A MATLAB app to assess, compare and validate new methods against their benchmarks

Shaul Ajo'[1, 2*], Davide Piaggio[1[0000-0001-5408-9360]], Mahir Taher[1], Franco Marinozzi[2], Fabiano Bini[2] and Leandro Pecchia[1** [0000-0002-7900-5415]]

[1] Applied Biomedical Signal Processing and Intelligent eHealth Lab, School of Engineering, University of Warwick, Coventry CV4 7AL, UK
[2] Department of Mechanical and Aerospace Engineering, Sapienza University of Rome, Rome 00184, Italy

*Email Address: shaul.ajo@gmail.com
** Email Address: l.pecchia@warwick.ac.uk

Abstract. Emerging technologies for physiological signals and data collection enable the monitoring of patient health and well-being in real-life settings. This requires novel methods and tools to compare the validity of this kind of information with that acquired in controlled environments using more costly and sophisticated technologies. In this paper, we describe a method and a MATLAB tool that relies on a standard sequence of statistical tests to compare features obtained using novel techniques with those acquired by means of benchmark procedures. After introducing the key steps of the proposed statistical analysis method, this paper describes its implementation in a MATLAB app, developed to support researchers in testing the extent to which a set of features, captured with a new methodology, can be considered a valid surrogate of that acquired employing gold standard techniques. An example of the application of the tool is provided in order to validate the method and illustrate the graphical user interface (GUI). The app development in MATLAB aims to improve its accessibility, foster its rapid adoption among the scientific community and its scalability into wider MATLAB tools.

Keywords: Surrogate Features, Statistical Analysis, Benchmarking.

1 Introduction

Biomedical engineers, among other experts of Science, Technology, Engineering and Mathematics (STEM), constantly attempt to develop novel approaches and techniques to acquire and/or analyze different kinds of data [1-3]. As an example, the unprecedented amount of data generated by the spread use of Internet of Things (IoT) is encouraging the deployment of alternative methods and tools. Normally these alternative methods are less time-consuming, less expensive, non-invasive and can work automatically, with minimum manual intervention of an expert.

This is a positive trend, which can, however, lead to the improper use and application of methodologies that are not yet correctly validated and whose results are consequently unreliable [4]. For instance, as widely discussed by Pecchia et al. [4], several studies employed only correlations to prove that ultra-short term heart rate variability (HRV) features behaved as short term ones, concluding that the former were good substitutes of the latter if significantly correlated among each other [5]. Nonetheless, this result is controversial, because “a correlate does not make a surrogate” [6].

Specifically, in [4], Pecchia et al. presented a protocol to evaluate whether ultra-short terms HRV features can be considered as a valid replacement for short term ones. In light of this, the authors of this paper propose the generalization of such algorithm and its implementation into a MATLAB app to define a precise procedure for statistical comparisons of two data vectors, respectively obtained using a new technique and the benchmark method. This MATLAB app can be used to assess, compare and validate new methods against their benchmarks.

2 Methods

2.1 Statistical analysis

Basing on Pecchia et al. [4], the authors generalized the proposed method as the sequence of four steps:

1. Normality test
2. Correlation analysis
3. Bland-Altman plots
4. Statistical hypothesis test

Following the above-mentioned steps, the authors have developed an app using MATLAB R2019b.

Normality test. In the world of inferential statistics, knowing the probability density function (PDF) underlying the data is essential for performing the correct analysis and avoiding erroneous and misleading outcomes. In particular, this is true with the normality of data: in fact, many statistical methods can be applied if and only if the data follow a Gaussian distribution. In all other cases, alternative methodologies should be taken into consideration.

For this reason, the first step of any good statistical analysis should be testing normality. In fact, many normality tests have been developed in the past, such as Shapiro-Wilk [7], Shapiro-Francia [8], Lilliefors [9], and Anderson-Darling [10]. For our case we selected the Shapiro test because it has more power than the others [11, 12] and is more appropriate for small sample sizes [13].

Correlation analysis. Investigating how strongly two variables or features are related, is the first step towards the identification of a good alternative method [4]. Consequently, our protocol uses Pearson’s r or Spearman’s rho coefficients, for normal and non-normal data respectively, and the related p-values to assess the correlation [14].

