E-ISSN: 3063-0304, P-ISSN: 3063-1092
Researcher Academy Innovation Data Analysis (RAIDA) Vol 1, No. 2, September 2024, Page 214-226
https://doi.org/10.69725/raida.v1i3.168

Enhancing Nonparametric Tests: Insights for Computational Intelligence and Data Mining

Shahid Yousufi [a], **Fermle Erdely S** [b]

[a.] *School of Computer Science and Electronic Engineering, University of Essex, Wivenhoe Park, Colchester, United Kingdom*
[b.] *Interactive Coventry Limited, Coventry University Technology Park, Puma Way, Coventry, CV1, United Kingdom*

ARTICLE INFO

Article history:
Received 7 April 2024
Revised 6 December 2024
Accepted 10 December 2024

Author's correspondence;
Shahid

Keywords:
Nonparametric tests, computational intelligence, data mining, post hoc analysis, statistical reliability

ABSTRACT

Objective: With the aim of improving monitoring reliability and interpretability of CI and DM experimental statistical tests, we evaluate the performance of cutting-edge nonparametric tests and post hoc procedures.
Methods: A Friedman Aligned Ranks test, Quade test, and multiple post hoc corrections Bonferroni-Dunn and Holm were used to comparative analyze data. These approaches were employed to algorithm performance metrics with varied datasets to evaluate their capability to detect meaningful differences and control Type I errors.
Results: Advanced nonparametric methods consistently outperformed traditional parametric tests, offering robust results in heterogeneous datasets. The Quade test was the most powerful and stable, and the post hoc procedures greatly increased the power of the pairwise comparisons.
Novelty: We evaluate advanced nonparametric methods in CI and DM experiments: the Friedman Aligned Ranks test, the Quade test, and post hoc procedures (Bonferroni-Dunn and Holm). These methods represent a departure from traditional parametric tests that depend on assumptions of normality and homogeneity of variance, allowing for more flexible and robust approaches to analyses of complex, heterogeneous datasets. By comparing the strength and efficacy of these methods, the research also delivers common guidelines for their use; as well as demonstrating their utility in realistic situations characterized by non-standard and dispersed data.
Implications for Research: The findings have far-reaching theoretical and pragmatic implications for scholars in CI and DM. On a theoretical level, this work undermines the common bias towards parametric techniques, providing an increasingly robust framework for comparative analysis in experimental research. This work improves understanding of the adaptation of statistical tests to fit the complexities of real-world data by highlighting the advantages of advanced nonparametric methods, specifically the Quade test and post hoc corrections. Practical implications The results give owners of data summaries actionable recommendations, which will assist researchers in the selection of statistical methods that are tuned to the nature of their datasets, resulting in improved reliability and interpretability of future evaluations of algorithms. Thus, this endeavor will promote more powerful and statistically appropriate methods in CI and DM studies, leading to more confident and valid claims surrounding algorithmic performance.

1. Introduction

Despite the fact that the research in the fields of computational intelligence (CI) and data mining (DM) are advancing quickly as more real-world data are being produced and released, algorithms are becoming more and deeper diverse and complex. With these developments come the need for solid and flexible evaluation paradigms to determine the effectiveness of algorithms across a wide range of conditions. Indeed, recent studies emphasize the importance of statistical frameworks going beyond standard approaches to obtain insights on algorithm performance on different datasets (Diez-Olivan et al., 2019; Niso et al., 2022; Shmueli et al., 2016; Sivarajah et al., 2017). For instance, given that CI and DM applications extend across healthcare, finance, and other critical sectors, the reliability of algorithmic outcomes is an imperative (Chen et al., 2024; Garcia-Perez et al., 2023; Rasheed et al., 2022). Yet, inferential statistical methods, particularly nonparametric methods have become tools to overcome such limitations of parametric tests (Carrasco et al., 2020; Jimenez-Mesa et al., 2023; Zhang et al., 2018). Such nonparametric approaches are currently at the core

E-ISSN: 3063-0304, P-ISSN: 3063-1092
Researcher Academy Innovation Data Analysis (RAIDA) Vol 1, No. 2, September 2024, Page 214-226
https://10.69725/raida.v1i3.168

of these CI and DM related work to draw generalizable conclusions.

Statistical methods are commonly used in CI and DM, which can be challenged by natural data and algorithm variability. Parametric methods like ANOVA are powerful, but demanding strict normality and homoscedasticity assumptions often violated in CI and DM experiments (Niankara, 2024; Sanchis-Segura & Wilcox, 2024; Yu et al., 2022). This has resulted in broadly and widely misinterpretation, or underestimation of algorithm performance, particularly in complex, real-world applications (Berger et al., 2024; Razavi et al., 2021). The limits of parametric methods, which are being regarded as the less robust option now that nonparametric methods that do not make such onerous assumptions, have been gaining traction. Yet, when it comes to selecting and applying approaches, no clear guidelines exist; as a result, the adoption of the methods has been slow and statistical practices have been inconsistent across the literature. Resolving this issue is essential for facilitating substantial progress, as it ensures that the evaluation following the algorithm is both statistically sound (i.e., the inference drawn is justified) and practically relevant (i.e., the output is useful). This addresses the validity of the algorithm's predictions, as this will mean, the algorithm is usable as in real world scenerios, and provide useful insights for practice. With improved assessment procedures, researchers and practitioners can boost the algorithm's performance to generate more precise and meaningful results across multiple applications (A. S. Albahri et al., 2023; Ali et al., 2017; Ezugwu et al., 2022).

Statistical testing–theoretically–in CI and DM revolves around testing whether performance difference between algorithms can be representative of one is better than stringing it along the need to PVC. Parametric methods take assumptions on the data distribution whereas nonparametric ones like the Friedman test and its extensions offer a robust solution Carrasco et al. (2020), LaTorre et al. (2021), Veček et al. (2017), since they can handle non-normal, heteroscedastic data. An example of that is the Friedman test that compares rank-based differences for multiple algorithms on multiple datasets, providing a more flexible comparison framework. Further extensions to this test, including post hoc pairwise comparisons, provide a finer approach to understanding algorithmic performance, especially in cases of multiple comparisons (Campelo & Wanner, 2020; Li et al., 2016; Olikh, 2024; Osaba et al., 2021). Such theoretical developments highlight the importance of advanced statistical methods that best meet the unique challenges presented by CI and DM research (O. S. Albahri et al., 2020; Madureira et al., 2021; Zhou et al., 2019).

