Obecnie żyjemy w epoce *big data*, co oznacza że nie tylko należy przygotować odpowiednie narzędzia informatyczne do analizy dużych wolumenów danych, ale także musimy być gotowi na przetwarzania danych o charakterze strumieniowej, czyli takich które napływają w trakcie eksploatacji modelu analitycznego. Wiąże się to przede wszystkim z potrzebą rozwiązania nowych problemów, takich jak zaproponowanie metod reakcji na zmiany w modelach danych, czy też duże dysproporcje pomiędzy obiektami z poszczególnych klas (niezbalansowanie danych). Projekt *Algorytmy klasyfikacji niezbalansowanych strumieni danych* stanowi próbę połączenia tzn. klasyfikacji danych strumieniowych oraz analizy danych niezbalansowanych. Jedynie nieliczni autorzy dostrzegają, że problem niezbalansowania danych strumieniowych ma zupełnie odmienny charakter, niż w przypadku gdy dysponujemy *a priori* pełnym zbiorem uczącym. Wiąże się to, przede wszystkim z faktem braku apriorycznej informacji o rozkładach klas oraz, szczególnie w początkowym okresie klasyfikacji strumienia danych, możliwości niewłaściwej identyfikacji klasy mniejszościowej. Utrudnieniem może być także występowanie zjawiska dryftu modelu (*concept drift*), mogącego wpływać np. na zmianę częstości występowania obiektów z rozważanych klas, czego przykładem jest diagnostyka techniczna, gdzie prawdopodobieństwo wystąpienia awarii wzrasta z czasem eksploatacji urządzenia, czy też wynika ze zjawisk związanych np. ze zmęczeniem materiałów, z których wykonano urządzenie. Takie zjawisko obserwować można także w zadaniach np. związanych z analizą mediów społecznościowych, m.in. popularnością tematów poruszanych na *Twitterze*. W ramach proponowanego projektu proponuje się następującą hipotezę badawczą:

Metody uczenia klasyfikatorów na podstawie danych strumieniowych, które są w stanie uwzględnić niezbalansowany rozkład obiektów w klasach, prowadzą do otrzymania modeli o wyższej jakości, niż algorytmy nie uwzględniające tej charakterystyki.

W ramach projektu zostaną zrealizowane następujące zadania badawcze

- *Metody klasyfikacji niezbalansowanych, stacjonarnych strumieni danych.* Główny nacisk położony będzie na opracowanie metod związanych z oceną stopnia niezbalansowania, co zostanie uwzględnione w trakcie opracowania nowych metod nad- i podpóbkowania, w tym zostanie rozważony problem generowania sztucznych obiektów klasy mniejszościowej, tak aby zmniejszały one trudności zadania klasyfikacji. Opracowane zostaną także wbudowane mechanizmy klasyfikacji niezbalansowanych strumieni danych, bazujące na koncepcji zespołów klasyfikatorów oraz wbudowane w pojedynczy model klasyfikatora. Wykorzystany zostanie także paradygmat aktywnego uczenia, który umożliwi tańsze (z punktu widzenia kosztu etykietyzacji) uczenie klasyfikatorów.
- *Metody klasyfikacji niezbalansowanych, niestacjonarnych strumieni danych.* Zaproponowane zostaną modyfikacje metod opracowanych w zadaniu poprzednim, które uwzględnią zjawisko dryftu modelu. Największy nacisk położony zostanie na zaproponowanie mechanizmów adaptacji modeli do zmian. W tym celu zostaną zaproponowane metody zapominania, uwzględniające niestacjonarność modelu oraz niezbalansowanych charakter danych. Podjęta także zostanie próba opracowania detektorów dryftu modelu dla takiego przypadku, w tym detektorów zmiany stopnia niezbalansowania.
- *Otwarta biblioteka programowa zawierająca opracowane metody.* Zostanie zaprojektowana i zaimplementowana otwarta biblioteka programistyczna w wybranym środowisku programistycznym, która posłuży do opracowania komputerowego systemu eksperymentowania. Kod implementacji algorytmów wchodzących w skład biblioteki programistycznej będzie otwarty i udostępniony w domenie publicznej.
- *Ocena opracowywanych algorytmów klasyfikacji.* Zostaną podjęte próby oceny własności opracowanych metod na drodze analitycznej. Jednakże, jako że ocena taka jest często ograniczona, bądź niemożliwa, ewaluacja opracowanych algorytmów zostanie dokonana głownie na drodze eksperymentu komputerowego. W ramach tego zadania zostanie podjęta także próba opracowania metryk jakości dedykowanych niezbalansowanym strumieniom danych.