In particular, the authors consider a p-value less than the significance level alpha (normally 0.05 or 0.01) significant and correlation coefficients with an absolute value above 0.7 as “high” [15]. However, since “a correlate does not make a surrogate”, as anticipated in the introduction, the further following steps are necessary.

Bland-Altman plots. Since their first introduction in literature [1, 16], Bland-Altman plots have been widely used in the clinical and medical fields for visually comparing two methods that measure the same phenomenon supporting the correlation analysis. Also in this case, the normality or non-normality of data affects the procedure to obtain the plot [16]. Moreover, in literature there is the erroneous tendency to omit the confidence intervals in the graph [17], which are suggested by Bland and Altman [1].

The Bland-Altman graph represents the differences among the variables acquired with the two different methods under test on the y-axis versus the averages of the same variables on the x-axis. The graph also plots three lines: the bias and the two 95 percent limits of agreement along with their confidence intervals. Such lines are important to understand whether there is a systematic error in one of the methodologies and the level of agreement between them. If the limits of agreement do not exceed the maximum allowed difference between methods, which depends on the specific context of application, the two methods are in agreement and may be used interchangeably (i.e., the narrower the limits of agreement, the closer the methodologies).

Statistical hypothesis test. Lastly, the similarity of two methods should be further confirmed by statistical hypothesis tests. The latter depend again on the type of distribution underlying the data: in our case, a t-test was applied for all the normally-distributed data and the non-parametric Wilcoxon signed-rank test was selected for all the other cases. In order to be more conservative [18], the authors suggest considering the results of the Wilcoxon signed-rank test in any case.

2.2 Validation

In order to validate the algorithm and present the graphical user interface (GUI) of the MATLAB tool the authors will show an example of its application to assess whether in rest condition ultra-short term heart rate variability (HRV) analysis (performed on excerpts shorter than 5 min) can be considered a good surrogate of the short term one (performed on 5 min excerpts), which is regarded as the reference method in this case [19].

HRV is currently one of the most investigated methods for assessing mental stress [19] and is effective in early detection of cardiovascular diseases worsening [20]. While short term HRV analysis has been considerably investigated, less work has been done on ultra-short term HRV analysis. The demands of latter for monitoring individual’s well-being status is increasing, due to the growing popularity of wearable sensors, smart phones and smart watches [4, 21]. In e-health monitoring, in fact, the conventional 5 min recordings might be unsuitable, due to real-time requirements.

3 Results

The algorithm implemented by the tool can be summarized by the block diagram shown in Fig. 1.

The app comprises of six tabs, which are illustrated below as part of the validation of the tool. The validation has been conducted on data collected for a previous study, as describe in [22]. In particular, the app has been tested to verify if the HRV feature meanRR, mean of the RR intervals, extracted via an ultra-short term HRV analysis (i.e. 3 min), which will be named meanRR_3min can substitute the correspondent feature evaluated on 5 min excerpts (benchmark), which will be called meanRR_5min.