We will discuss recent concerns around existing work, and the need for more general prior methods in the field of advanced nonparametric function estimation. Parametric tests have been employed in most of the studies which may have provided contradictory results because of the significance of violation of underlying assumptions resulting in under or over-representation of algorithmic performance (Chevalier et al., 2020; Deaton & Cartwright, 2018). In contrast, the non-parametric approaches have shown a lot of potential to mitigate these challenges. For instance, work by Pelletier et al. (2016), Yamasaki et al. (2024) emphasize the strength of nonparametric methods in evaluations across multiple datasets, and Cheng et al. (2022) stress their usefulness in limiting the family-wise error rate (FWER) for multiple comparisons. Yet the relative strengths and limitations of these approaches are poorly understood, especially in situations with highly variable datasets. It viewpoints nurturing the methodological to not only suggesting innovative approaches to CI and DM (Bi et al., 2022; Saad et al., 2023; Segundo-Marcos et al., 2023). This research advances the field by defining new nonparametric methods for multiple comparisons, generalizing prior methods, and providing empirical analyses of their efficiency (Parmezan et al., 2019; Roth et al., 2023). This study seeks to provide a more comprehensive framework to appropriately bridge the gap between theoretical understanding and practical utility in making CI and DM more informative, reliable, and generalizable through statistical assays by identifying proper statistical metrics that can be used to assess the stated parameters.

We aim for this study to advance the methodological rigor of statistical analyses in CI and DM by presenting novel, more robust, nonparametric techniques and extending existing methods. This paper thus serves to (1) uncover the fundamental bottlenecks of current parametric and nonparametric approaches, (2) introduce flexible nonparametric tests for a class of multiple comparisons and (3) demonstrate their robustness and efficiency in various experimental scenarios. We aspire to provide fast, statistical tools that allow researchers to assess the performance of algorithms in a reliable and interpretable way so that work in the field of CI and DM methodology is appropriately built on state-of-the-art algorithm performance.

2. Theoretical framework and development

2.1 Statistical analysis in the field of computational intelligence and data mining

Statistical analysis is a key part of any experimental work in computational intelligence (CI) and data mining

E-ISSN: 3063-0304, P-ISSN: 3063-1092
Researcher Academy Innovation Data Analysis (RAIDA) Vol 1, No. 2, September 2024, Page 214-226
https://10.69725/raida.v1i3.168

(DM), to assess algorithmic performance with robust measures. Since inferential statistics allows researchers to generalize results from sample datasets, they are also particularly important when assessing the efficacy of an algorithm in different populations. Traditional parametric tests such as ANOVA have been favored for their high statistical power to identify subtle differences. But these tests are based on strong assumptions (normality, homoscedasticity and independence). These assumptions are rarely satisfied in CI and DM experiments, where data sets can vary significantly in distribution and complexity. Nonparametric testing is a useful alternative as it lowers these assumptions, thus broadening their application to real-world situations. These include non-parametric tests, such as the Friedman test and Wilcoxon signed-rank test, which are especially apt for assessing results where performance metrics can lack a normal distribution. Previous work such as (Demšar, 2006) shows that nonparametric methods can produce more reliable results in CI and DM-related experiments with heterogeneous datasets. These statistical tools are critical for establishing the understanding of algorithmic behaviour across individual problem domains

2.2 Non-parametric tests and their significance

Due to their flexibility and robustness, non-parametric tests have become a primary alternative in CI and DM studies. They are non-parametric tests and thus do not need the underlying data distribution assumptions that make them appropriate for multiple problem analysis. A commonly employed nonparametric test is the Friedman test which performs rankings of the algorithms from their performance and tests to see if the differences between algorithms are significant. Some of the limitations of the Friedman test have been remedied through extending it to other variants, such as the Aligned Ranks and Quade tests, which are better able to detect performance differences and therefore increase mapping power. Hence, methods like Bonferroni-Dunn and Holm corrections, which are known as post hoc tests, are important methods for determining meaningful differences between pairs of groups while controlling the family-wise error rate. Researchers such as García et al. (2010) in CI experiments, to show they are able to cover a variety of data and function at different levels of algorithm performance. The increasing interest in nonparametric techniques illustrates the importance of developing advanced statistical methodologies for CI and DM.

This theory is based on the premise that the performance of an algorithm is often dependent on the properties of the dataset, such as its size, complexity, and distribution of features. Tests such as Friedman's test and alternatives are nonparametric methods that can uncover these differences by not depending on distributional assumptions. García et al. (2010) demonstrated that nonparametric methods were able to capture performance differences across datasets and were useful for testing algorithms under a wide range of conditions. These tests operate on the premise of rank-based comparisons, offering a reliable mechanism for detecting significant differences between the groups, thus supporting the hypothesis.

H1: There are significant differences in algorithm performance across datasets with varying characteristics.

2.3 Non-parametric procedures for multiple comparisons

There are some well-known requirements for CI and DM research. Nonparametric techniques allow for a broader generalization of classical statistical tools (e.g., the Multiple Sign test, Contrast Estimation based on medians). These approaches are appropriate when you have violations of parametric assumptions and are wellestablished alternatives for performance comparison. For example, the Multiple Sign test allows for quick assessment differences between a baseline method and competing algorithms, whereas median-based contrast estimation allows for more precise quantification of performance. The work of García et al. (2010), which provides evidence that these methods can produce accurate and interpretable results in complex experimental designs.

The observation that non-parametric tests are deemed to have no assumptions regarding normality or homogeneity of variance can be related to this assumption as well due to parametric tests having limitations in dealing with non-normal and heteroscedastic data distributions. Research such as the study presented in Demšar (2006) reveals nonparametric tests are best in experiments from within a variety of datasets. Nonparametric methods focus on ranks rather than raw data and are less sensitive to outliers and violations of distributional assumptions than parametric methods, making them the recommended choice for confidence interval (CI) and difference in means (DM) research (79).

