JOHANNES KEPLER
UNIVERSITY LINZ

Philipp Renz, MSc
Institute for
Machine Learning

P +43 732 2468 4481
F +43 732 2468 4539
ramsauer@ml.jku.at

Office:
Birgit Hauer
Ext.4520
birgit@ml.jku.at

Linz, September 6th, 2019

Opponent Review "Imputation Of Missing Values In Clinical Data" by Micha Birklbauer

Reading "Imputation Of Missing Values In Clinical Data" by Micha Birklbauer was very interesting.
The thesis is concerned with the problem that predictive models often can only be trained on datasets without missing values, which is not given in all applications.
To mitigate this problem one often uses data imputation techniques. In this work the quality of some of these techniques are compared, as measured by the performance of a downstream prediction task.

The thesis treats a relevant topic and the experimental execution is rigorous.

There are some flaws in the presentation of the used imputation methods.
Firstly, two of the methods are introduced by copying relevant sections from cited publications, which is not usual. This amounts to about three pages in total. Some variables (R, φ) used in the explanation of "Fully conditional specification" are not defined, which hinders the understanding of Algorithm 2. "MICE" and "MissForest" are not cited the first time they are mentioned in the main text. In the first sentence of section 4 citations are missing. The figure of the RBM on page 17 is not cited, but can be found on the internet (and was most likely not made by the author). In the explanation of RBMs it is also stated that "hidden units are conditionally independent" which makes no sense without also stating the condition.

The result analysis using the t-test could easily overlook differences, as for example when doing a test using BACC-values one also compares values from models optimized for ACC or AUC. As one can see from the results the BACC values depend strongly on what the models have been optimized for. Also the comparison across different amounts of missingness can overlook positive results as the performance metrics also depend on this variable.

JOHANNES KEPLER
UNIVERSITY LINZ
Altenberger Str. 69
4040 Linz, Austria
www.jku.at
DVR 0093696

JOHANNES KEPLER
UNIVERSITY LINZ

Another point that it missing in my view is information about the sizes of the positive/negative class of the prediction task, which is useful to interpret the performance metrics.

Although I think that there are a few minor problems, they are not able to outweigh the positive aspects of this work. Therefore I recommend acceptance of this Bachelor's thesis.

Philipp Renz, MSc

Renz P.

Johannes Kepler Universität Linz
Institut für Machine Learning
Altenberger Straße 69
A-4040 Linz, Austria

JOHANNES KEPLER
UNIVERSITY LINZ
Altenberger Str. 69
4040 Linz, Austria
www.jku.at
DVR 0093696