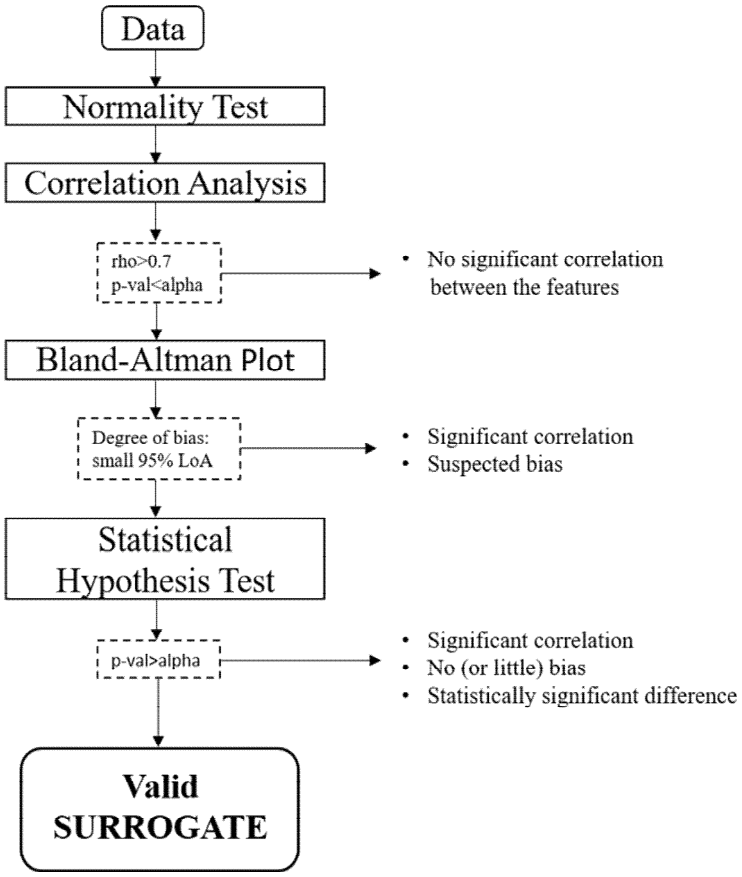

Fig. 1. The algorithm implemented by the MATLAB tool to assess if the features extracted by the novel method can be considered a good surrogate for the features obtained applying the reference technique (benchmark). rho: correlation coefficient; p-val: p-value associated with the correlation analysis; alpha: the significance level chosen by the user; LoA: line of agreement in Bland-Altman plot.

3.1 First tab: overview

The first tab (Fig.2) briefly describes the purpose of the app, reports the block diagram described in Fig. 1 and indicates the steps the users need to follow in order to run the app. Lastly, it allows the user to download and save a template of a spreadsheet.

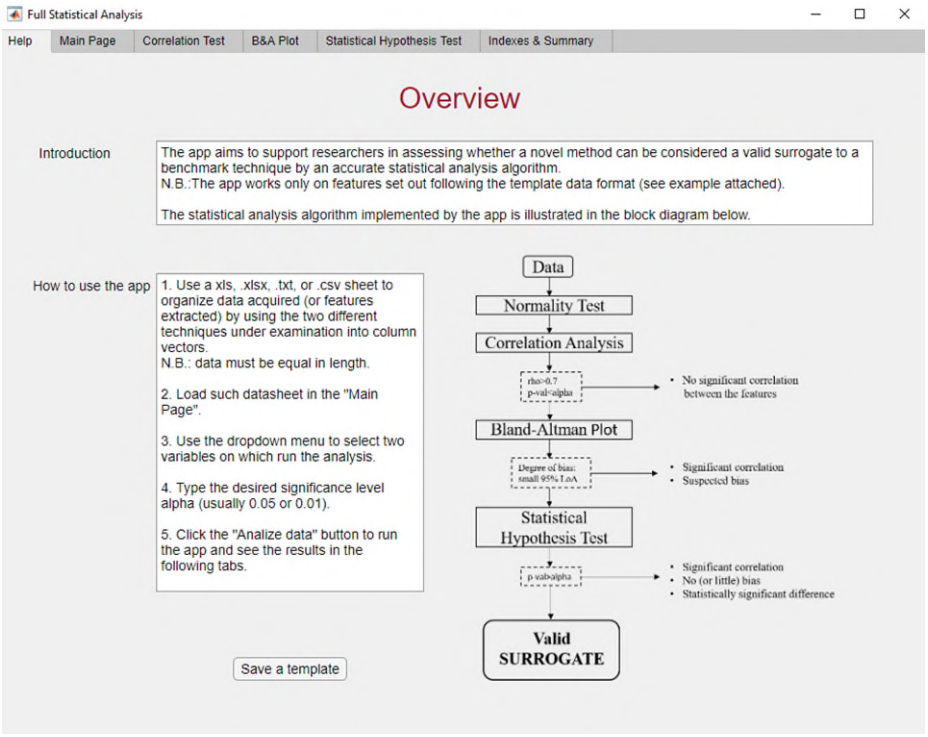

Fig. 2. First tab. Overview and instructions on how to run the app.