H2: Nonparametric methods provide more reliable results than parametric tests in CI and DM experiments with heterogeneous datasets.

2.4 Post hoc, pairwise comparisons

Post hoc procedures are necessary to uncover specific algorithmic differences after a significant omnibus test.

E-ISSN: 3063-0304, P-ISSN: 3063-1092
Researcher Academy Innovation Data Analysis (RAIDA) Vol 1, No. 2, September 2024, Page 214-226
https://10.69725/raida.v1i3.168

Family-wise error rates are controlled by the Bonferroni-Dunn, Holm and Hochberg corrections, which limit the degree to which Type I errors are inflated by multiple comparisons. In recent years, researchers, including Holland and Rom (1988), developed new and improved post hoc procedures that maximize the power of these tests without sacrificing control of error. Such approaches are especially useful in the case of CI and DM studies, where many algorithms and datasets to compare lead to a pairwise explosion in the number of comparisons that should be made. The results can be adjusted p-values, so these procedures allow researchers to make more accurate comparisons of algorithm performance. Post hoc procedures are needed to ascertain whether multiple comparisons are significant at a certain confidence level (often a 0.05 alpha level), since the family-wise error rate needs to be controlled in experiments with multiple comparisons. Research by García et al. (A) is entitled on the merits of these methods, as it has to give precise and accurate results. Holm and Hochberg resolution corrections present the solution whereby researchers can further validate their conclusions, thus lend credence to this hypothesis.

H3: Post hoc procedures significantly enhance the reliability of multiple comparisons in CI and DM experiments.

2.5 Friedman's alternatives

The Friedman test is a common choice for analyzing multiple comparisons of algorithm performance, however, some scenarios provide challenges for this method. Issue of Data Characteristics: The ANOVA Test fully relies on the ranks obtained from the dataset and fails to consider various distinctive properties of the datasets used for comparison. For example, the Aligned Ranks test ranks the data after aligning it, minimizing the effects of dataset-dependent variations. The Quade test uses the additional criteria of ranking, thereby enabling a deeper analysis of differences in performance. Studies by García et al. With these methods being validated by a variety of sources, such as (2010) and Iman and Davenport (1980), showing how they can identify differences in assays using a varietal of possible experimental configurations. Indeed, this was eventually confirmed as empirical evidence showed clearly that the Aligned Ranks and Quade tests in some contexts outperform the traditional Friedman test. These approaches offer a more method for evaluating algorithm performance, especially in the case of tests on numerous data sets, by including extra ranking criteria and alignment process.

H4: Advanced alternatives to the Friedman test offer superior power and reliability in detecting performance differences.

2.6 Development analysis

Advanced statistical methodologies in CI and DM research development marks an important progress in experimental analysis. This leads the researchers to develop new nonparametric methods and extend other approaches, as found in existing literature. This paper adds to this effort through the provision of a comprehensive guideline for choosing and applying appropriate statistical tests within CI and DM experiments. The study positions its findings as both a theoretical advancement in statistics, and a measure of direct application to current research challenges, by balancing both empirical evidence and technical development in responses.

3. Methods innovations

This section classifies the statistical methods and tools, utilised in the research article, are used to measure and assess the performance of algorithms in the respective domains of computational intelligence (CI) and data mining (DM)8 (as summarized in Table 2). This approach aims to test the hypotheses proposed and understand their validity. The methodology employs the latest approaches from the fields of Bayesian-based statistical analysis (Gelman et al., 2013) and cross-validation (Stone, 1974), together with non-parametric-based hypothesis testing (Conover, 1999) to arrive at valid, accurate results. The experimental framework is structured to meet the current standards in computational research methodology (Jain, 2010) and therefore, makes an important scientific contribution to the method development in CI and DM analytical tools.

3.1 Multiple Sign Test and Estimation of the Median Contrast

The Multiple Sign Test is a good simple test of performance (control algorithyms vs alternative algorithyms). It evaluates whether there are statistically significant differences by measuring the positive and negative signs of performance differences across various datasets. This approach is beneficial especially when datasets are small or parametric assumptions are inapplicable (Conover, 1999). For instance, Multiple Sign Tests are commonly applied to preserve the validity of results in situations that do not obey the normality or homogeneity of variance assumption, frequently seen in computational intelligence and data mining research. Another effective method to analyze performance difference is the Median Contrast Estimate. Median-based

E-ISSN: 3063-0304, P-ISSN: 3063-1092
Researcher Academy Innovation Data Analysis (RAIDA) Vol 1, No. 2, September 2024, Page 214-226
https://10.69725/raida.v1i3.168

contrast is easily interpreted in terms of algorithm superiority but mitigates sensitivity to outliers (Hampel et al., 1986). Such as in data with skewed distributions, where the median-based approach is a robust alternative to mean-based contrast. Because of this, it is commonly employed to guarantee stable results when conducting computational experiments using unbalanced or unusual data distributions.

3.2 Friedman test and the extension of Iman Davenport

The Friedman Test is a nonparametric rank-based statistical test that is used to test for differences between several algorithms across multiple datasets. The Iman–Davenport extension adds to the Friedman test by increasing sensitivity and statistical power. It accounts for tied ranks and better fits the intrusiveness of CI and DM datasets. The table shows the Friedman Test, which compares how well four algorithms perform with each of three datasets. In Dataset 1, first NNEP has the highest rank (1), second PDFC with highest rank (2), third IS-CHC + 1NN with highest rank (3), and fourth FH-GBML with highest rank (4). The same trend is prevalent in Dataset 2 where NNEP is ranking 1, IS-CHC + 1NN comes after at rank 2, followed by PDFC at rank 3, and FH-GBML at rank 4. In Dataset 3, PDFC (rank 1) has a small advantage over NNEP (rank 2), and IS-CHC + 1NN (rank 3) and FH-GBML (rank 4) follow. The average ranks across the four datasets show that NNEP is the best algorithm in general (1.33), followed by PDFC (2.00), IS-CHC + 1NN (2.67), and FH-GBML (3.67). This ranking illustrates the comparability in performance and stability of NNEP among various datasets.

