Aspekt	Wnioski	Artykuł/Źródło
Manipulacja zachowaniem modeli specjalnie przygotowanymi aktywacjami	Można łatwo badać i wpływać na zachowanie modeli stosując trzystopniowe podejście: 1. zebranie aktywacji z warstw ukrytych, 2. zmodyfikowanie wejść i ponowne zebranie tym razem zmodyfikowanych aktywacji, 3 kontrolowana manipulacja zachowaniem modeli podstawiając za wyjście każdej warstwy ukrytej jej odpowiadające zmodyfikowane aktywacje podczas inferencji	Activation Patching - https://nnsight.net/notebooks/tutorials/causal_mediation_analysis/activation_patching/
Optymalizacja czasu wykonania procesu Activation Patching	Zamiast wykonywać oddzielny forward pass dla każdej zastępowanej paczki aktywacji z danej warstwy ukrytej wykonuje się dwa przejścia przez model w przód (jedno ze zwykłymi danymi, drugie ze zmodyfikowanymi) oraz jedno przejście w tył (backward pass) stosując liniowe aproksymacje gradientów do ustalenia wpływu zmodyfikowanych aktywacji na wyjścia.	Attribution Patching - https://nnsight.net/notebooks/tutorials/causal_mediation_analysis/attribution_patching/
Uczenie się wysokopoziomowej reprezentacji cech zawartych w aktywacjach z warstw ukrytych	Wykorzystując rzadkie autoenkodery (np. VAE z niewielką przestrzenią ukrytą) można się nauczyć reprezentacji cech zawartych w aktywacjach. Pozwala to na późniejszą wizualizację wpływu poszczególnych neuronów z aktywacji na wyodrębnione, wysokopoziomowe cechy.	Dictionary Learning - https://nnsight.net/notebooks/tutorials/steering/dict_learning/
Wykorzystanie spektralnych oraz przestrzennych cech obiektów z wejść sieci wzbogacając tym samym możliwe do	Wedle artykułu wykorzystanie całego modelu 3D, z nie tylko zdjęć skanu, pozwala osiągać lepsze wyniki w predykcji.	Singh, S.P.; Wang, L.; Gupta, S.; Goli, H.; Padmanabhan, P.; Gulyás, B. “3D Deep Learning on Medical Images: A Review”.

zaobserwowania wysokopoziomowe cechy, którymi można manipulować	Wykorzystywany w artykule model jest siecią konwolucyjną 3D, która jednocześnie jest w stanie wyodrębniać sprektralne oraz przestrzenne zależności w kolejnych warstwach ukrytych. Pozwala to na uchwycenie dodatkowych zależności (stosując na przykład dodatkową atencję w przestrzeni cech), które umożliwiają badanie skanów 3D	Sensors 2020, 20, 5097. https://doi.org/10.3390/s20185097
Metody wyjaśnialności LRP oraz Grad-CAM	LRP - Wyjście z ostatniej warstwy to finalna kontrybucja neuronów z warstw ukrytych, która ulega propagacji wstecz, aby stworzyć mapę kontrybucji neuronów ze wszystkich warstw (suma kontrybucji neuronów z warstw ma ograniczenie - każda warstwa ma kontrybucję o wartości stałej, cała kontrybucja to liczba warstw razy kontrybucja dla warstwy). Grad-CAM - Skupia się na obliczaniu gradientu wchodzącego do ostatniej warstwy z podziałem na klasy, na podstawie którego tworzy mapę ważności poszczególnych wejść do sieci	Kim, Seonggyeom, and Dong-Kyu Chae. "Exmeshcnn: An explainable convolutional neural network architecture for 3d shape analysis." Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining. 2022. https://doi.org/10.1145/3534678.3539463
Wykorzystanie transformerów do segmentacji (i pośrednio do rozpoznawania cech) wolumetrycznych danych medycznych - użycie transformerów do wyodrębniania reprezentacji	Przetwarzanie za pomocą transformerów niskopoziomowych cech wyodrębnionych przez konwolucje pozwala na stworzenie mapy relacji pomiędzy grupami lokalnych cech, umożliwiając modelowania globalno-lokalnych zależności w dynamiczny sposób dając nowe możliwości interpretacji cech	Zhou, Hong-Yu, et al. "nnformer: Volumetric medical image segmentation via a 3d transformer." IEEE transactions on image processing 32 (2023): 4036-4045.