3.2 Second tab: data uploading and normality test

The second tab (Fig. 3) asks the user to upload a spreadsheet (.xls, .xlsx, .txt or .csv) containing the data obtained by using different methods. Moreover, a dropdown menu is available to select the two variables that will be compared following the pipeline shown in Fig 1. Furthermore, it is possible to choose the desired significance level alpha (such as 0.05 or 0.01) which will be used throughout the analysis. Once the run button (i.e. analyze data button) is clicked, a message box appears and indicates if the data are normally distributed. The normality test is based on the MATLAB file by BenSaida [23], available for free on the internet. For this specific validation, a significance level α equal to 0.05 was chosen. The two variables proved to be both normally distributed.

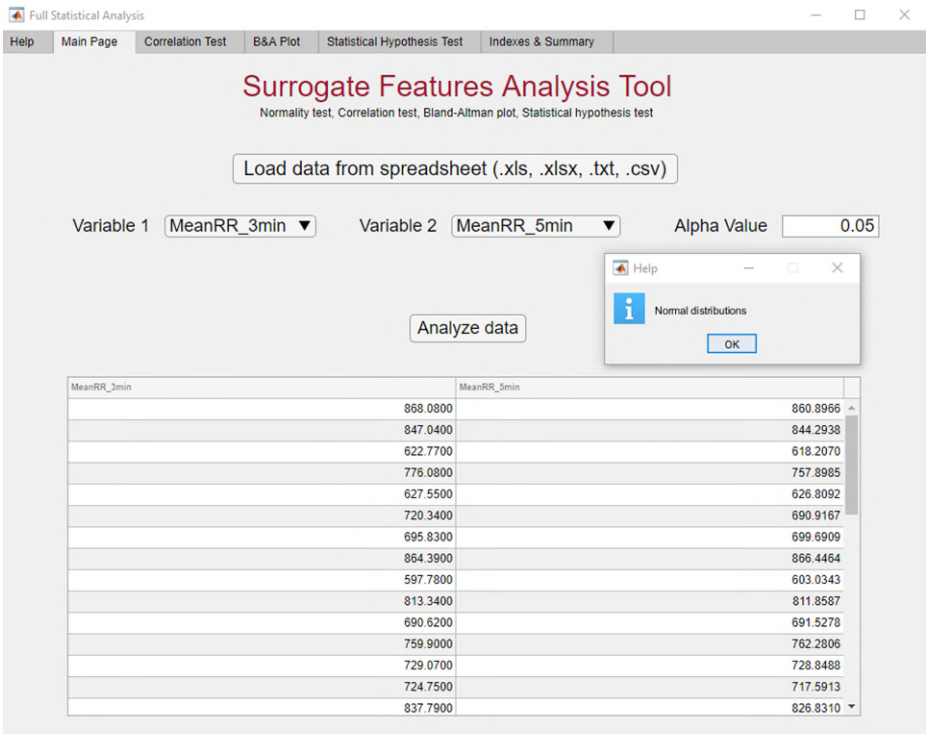

Fig. 3. Second tab. Main page of the tool.

3.3 Third tab: results of the correlation analysis

The third tab (Fig. 4) shows the results of the correlation analysis (correlation coefficient and p-value) and a textbox, which states whether the two input variables are significantly correlated (p-value < alpha AND |rho| > 0.7) or not. The software performs the Pearson's correlation analysis or the Spearman's correlation analysis depending on the PDFs underlying the data.

According to the Pearson's correlation analysis meanRR_3min and meanRR_5min are significantly associated with a correlation coefficient r equal to 0.9922 and a p-value less than 0.01.