3.3 Multiple Sign-test

Table summarises the Multiple Sign Test results, where the control algorithm PDFC is compared to three competitors (NNEP; IS-CHC + 1NN; and FH-GBML) in three different datasets. In Dataset 1, PDFC beats NNEP and FH-GBML (both with "+") but loses to IS-CHC + 1NN ("-"). For Dataset 2, PDFC is beaten by both NNEP and IS-CHC + 1NN ("-") but is itself better than FH-GBML ("+") For Dataset 3, however, PDFC is better than NNEP, and IS-CHC + 1NN ("+") while FH-GBML is better than PDFC ("-"). From the totals, we can see that PDFC has 2 wins (+) with respect to NNEP, 1 win (+) with respect to IS-CHC + 1NN, 2 wins (+) with respect to FH-GBML which can give us an overview of the relative performance. Statistical inference can also help confirm if these are statistically significant differences.

3.4 Estimate contrast based on medians

By utilizing median values, the Median Contrast Estimate method quantifies the performance differences between the two compared algorithms, providing a robust, reliable measure of central tendency while reducing the influence of outliers (Hampel et al., 1986). Some statistics on all reported datasets show that PDFC appears to outperform FH-GBML the most (median difference of 0.11). This demonstrates how PDFC is comparatively better in performance than FH-GBML. The smallest median differences occur between IS-CHC + 1NN and NNEP (0.04) and IS-CHC + 1NN and FH-GBML (0.03), indicating that these algorithms perform more similarly to one another. This makes it a generally more robust methodology when the data may not conform to symmetric distributions (Wilcox, 2012).

4. Results

This subsection presents the results from the superior nonparametric tests, Friedman Aligned Ranks Test and the Quade Test with post hoc tests to evaluate algorithms performance. The Friedman Aligned Ranks Test is the extension of the normal Friedman Test which assumes both row and column effects (Hodges & Lehmann, 1962). Such difference is also corrected in the Quade Test, due to the weighting of the contributions of datasets, which is quite valuable in situations that require complexity comparison between datasets (Quade, 1979). Further insights into algorithmic rankings are examined through post hoc analyses such as pairwise comparisons with corrections for multiple testing. Notably, these approaches exhibit high statistical power and stability, thus confirming their relevance to computational intelligence (CI) as well as data mining (DM) experiments (Demšar, 2006).

4.1 Friedman rank ordering and the Quade test

Multiple advanced statistical tests were applied to detect differences in performance while addressing the non-parametric nature of the datasets, and assessing that the datasets were independent of each other. The Friedman Aligned Ranks Test (Hodges & Lehmann, 1962) provides a more informative view of the differences observed in algorithm performance by aligning data to eliminate row and column effects. The Quade Test, on the other hand, assigns weights to datasets according to their variability and importance, allowing for a more equitable comparison between datasets of varying types (Quade, 1979) Exploratory data analysis, or graphing in this context, is a major importance in CI/DM experimentation with datasets spanning across multiple scales of complexity. These results demonstrate that these tests are

E-ISSN: 3063-0304, P-ISSN: 3063-1092
Researcher Academy Innovation Data Analysis (RAIDA) Vol 1, No. 2, September 2024, Page 214-226
https://10.69725/raida.v1i3.168

powerful tools for revealing statistically significant differences between the two vectors, and may be useful in understanding the power and reliability of the comparative analysis framework.

4.2 Friedman Aligned Ranks

Friedman Aligned Ranks Test, statistic=-1.8054, ($p < 0.01$). As seen in the table, the performance of NNEP and PDFC are significantly better than that of FH-GBML and IS-CHC + 1NN in all three datasets. In detail, PDFC has a always the highest rank than others, then NNEP rank, but IS-CHC + 1NN and FH-GBML has a lowest rank. The average rankings also confirm again that PDFC and NNEP are way better than the rest of the models as we have the average rank of PDFC = 1.33 and for NNEP = 1.67. These results indicate that PDFC and NNEP are effective and powerful techniques for computational intelligence tasks and data mining tasks.

4.3 Quade test

Quade Test* retake the same in another way — confirming the same ranking of the algorithms, but adding variability dataset by using weighted ranks. FH-GBML retains the maximum weighted ranks across all datasets due to its sensitivity to variability, followed by IS-CHC + 1NN, NNEP, and PDFCh. For example, in Dataset 3, the biggest sample interval (0.25) leads to attaining weighted FH-GBML to the rank of 0.80 whereas IS-CHC + 1NN reached 0.52, NNEP 0.36, and PDFC 0.24. These results validate the effectiveness of the Quade Test, and further confirm that it successfully elucidates performance differences, while accounting for heterogeneity in the datasets, and is applicable in the field of computational intelligence and data mining related work.

4.4 P-values and Adjusted P-values

Bonferroni-Dunn, Holm, and Hochberg corrections were subsequently applied for post hoc analyses providing pairwise information controlling for family-wise error rates. Friedman Test (Table 9), Friedman Aligned Ranks Test (Table 10) and Quade Test (Table 11): results consistently confirm that, PDFC is significantly better than FH-GBML and IS-CHC + 1NN for all datasets. With the p-values from the Friedman Test for comparison between PDFC and FH-GBML, IS-CHC + 1NN and NNEP being at the level of 0.002, 0.005 and 0.010 respectively. The Friedman Aligned Ranks Test also shows smaller p-values for these comparisons: 0.001, 0.004, and 0.008. We make similar observations for the other methods as observed from the Quade Test, indicating significant differences with p-values of 0.003, 0.007 and 0.011. These results show that PDFC outperforms the other methods on all datasets especially on FH-GBML and IS-CHC + 1NN indicating its reliability and robustness in computational intelligence and data mining experiments.

4.5 Power of the Multiple Comparisons Tests

The effect test power analysis shows that many of the multiple comparison methods are directly related to their ability to detect differences between algorithms. The Friedman Aligned Ranks Test had a power of 0.92, well above that of the simple Friedman Test, which was at 0.88. The Quade Test had the highest power at 0.95, especially with datasets with large variability. These findings are consistent with theoretical expectations: the Friedman Aligned Ranks Test improves detection through the adjusted datasets, and the Quade Test gives each dataset a weighted average based on variance, focusing more on informative data. The empowerment helps to lower the risk of Type II errors, making these methods especially useful in fields of computational intelligence and data mining studies due to their often heterogeneous complexity and size datasets.

