blood Cell Count (WBC) and Differential," [Online]. Available: https://www.rnceus.com/cbc/cbcwbc.html. [Accessed 09 09 2022].

[3] R. C. Wilkins, D. Wilkinson, H. P. Maharaj, P. V. Bellier, M. B. Cybulski and J. R. N. McLean, "Differential apoptotic response to ionizing radiation in subpopulations of human white blood cells," *Mutation Research/Genetic Toxicology and Environmental Mutagenesis, 513(1-2),* pp. 27-36, 2002.

[4] K. Sieradzka and J. Polańska, "Feature selection methods for classification purposes," *Recent Advances in Computational Oncology and Personalized Medicine Volume 2: The challenges of the future, Publishing House of the Silesian University of Technology,* pp. 169-189, 2022.

[5] A. Jović, K. Brkić and N. Bogunović, "A review of feature selection methods with applications.," *In 2015 38th international convention on information and communication technology, electronics and microelectronics (MIPRO),* pp. 1200-1205, 2015.

[6] G. Chandrashekar i F. Sahin, „A survey on feature selection methods.," *Computers & Electrical Engineering, 40(1),* pp. 16-28, 2014.

[7] M. Mera-Gaona, D. M. López, R. Vargas-Canas and U. Neumann, " Framework for the ensemble of feature selection methods," *Applied Sciences, 11(17),* p. 8122, 2021.

[8] A. J. Ferreira and M. A. Figueiredo, "Efficient feature selection filters for high-dimensional data," *Pattern recognition letters, 33(13),* pp. 1794-1804, 2012.

[9] P. Křížek, J. Kittler and V. Hlaváč, "Improving stability of feature selection methods. In International Conference on Computer Analysis of Images and Patterns," *Springer, Berlin, Heidelberg,* pp. 929-936, 2007.

[10] L. McInnes, J. Healy and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426,* 2018.

[11] R. J. Campello, D. Moulavi and J. Sander, "Density-based clustering based on hierarchical density estimates.," in *Pacific-Asia conference on knowledge discovery and data mining (pp. 160-172)*, Springer, Berlin, Heidelberg, 2013.

[12] D. Jurafsky and J. H. Martin, "Logistic Regression," in *Speech and Language Processing, Chapter 5*, Draft of January 7, 2023.

[13] "The Sequential model," [Online]. Available: https://www.tensorflow.org/api_docs/python/tf/keras/Sequential. [Accessed 10 02 2023].

[14] "Welcome to the SHAP documentation," [Online]. Available: https://shap.readthedocs.io/en/latest/. [Accessed 10 02 2023].

[15] Y. G. Lee, J. Y. Oh, D. Kim and G. Kim, "SHAP Value-Based Feature Importance Analysis for Short-Term Load Forecasting," *Journal of Electrical Engineering & Technology,* pp. 1-10, 2022.

[16] "GeneCards: The Human Gene Database," [Online]. Available: https://www.genecards.org/. [Accessed 31 08 2022].

[17] H. Budworth, A. M. Snijders, F. Marchetti, B. Mannion, S. Bhatnagar, E. Kwoh and A. J. Wyrobek, "DNA repair and cell cycle biomarkers of radiation exposure and inflammation stress in human blood.," *PloS one, 7(11), e48619,* 2012.

[18] L. Cruz-Garcia, G. O'Brien, B. Sipos, S. Mayes, M. I. Love, D. J. Turner and C. Badie, "Generation of a transcriptional radiation exposure signature in human blood using long-read nanopore sequencing," *Radiation research, 193(2),* pp. 143-154, 2020.

[19] M. Moreno-Villanueva, Y. Zhang, A. Feiveson, B. Mistretta, Y. Pan, S. Chatterjee and H. Wu, "Single-cell RNA-sequencing identifies activation of TP53 and STAT1 pathways in human T lymphocyte subpopulations in response to ex vivo radiation exposure," *International journal of molecular sciences, 20(9),* p. 2316, 2019.

[20] H. Kaatsch, B. V. Becker, S. Schüle, P. Ostheim, K. Nestler, J. Jakobi and R. Ullmann, "Gene expression changes and DNA damage after ex vivo exposure of peripheral blood cells to various CT photon spectra," *Scientific Reports, 11(1),* pp. 1-9, 2021.

[21] E. Macaeva, M. Mysara, W. H. De Vos, S. Baatout and R. Quintens, "Gene expression-based biodosimetry for radiological incidents: Assessment of dose and time after radiation exposure," *International Journal of Radiation Biology, 95(1),* pp. 64-75, 2019.

[22] A. Tichy, S. Kabacik, G. O'Brien, J. Pejchal, Z. Sinkorova, A. Kmochova and C. Badie, "The first in vivo multiparametric comparison of different radiation exposure biomarkers in human blood," *PLoS One, 13(2), e0193412,* 2018.

[23] S. A. Ghandhi, L. Smilenov, I. Shuryak, M. Pujol-Canadell and S. A. Amundson, "Discordant gene responses to radiation in humans and mice and the role of hematopoietically humanized mice in the search for radiation biomarkers," *Scientific reports, 9(1),* pp. 1-13, 2019.

[24] H. Lyng, K. S. Landsverk, E. Kristiansen, P. M. DeAngelis, A. H. Ree, O. Myklebost and T. Stokke, "Response of malignant B lymphocytes to ionizing radiation: gene expression and genotype," *International journal of cancer, 115(6),* pp. 935-942, 2005.

[25] M. Iwakawa, T. Ohno, K. Imadome, M. Nakawatari, K. I. Ishikawa, M. Sakai and T. Imai, "The radiation-induced cell-death signaling pathway is activated by concurrent use of cisplatin in sequential biopsy specimens from patients with cervical cancer," *Cancer biology & therapy, 6(6),* pp. 905-911, 2007.

[26] L. Cruz-Garcia, G. O'brien, E. Donovan, L. Gothard, S. Boyle, A. Laval and C. Badie, "Influence of confounding factors on radiation dose estimation in in vivo validated transcriptional biomarkers," *Health physics, 115(1),* p. 90, 2018.

[27] C. Girardi, C. De Pittà, S. Casara, G. Sales, G. Lanfranchi, L. Celotti and M. Mognato, "Analysis of miRNA and mRNA expression profiles highlights alterations in ionizing radiation response of human lymphocytes under modeled microgravity," *PLoS One, 7(2), e31293,* 2012.

[28] Y. Feng, Z. Wang, N. Yang, S. Liu, J. Yan, J. Song and Y. Zhang, "Identification of biomarkers for cervical cancer radiotherapy resistance based on RNA sequencing data," *Frontiers in Cell and Developmental Biology, 9,* p. 724172, 2021.

[29] Z. C. Fu, F. M. Wang and J. M. Cai, "Gene expression changes in residual advanced cervical cancer after radiotherapy: indicators of poor prognosis and radioresistance?," *Medical Science Monitor: International Medical Journal of Experimental and Clinical Research, 21,* p. 1276, 2015.

[30] M. A. Chaudhry, "Analysis of gene expression in normal and cancer cells exposed to [gamma]-radiation," *BioMed Research International,* 2008.

[31] A. C. Wilkins, E. C. Patin, K. J. Harrington and A. A. Melcher, "The immunological consequences of radiation-induced DNA damage," *The Journal of Pathology, 247(5),* pp. 606-614, 2019.

[32] Y. C. Kim, M. Barshishat-Kupper, E. A. McCart, G. P. Mueller and R. M. Day, "Bone marrow protein oxidation in response to ionizing radiation in C57BL/6J mice," *Proteomes, 2(3),* pp. 291-302, 2014.