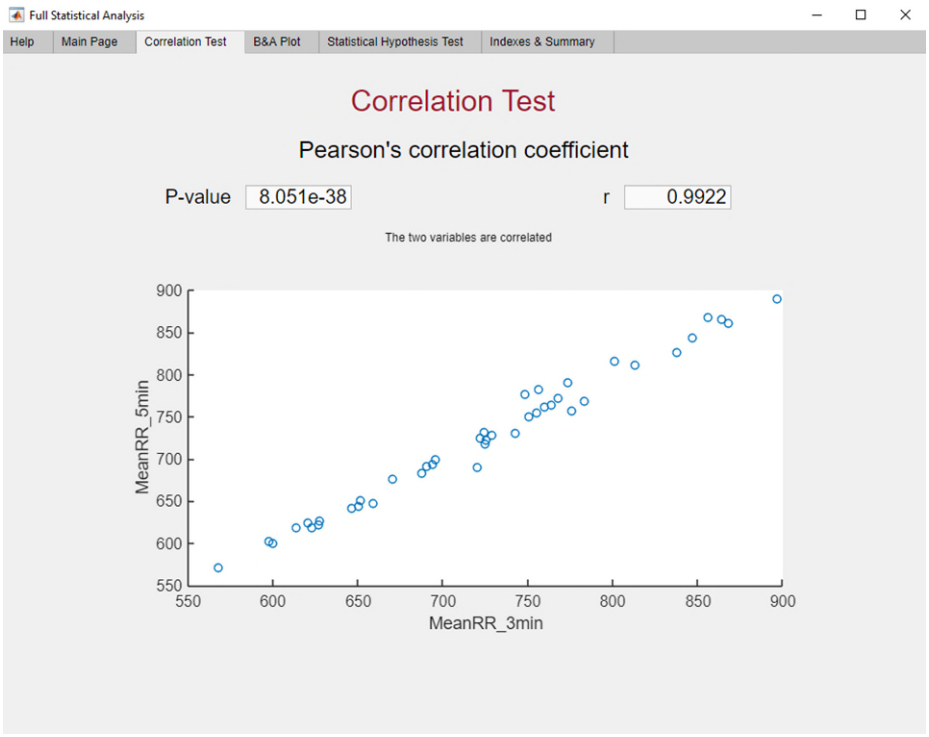

Fig. 4. Third tab. Results of the correlation analysis.

3.4 Fourth tab: bias, limits of agreement and Bland-Altman plots

The fourth tab (Fig. 5) represents the Bland-Altman plots, inclusive of the limits of agreement, the bias and the confidence intervals, to support the results from the correlation analysis by a visual inspection. The first graph plots the differences among the variables acquired with the two different methods under test on the y-axis versus the averages of the same variables on the x-axis, while the second one expresses the differences as percentages of the observations represented on the x-axis. The authors want to underline that the limits of agreement and the estimation of their confidence intervals when the data are not normally distributed can be considered only a first approximation since they are evaluated under the hypothesis of normality. Moreover, the two graphs can be saved or copied as images (.png, .jpg, .tif, .pdf), or copied as a vector graphic.
By looking at the Bland-Altman plot it is possible to identify two outliers, a very small bias and 95 percent limits of agreement very close to each other. Overall, no specific trend stands out and therefore the two features can be considered in agreement.