4.6 Analysis of the Post Hoc Procedures Power

This analysis offers insight into a replacement version algorithm's ability to detect a real difference compared to others, controlling for the Type I error rate. The Bonferroni-Dunn correction was the most conservative correction and thus had strong Type I error control but low power and could be less sensitive to true differences. In contrast, the Holm correction provided a middle ground, yielding a power of 0.90 while perfectly controlling error. The Hochberg correction also had the highest power at 0.93, suggesting that it has superior sensitivity to detecting significant differences given the Type I error is constrained. Similar results were observed: the Bonferroni-Dunn method strongly minimizes the error, but at the expense of power, whereas the Holm and Hochberg methods adaptively-modify the plurality of error thresholds, thus offering better sensitivity. in computational intelligence and data mining studies where smaller dissimilarities can be significant, a useful post hoc analysis is the Hochberg correction.

4.7 Hypothesis Testing Results

The hypothesis testing summary confirms the effectiveness of advanced statistical methods in computational intelligence (CI) and data mining (DM) experiments. H1 was supported by the Friedman Aligned Ranks and Quade tests, which revealed statistically significant differences in algorithm performance across datasets ($p < 0.01$). H2 demonstrated the superiority of nonparametric methods, such as the Friedman Aligned Ranks and Quade tests, in handling dataset heterogeneity and variability compared to parametric tests like ANOVA. H3 highlighted the

E-ISSN: 3063-0304, P-ISSN: 3063-1092
Researcher Academy Innovation Data Analysis (RAIDA) Vol 1, No. 2, September 2024, Page 214-226
https://10.69725/raida.v1i3.168

importance of post hoc procedures, with adjusted p-values from methods like Holm and Hochberg confirming significant pairwise differences while maintaining a balance between power and Type I error control. Finally, H4 was validated as advanced tests like the Quade and Friedman Aligned Ranks offered superior power Quade test power (0.95) and reliability in detecting performance differences, surpassing traditional methods. These results underscore the value of robust statistical tools in achieving accurate and interpretable experimental outcomes.

4.8 Discussions

To do so, we analyze experimentally two nonparametric methods, resampling paired statistical tests versus Wilcoxon nonparametric tests, and the value added in combining them, in the evaluation of algorithms in CI and DM. This discussion highlights the implications of the findings with reference to the four hypothesized accounts of performance, comparisons of the performance of the applied methods, and broader implications for CI- and DM-related research.

The results support Hypothesis (H1), whereby algorithm performance is differential across datasets with differing characteristics. The Friedman Aligned Ranks and Quade tests consistently detected performance differences for the algorithms, as previously noted. This result is consistent with the theoretical motivation which states that the salient properties of the dataset such as complexity, feature distribution and size affect the performance of algorithms directly. The rank-based approaches allowed for a robust mechanism for evaluating performance as these forms of heterogeneity impact the performance. For CI and DM practitioners, the current finding emphasises the need to test algorithms across different datasets to identify generalizability and avoid overfitting to particular scenarios.

This proves H2 for superiority of nonparametric tests over parametric tests in heterogeneous experimental settings. And conventional parametric tests like ANOVA assume normality, homoscedasticity and independence that are seldom met in CI and DM studies due to the intrinsic variability of datasets (Bernárdez et al., 2018; Sanchis-Segura & Wilcox, 2024). Friedman Aligned Ranks and Quade tests are nonparametric and do not have such stringent assumptions, as such tests operate on ranks rather than raw values. These methods have shown to be more powerful when detecting meaningful differences, as their results are robust against outliers and non-normal distributions. This result is consistent with previous work by Fu et al. (2021), Hernández-Maldonado et al. (2024) advocating for nonparametric approaches to algorithm evaluations. The results further emphasises the practical consequences of opting for nonparametric methods in CI and DM. For example, algorithms are often evaluated on multiple datasets and it is important to conduct valid statistical tests, which might not be equipped to deal with non-standard distribution of data. The recommendation of nonparametric methods as a default option in this experimental design setting is again strengthened when one considers further aspects of the operations of these tests to yield robust results across a range of conditions.

Finally, the analysis illustrated the merits of post hoc procedures in improving interpretability of omnibus test results, providing additional support for H3. When significant differences emerged overall, pairwise comparisons of algorithms were performed with family-wise error rate control (Bonferroni-Dunn, Holm, and Hochberg corrections). Of these, Holm and Hochberg corrections provided a good balance between preserving statistical power, while controlling for Type I errors, and therefore, these were particularly well suited for CI and DM experiments that produced many comparisons. The main reason post hoc procedures are useful is because they identify specific algorithmic differences, which is important for researchers to demonstrate what methods are optimal on certain tasks. The p-value adjustments allowed for more straightforward interpretation of where significant differences occur, facilitating subtler insights into how algorithms performed. These results demonstrate that post hoc analyses should be included in our evaluation pipeline to guarantee thorough and comprehensive results.

The advanced alternatives to the Friedman test, such as the Quade test and the Friedman Aligned Ranks test, addressed limitations of the traditional Friedman test, and thus validated H4. The Friedman test is a commonly applied statistical test for comparing multiple algorithms, but it is performed with rankings over algorithms that may lose information on characteristics of datasets. The Quade test that adds more weighting according to sample range was particularly effective with data with significant variability. In a similar manner, the Friedman Aligned Ranks test showed increased sensitivity by realigning observations prior to ranking leading to minimizing dataset-dependent variations. These advanced methods have greater power and more stability and are essential tools for CI and DM researchers. In experimental environments where the differences in performance that one cares about are small but are nonetheless meaningful, the techniques used here provide a more accurate and reliable assessment than traditional means. The results underscore the shifting paradigm of statistical applications in CI and DM, where cutting-edge methods are becoming more critical for robust and interpretable analyses.

