SILESIAN UNIVERSITY OF TECHNOLOGY
Faculty of Automatic Control, Electronics
and Computer Science

Machine learning methods in support of multiomics signature identification for breast cancer patient subpopulations

Doctoral Dissertation

Joanna Tobiasz

Supervisor:
Professor Joanna Polańska, PhD, DSc

Co-supervisor:
Christos Hatzis, PhD

2023
Gliwice, POLAND

Abstract

Breast cancer is a highly heterogeneous disease with a diverse molecular portrait. The commonly used clinical classification of breast cancer subtypes relies on levels of several protein markers, while molecular classification is defined based on gene expression profiling. Both those divisions remain unchanged for years and do not sufficiently reflect the complex structure of the disease and observed clinical experience diversity. With the rapid progress in molecular biology, more accurate characterization of breast cancer subtypes may support the search for new therapeutic targets. This dissertation aims to develop machine learning-based methods to identify novel subpopulations of breast cancer patients and examine their unique molecular and clinical characteristics.

The tested combinations of feature engineering and clustering approaches and proposed comparison methods allow the division of patients based on their proteomic profiles. Six subpopulations were identified. They were evaluated demographically, clinically, and molecularly based on their protein and transcriptomic profiles. Suitable classic statistical analysis methods supported by effect size estimates and machine learning algorithms allowed for dealing with the comparison groups' different, sometimes insufficient, sizes.

Three of the six subpopulations derived from the proteomic profile were highly consistent with commonly used transcriptomic-based subtypes: basal, HER2-enriched, and luminal B. Nevertheless, the transcriptomic-based luminal A subtype was highly heterogeneous and divided into three subgroups in this work. Revealed subpopulations vary in survival experience and proteomic and transcriptomic profiles. Novel luminal subtypes are less differentiated at the transcriptomic level than in proteomic space. The sets of markers specific for certain subpopulations and the signature enabling distinction between all subtypes were obtained.

The obtained profiles of revealed subpopulations, especially the proteomic one, may potentially complement the used classifications of breast cancer and support the search for novel targeted therapies with the development of personalized medicine. Nevertheless, the independent validation of those findings is required to assess clinical applications further.