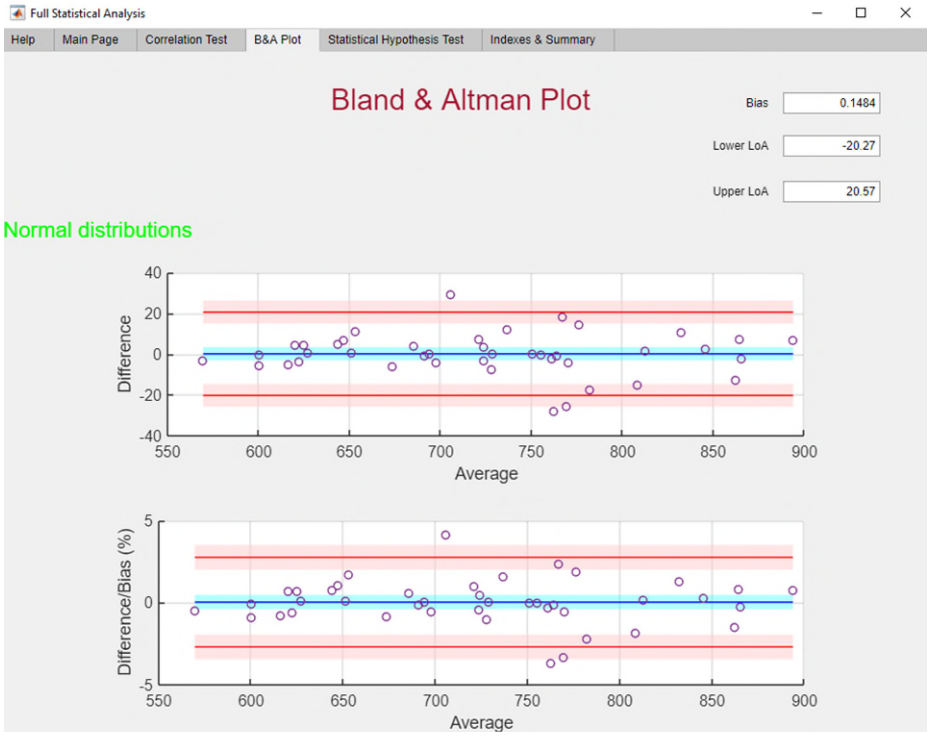

Fig. 5. Fourth tab. Bland & Altman plots.

3.5 Fifth tab: results of the statistical hypothesis test

The fifth tab (Fig. 6) provides the result of the statistical hypothesis test and a text-box, which states if the two input variables are significantly different or not.
Based on the Student's t-test the two features under examination are not significantly different. However, the authors suggest that using Wilcoxon signed-rank test even in the case of normality can lead to more conservative results.

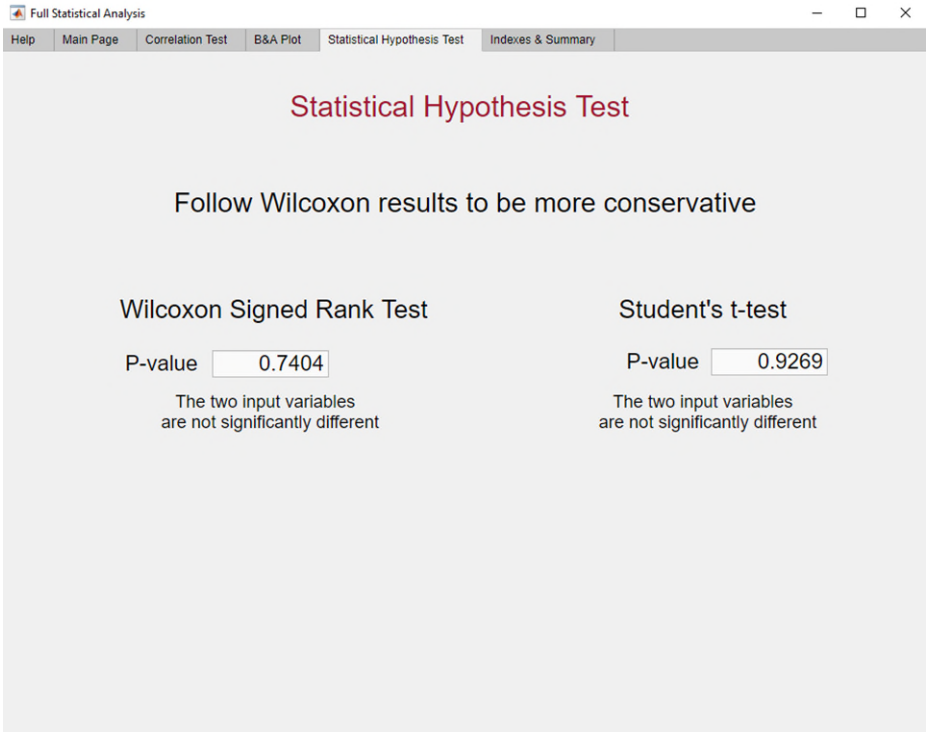

Fig. 6. Fifth tab. Results of the statistical hypothesis test.

3.6 Sixth tab: summary and statistical indexes

The sixth and last tab (Fig. 7) of the tool summarizes the most relevant values of the previous steps and indicates median (MD), standard deviation (SD), 25^{th} and 75^{th} percentiles of the two variables under examination. Hence, the tab allows the user to draw conclusions and state if the two variables, obtained via two different methods, can be used interchangeably.