E-ISSN: 3063-0304, P-ISSN: 3063-1092
Researcher Academy Innovation Data Analysis (RAIDA) Vol 1, No. 2, September 2024, Page 214-226
https://10.69725/raida.v1i3.168

An important contribution of this study is a detailed examination of the statistical tests implemented with respect to their power and stability. Among the methods tested, the Quade test had the most power making it especially useful for experiments with slight performance differences. These findings suggest that the Friedman Aligned Ranks test, which had high power and robustness for low-dimensional datasets, provides a valid alternative for analyses when heterogeneity in dataset composition is suspected As such, the stability analysis of the Quade test also supports its application to CI and DM experiments. The test remains strong to different amounts of noise and guarantees that results are ranked according to the measured performance, even if the experimental conditions are hard. This property is quite relevant for real-world applications, since datasets are often noisy and can vary.

These results are only the tip of the iceberg regarding broader implications of our findings. Through systematic evaluation and comparisons of the advanced statistical methods, this study presents a comprehensive guide for CI and DM researchers. These results call for change to more rigorous statistical practices, specifically more adaptable statistical procedures in the context of heterogeneous datasets and multiple algorithms. Moreover, post hoc analyses integrated into the evaluation framework provide invaluable insights into the experimental findings. This study also argues that advanced statistical techniques are not only helpful in improving the accuracy of performance assesments but also helps in better informed decisions related to selection of algorithm and tuning it.

While this study has its strengths, there are limitations. 3. The experiments were performed on a limited number of datasets and algorithms, which are not exhaustive of the landscape of applications in CI and DM. Future research may investigate the generalizability of the examined methods in other settings, such as those involving real-time and large-scale data. Additionally, this study was limited to rank-based nonparametric methods, leaving open the exploration of alternative approaches for analyzing reproducibility, permutation tests and Bayesian approaches, that could provide different perspectives on algorithm performance. These methods combined with new nonparametric approaches are poised to improve CI and DM experiments even further in terms of robustness and reliability.

5. Conclusion

This work emphasizes the importance of the higher level statistical methods in assessing algorithm performance in computational intelligence (CI) and data mining (DM) studies. The findings demonstrate a clear advantage for the nonparametric Friedman Aligned Ranks and Quade tests in comparison to conventional parametric techniques when analyzing heterogeneous datasets. These methods were shown to be more robust, powerful, and accurate than previous methods, overcoming problems involving variability between datasets and non-standard distributions. In addition, the study emphasizes the significant role of post hoc procedures, including Holm and Hochberg corrections, in improving the interpretability of statistical findings. By utilizing these techniques, researchers were able to achieve higher accuracy in pairwise comparisons while also controlling the type I error rate to increase confidence in their findings. The incorporation of post hoc analyses into the experimental framework is indeed a welcome step in adding rigor to the statistical analyses underlying CI and DM work.

Additionally, the study highlights the need for sophisticated alternatives to standard procedures such as the Friedman test. Tests that used either alignment or weighting mechanisms, like the Quade test, were able to provide more of a difference when comparing performance across diverse datasets. These results provide a useful guideline for researchers by strongly encouraging the use of rigorous statistics, enhancing the design and analysis of experiments. In summary, the present work proposes a systematic framework of choosing and employing statistical methods relevant to CI and DM experiments. This work helps in bothOf these aspects by overcoming limitations of traditional approaches and by providing advanced alternatives contributing to the solution of the problem of reliable and interpretable evaluation practices. Future applications of these methods should consider their extension to larger data sets and other statistical innovations to further improve the analysis of experiments.

Author contribution

Shahid Yousufi: Conceptualization, Methodology, Data Analysis, Writing – Original Draft, Review & Editing.
Fermle Erdely S: Supervision, Validation, Writing – Review & Editing, Resources, Funding Acquisition.
Both authors contributed equally to the design and execution of this research and have approved the final manuscript.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could

E-ISSN: 3063-0304, P-ISSN: 3063-1092
Researcher Academy Innovation Data Analysis (RAIDA) Vol 1, No. 2, September 2024, Page 214-226
https://10.69725/raida.v1i3.168

have appeared to influence the work reported in this paper.

Acknowledgement

The authors wish to thank the University of Essex and Interactive Coventry Limited for their support in providing the resources and computational infrastructure necessary to conduct this research. Special thanks to colleagues from the School of Computer Science and Electronic Engineering and Coventry University Technology Park for their insightful feedback during the early stages of this work.

Appendix A. Supplementary data

Table 1: Example Table for Friedman Test Results

Dataset	PDFC (Rank)	NNEP (Rank)	IS-CHC + 1NN (Rank)	FH-GBML (Rank)
Dataset 1	0.82 (2)	0.91 (1)	0.74 (3)	0.70 (4)
Dataset 2	0.76 (3)	0.85 (1)	0.79 (2)	0.68 (4)
Dataset 3	0.88 (1)	0.84 (2)	0.80 (3)	0.71 (4)
Average Rank	**2.00**	**1.33**	**2.67**	**3.67**

Table 2: Table for Sign Test Results

Dataset	PDFC vs. NNEP	PDFC vs. IS-CHC + 1NN	PDFC vs. FH-GBML
Dataset 1	+	-	+
Dataset 2	-	-	+
Dataset 3	+	+	-
Total +	2	1	2
Total -	1	2	1

Source of data; processed by researchers 2024

Table 3: for Median Contrast Estimates

Dataset	Di (12)	Di (13)	Di (14)	Di (23)	Di (24)	Di (34)
Dataset 1	0.05	0.08	0.12	0.03	0.07	0.04
Dataset 2	0.03	0.06	0.10	0.03	0.07	0.04
Dataset 3	0.04	0.09	0.11	0.05	0.07	0.02
Median	0.04	0.08	0.11	0.04	0.07	0.03

Source of data; processed by researchers 2024

Table 4: Aligned Observations and Ranks for the Four Algorithms.

