جامعة محمد الخامس بالرباط
Université Mohammed V de Rabat

École Nationale Supérieure d'Informatique et d'Analyse des Systèmes
Centre d'Études Doctorales en Sciences des Technologies de l'Information et de l'Ingénieur

AVIS DE SOUTENANCE DE THÈSE DE DOCTORAT

Madame Imane CHLIOUI

soutiendra publiquement sa thèse de Doctorat en Informatique

Le Mercredi 28 Juillet 2021 à 10H au Grand Amphi à l'ENSIAS

Intitulé de la thèse

Missing data techniques for breast cancer classification

UNIVERSITE MOHAMMED V - RABAT
ENSIAS

Devant le Jury composé de :

Président :

Pr. Hassan BERBIA, PES, ENSIAS, Université Mohammed V de Rabat

Directeur de thèse :

Pr. Ali IDRI, PES, ENSIAS, Université Mohammed V de Rabat

Rapporteurs :

Pr. Abdellah MASSAQ, PES, ENSA, Université Cadi Ayyad de Marrakech

Pr. Redouane EZZAHIR, PES, ENSA, Université Ibn Zohr d'Agadir

Pr. Atman JBARI, PH, ENSAM, Université Mohammed V de Rabat

Examinateur :

Pr. Tamou NASSER, PES, ENSIAS, Université Mohammed V de Rabat

Invité :

Pr. Ibtissam ABNANE, PESA, ENSIAS, Université Mohammed V de Rabat

MISSING DATA TECHNIQUES FOR BREAST CANCER CLASSIFICATION

Abstract: Nowadays, the emergence of Data mining (DM) has helped to assist doctors in several subfields of medicine such as cardiology, endocrinology, neurology and oncology. Specifically, DM techniques have been actively used to assist doctors in the process of BC diagnosis. Usually, classification models for BC diagnosis are built using medical data collected from hospitals. However, medical data often contain Missing Data (MD) which reduce the number of available cases for analysis. Missing Data MD a substantial problem that faces researcher when applying Data Mining techniques, notably when it comes to medical datasets. In the light of the different proposed MD techniques, it is important to determine which one is the more suitable to apply. The development of MD techniques was therefore frequently motivated by the requirements for dealing with these characteristics and filling the gaps identified.

This thesis aims to: (1) perform a systematic mapping study on the use of data mining techniques in breast cancer to analyze and synthesize evidence about this field of research. (2) perform a second mapping study on the use of preprocessing techniques in BC to identify the gaps in this area. (3) evaluate and compare existing MD techniques for BC classification: deletion, mean, Expectation-maxmization imputation, k-nearest neighbors' imputation, and support vector regression imputation with five classifiers: decision tree, random forest, support vector machine, case based reasoning, and multilayer perceptron. (4) propose a new approach to deal with MD for BC classification based on ensembles (homogeneous and heterogeneous) using multiple base models.

The finding of this thesis proved that there is no better missing data technique for every classifier and every dataset. In fact, every single technique has its strength and weaknesses. Thus, the use of ensemble imputation techniques yield to the best classification results for BC.

Keywords: Knowledge Data Discovery, Data mining, classification techniques, missing data, breast cancer, ensembles.