Finally, it is possible to conclude that the HRV feature mean RR is resilient: the mean RR calculated on 3 min excerpts is a good surrogate of the benchmark mean RR (i.e. calculated on 5 min excerpts). In light of this positive result, this investigation could be repeated to compare other HRV features and provide additional evidence that the ultra-short term HRV analysis based on 3 min recording is a valid replacement for the short term one.

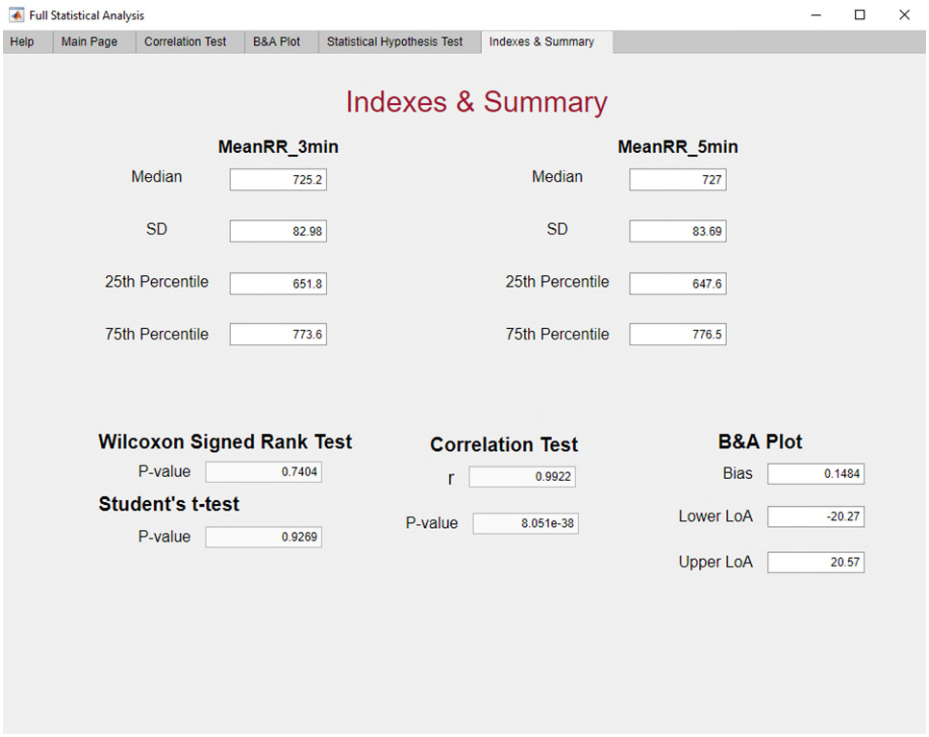

Fig. 7. Sixth tab. Summary and statistical indexes.

4 Conclusion

The current study proposed the implementation of the algorithm shown in Fig. 1 into a MATLAB app to quickly assess, compare and validate new methods for extracting features with their benchmarks. This tool will help researchers to use a standard sequence of statistical methods (i.e., not just statistical hypothesis or correlation tests) to explore whether new features can be considered robust substitutes of the current benchmarks. In fact, the tool guides the users into a correct statistical procedure, which includes normality test, correlation test, Bland-Altman plots, and statistical hypothesis test.

The tool offers a smaller range of statistical methods compared to some statistical software available on the market, such SPSS, which can also deal with very complex datasets. Nevertheless, the tool is conceived as a simple means to lead the user throughout the steps of a rigorous and standardized statistical analysis, since many studies available in the literature employed partially or completely unreliable methods.

Future versions of the app will allow the user to select the preferred normality and hypothesis tests and will include an effect size tab (e.g., Cohen's d) in order to complement the results of the statistical hypothesis test and provide further evidence of the agreement between two variables.

5 Conflict of Interest

The authors declare that they have no conflict of interest.