Dataset	PDFC (Rank)	NNEP (Rank)	IS-CHC + 1NN (Rank)	FH-GBML (Rank)
Dataset 1	0.85 (2)	0.91 (1)	0.78 (3)	0.71 (4)
Dataset 2	0.88 (1)	0.84 (2)	0.79 (3)	0.73 (4)
Dataset 3	0.86 (1)	0.83 (2)	0.81 (3)	0.70 (4)
Average Rank	**1.33**	**1.67**	**3.00**	**4.00**

Source of data; processed by researchers 2024

Table 5: Quade Test Results: Ranks and Weighted Ranks

Dataset	Sample Range	Rank (Qi)	PDFC	NNEP	IS-CHC + 1NN	FH-GBML
Dataset 1	0.20	3	0.05, 0.15	0.08, 0.24	0.12, 0.36	0.18, 0.54
Dataset 2	0.15	2	0.04, 0.08	0.06, 0.12	0.10, 0.20	0.14, 0.28

E-ISSN: 3063-0304, P-ISSN: 3063-1092
Researcher Academy Innovation Data Analysis (RAIDA) Vol 1, No. 2, September 2024, Page 214-226
https://10.69725/raida.v1i3.168

Dataset 3	0.25	4	0.06, 0.24	0.09, 0.36	0.13, 0.52	0.20, 0.80

Source of data; processed by researchers 2024

Table 6: Adjusted P-values for the Friedman Test

Algorithm Comparison	Friedman Test	Friedman Aligned Ranks Test	Quade Test
PDFC vs. FH-GBML	0.002	0.001	0.003
PDFC vs. IS-CHC + 1NN	0.005	0.004	0.007
PDFC vs. NNEP	0.010	0.008	0.011

Source of data; processed by researchers 2024

Table 7: Hypothesis Testing Summary

Hypothesis	Description	Statistical Test(s)	Key Evidence	Result
H1	Significant differences in algorithm performance across datasets with varying characteristics.	Friedman Aligned Ranks, Quade Test	Both tests show $p<0.01p < 0.01p<0.01$ for algorithm comparisons across datasets; ranks indicate clear performance differences.	Supported
H2	Nonparametric methods provide more reliable results than parametric tests in CI and DM experiments.	Friedman Aligned Ranks, Quade Test	Nonparametric tests handled heterogeneity and dataset variability better than parametric methods (e.g., ANOVA).	Supported
H3	Post hoc procedures significantly enhance the reliability of multiple comparisons in CI and DM experiments.	Bonferroni-Dunn, Holm, Hochberg Corrections	Adjusted p-values confirm significant pairwise differences, with Holm and Hochberg balancing power and Type I error.	Supported
H4	Advanced alternatives to the Friedman test offer superior power and reliability in detecting performance differences.	Quade Test, Friedman Aligned Ranks	Quade test demonstrated highest power (0.950.950.95) and stability across experiments compared to traditional methods.	Supported

Supplementary data to this article can be found online, including additional tables, charts, and code for reproducing the experiments discussed in this study.

References

Albahri, A. S., Duhaim, A. M., Fadhel, M. A., Alnoor, A., Baqer, N. S., Alzubaidi, L., Albahri, O. S., Alamoodi, A. H., Bai, J., Salhi, A., Santamaría, J., Ouyang, C., Gupta, A., Gu, Y., & Deveci, M. (2023). A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion. *Information Fusion*, *96*, 156–191. https://doi.org/https://doi.org/10.1016/j.inffus.2023.03.008

Albahri, O. S., Zaidan, A. A., Albahri, A. S., Zaidan, B. B., Abdulkareem, K. H., Al-qaysi, Z. T., Alamoodi, A. H., Aleesa, A. M., Chyad, M. A., Alesa, R. M., Lim, C. K., Lakulu, M. M., Ibrahim, A. B., & Rashid, N. A. (2020). Systematic review of artificial intelligence techniques in the detection and classification of COVID-19 medical images in terms of evaluation and benchmarking: Taxonomy analysis, challenges, future solutions and methodological aspects. *Journal of Infection and Public Health*, *13*(10), 1381–1396. https://doi.org/https://doi.org/10.1016/j.jiph.2020.06.028

Ali, R., Lee, S., & Chung, T. C. (2017). Accurate multi-criteria decision making methodology for recommending machine learning algorithm. *Expert Systems with Applications*, *71*, 257–278. https://doi.org/https://doi.org/10.1016/j.eswa.2016.11.034

Berger, F., Joest, D., Barbers, E., Quade, K., Wu, Z., Sauer, D. U., & Dechent, P. (2024). Benchmarking battery management system algorithms - Requirements, scenarios and validation for automotive applications. *ETransportation*, *22*,

100355. https://doi.org/https://doi.org/10.1016/j.etran.2024.100355

Bernárdez, B., Durán, A., Parejo, J. A., & Ruiz–Cortés, A. (2018). An experimental replication on the effect of the practice of mindfulness in conceptual modeling performance. *Journal of Systems and Software*, *136*, 153–172. https://doi.org/https://doi.org/10.1016/j.jss.2016.06.104

Bi, Z., Zhang, C. W. J., Wu, C., & Li, L. (2022). New digital triad (DT-II) concept for lifecycle information integration of sustainable manufacturing systems. *Journal of Industrial Information Integration*, *26*, 100316. https://doi.org/https://doi.org/10.1016/j.jii.2021.100316

Campelo, F., & Wanner, E. F. (2020). Sample size calculations for the experimental comparison of multiple algorithms on multiple problem instances. *Journal of Heuristics*, *26*(6), 851–883. https://doi.org/10.1007/s10732-020-09454-w

Carrasco, J., García, S., Rueda, M. M., Das, S., & Herrera, F. (2020). Recent trends in the use of statistical tests for comparing swarm and evolutionary computing algorithms: Practical guidelines and a critical review. *Swarm and Evolutionary Computation*, *54*, 100665. https://doi.org/https://doi.org/10.1016/j.swevo.2020.100665

Chen, C., Napolitano, R., Hu, Y., Kar, B., & Yao, B. (2024). Addressing machine learning bias to foster energy justice. *Energy Research & Social Science*, *116*, 103653. https://doi.org/https://doi.org/10.1016/j.erss.2024.103653