References

1. Bland, JM., Altman, DG.: Measuring agreement in method comparison studies. Statistical Methods in Medical Research 8(2), 135-160 (1999).
2. Serrao, M., Chini, G., Bergantino, M., Sarnari, D., Casali, C., Conte, C., Ranavolo, A., Marcotulli, C., Rinaldi, M., Coppola, G., Bini, F., Pierelli, F., Marinozzi, F.: Identification of specific gait patterns in patients with Cerebellar Ataxia, Spastic Paraplegia and Parkinson's Disease: a non-hierarchical cluster analysis. Hum. Movement Sci. 57, 267-279 (2018).
3. Marinozzi, F., Bini, F., Marinozzi, A., Zuppante, F., De Paolis, A., Pecci, R., Bedini. R.: Technique for bone volume measurement from human femur head samples by classification of micro-CT image histograms. Ann. I. Super. Sanità 49(3), 300-305 (2013).
4. Pecchia, L., Castaldo, R., Montesinos, L., Melillo, P.: Are ultra-short heart rate variability features good surrogates of short-term ones? State-of-the-art review and recommendations. Healthcare Technology Letters 5(3), 94-100 (2018).
5. Brisinda, D., Venuti, A., Cataldi, C., Efremov, K., Intorno, E., Fenici, R.: Real-time imaging of stress-induced cardiac autonomic adaptation during realistic force-on-force police scenarios. Journal of Police and Criminal Psychology 30(2), 71-86 (2015).
6. Fleming TR., DeMets, DL.: Surrogate end points in clinical trials: are we being misled? Annals of Internal Medicine 125(7), 605-613 (1996).
7. Shapiro, SS., Wilk, MB.: An analysis of variance test for normality (complete samples). Biometrika 52(3-4), 591-611 (1965).
8. Shapiro, SS., Francia, RS.: An approximate analysis of variance test for normality. Journal of the American Statistical Association 67(337), 215-216 (1972).
9. Lilliefors HW.: On the Kolmogorov-Smirnov test for normality with mean and variance unknown. Journal of the American Statistical Association 62(318), 399-402 (1967).
10. Anderson TW., Darling, DA.: Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. The Annals of Mathematical Statistics 23(2), 193-212 (1952).
11. Thode, HC.: Testing for normality. Marcel Dekker, New York (2002).
12. Razali, NM., Yap BW.: Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. Journal of Statistical Modeling and Analytics 2(1), 21-33 (2011).
13. Ahad, NA., Yin, TS., Othman, AR., Yaacob, CR.: Sensitivity of Normality Tests to Non-normal Data. Sains Malaysiana 40(6), 637-641 (2011).
14. Mukaka, MM.: Statistics corner: a guide to appropriate use of correlation coefficient in medical research. Malawi Medical Journal 24(3), 69-71 (2012).
15. Hinkle, DE., Wiersma W., Jurs SG.: Applied statistics for the behavioural sciences. Houghton Mifflin, Boston, Massachusetts (2003).
16. Bland, JM., Altman, DG.: Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1(8476), 307-310 (1986).

17. Sedgwick, P.: Limits of agreement (Bland-Altman method). *BMJ* 346(1630), (2013).
18. Nahm, FS.: Nonparametric statistical tests for the continuous data: the basic concept and the practical use. Korean Journal of Anesthesiology 69(1), 8-14 (2016).
19. Castaldo, R., Melillo, P., Bracale, U., Caserta, M., Triassi, M., Pecchia, L.: Acute mental stress assessment via short term HRV analysis in healthy adults: A systematic review with meta-analysis. Biomedical Signal Processing and Control 18, 370-377 (2015).
20. Pecchia, L., Melillo, P., Sansone, M., Bracale, M.: Discrimination power of short-term heart rate variability measures for CHF assessment. IEEE Transactions on Information Technology in Biomedicine 15(1), 46-46 (2011).
21. Athavale, Y., Krishnan, S.: Biosignal monitoring using wearables: observations and opportunities. Biomedical Signal Processing and Control 38, 22-33 (2017).
22. Melillo, P., Bracale, M., Pecchia, L.: Nonlinear heart rate variability features for real-life stress detection. Case study: students under stress due to university examination. BioMedical Engineering OnLine 10(96) (2011).
23. MathWorks, https://uk.mathworks.com/matlabcentral/fileexchange/13964-shapiro-wilk-and-shapiro-francia-normality-tests, last accessed 2019/12/28.