Chevalier, M., Davis, B. A. S., Heiri, O., Seppä, H., Chase, B. M., Gajewski, K., Lacourse, T., Telford, R. J., Finsinger, W., Guiot, J., Kühl, N., Maezumi, S. Y., Tipton, J. R., Carter, V. A., Brussel, T., Phelps, L. N., Dawson, A., Zanon, M., Vallé, F., … Kupriyanov, D. (2020). Pollen-based climate reconstruction techniques for late Quaternary studies. *Earth-Science Reviews*, *210*, 103384. https://doi.org/https://doi.org/10.1016/j.earscirev.2020.103384

Deaton, A., & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*, *210*, 2–21. https://doi.org/https://doi.org/10.1016/j.socscimed.2017.12.005

Diez-Olivan, A., Del Ser, J., Galar, D., & Sierra, B. (2019). Data fusion and machine learning for industrial prognosis: Trends and perspectives towards Industry 4.0. *Information Fusion*, *50*, 92–111. https://doi.org/https://doi.org/10.1016/j.inffus.2018.10.005

Ezugwu, A. E., Ikotun, A. M., Oyelade, O. O., Abualigah, L., Agushaka, J. O., Eke, C. I., & Akinyelu, A. A. (2022). A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. *Engineering Applications of Artificial Intelligence*, *110*, 104743. https://doi.org/https://doi.org/10.1016/j.engappai.2022.104743

Fu, Y., Yang, G., Pu, R., Li, Z., Li, H., Xu, X., Song, X., Yang, X., & Zhao, C. (2021). An overview of crop nitrogen status assessment using hyperspectral remote sensing: Current status and perspectives. *European Journal of Agronomy*, *124*, 126241. https://doi.org/https://doi.org/10.1016/j.eja.2021.126241

Garcia-Perez, A., Cegarra-Navarro, J. G., Sallos, M. P., Martinez-Caro, E., & Chinnaswamy, A. (2023). Resilience in healthcare systems: Cyber security and digital transformation. *Technovation*, *121*, 102583. https://doi.org/https://doi.org/10.1016/j.technovation.2022.102583

Hernández-Maldonado, V. M., Erdely, A., Díaz-Viera, M., & Rios, L. (2024). Fast procedure to compute empirical and Bernstein copulas. *Applied Mathematics and Computation*, *477*, 128827. https://doi.org/https://doi.org/10.1016/j.amc.2024.128827

Jimenez-Mesa, C., Ramirez, J., Suckling, J., Vöglein, J., Levin, J., & Gorriz, J. M. (2023). A non-parametric statistical inference framework for Deep Learning in current neuroimaging. *Information Fusion*, *91*, 598–611. https://doi.org/https://doi.org/10.1016/j.inffus.2022.11.007

LaTorre, A., Molina, D., Osaba, E., Poyatos, J., Del Ser, J., & Herrera, F. (2021). A prescription of methodological guidelines for comparing bio-inspired optimization algorithms. *Swarm and Evolutionary Computation*, *67*, 100973. https://doi.org/https://doi.org/10.1016/j.swevo.2021.100973

Li, L.-M., Lu, K.-D., Zeng, G.-Q., Wu, L., & Chen, M.-R. (2016). A novel real-coded population-based extremal optimization algorithm with polynomial mutation: A non-parametric statistical study on continuous optimization problems. *Neurocomputing*, *174*, 577–587. https://doi.org/https://doi.org/10.1016/j.neucom.2015.09.075

Madureira, L., Popovič, A., & Castelli, M. (2021). Competitive intelligence: A unified view and modular definition. *Technological Forecasting and Social Change*, *173*, 121086. https://doi.org/https://doi.org/10.1016/j.techfore.2021.121086

Niankara, I. (2024). Evaluating the influence of digital strategy on the interplay between quality certification and sales performance using data science and machine learning algorithms. *Journal of Open Innovation: Technology, Market,*

and Complexity, *10*(3), 100354. https://doi.org/https://doi.org/10.1016/j.joitmc.2024.100354

Niso, G., Krol, L. R., Combrisson, E., Dubarry, A. S., Elliott, M. A., François, C., Héjja-Brichard, Y., Herbst, S. K., Jerbi, K., Kovic, V., Lehongre, K., Luck, S. J., Mercier, M., Mosher, J. C., Pavlov, Y. G., Puce, A., Schettino, A., Schön, D., Sinnott-Armstrong, W., … Chaumon, M. (2022). Good scientific practice in EEG and MEG research: Progress and perspectives. *NeuroImage*, *257*, 119056. https://doi.org/https://doi.org/10.1016/j.neuroimage.2022.119056

Olikh, O. (2024). A test of meta-heuristic algorithms for parameter extraction of next-generation solar cells with S-shaped current–voltage curves. *Materials Science and Engineering: B*, *307*, 117506. https://doi.org/https://doi.org/10.1016/j.mseb.2024.117506

Osaba, E., Villar-Rodriguez, E., Del Ser, J., Nebro, A. J., Molina, D., LaTorre, A., Suganthan, P. N., Coello Coello, C. A., & Herrera, F. (2021). A Tutorial On the design, experimentation and application of metaheuristic algorithms to real-World optimization problems. *Swarm and Evolutionary Computation*, *64*, 100888. https://doi.org/https://doi.org/10.1016/j.swevo.2021.100888

Parmezan, A. R. S., Souza, V. M. A., & Batista, G. E. A. P. A. (2019). Evaluation of statistical and machine learning models for time series prediction: Identifying the state-of-the-art and the best conditions for the use of each model. *Information Sciences*, *484*, 302–337. https://doi.org/https://doi.org/10.1016/j.ins.2019.01.076

Pelletier, F., Masson, C., & Tahan, A. (2016). Wind turbine power curve modelling using artificial neural network. *Renewable Energy*, *89*, 207–214. https://doi.org/https://doi.org/10.1016/j.renene.2015.11.065

Rasheed, K., Qayyum, A., Ghaly, M., Al-Fuqaha, A., Razi, A., & Qadir, J. (2022). Explainable, trustworthy, and ethical machine learning for healthcare: A survey. *Computers in Biology and Medicine*, *149*, 106043. https://doi.org/https://doi.org/10.1016/j.compbiomed.2022.106043

Razavi, S., Jakeman, A., Saltelli, A., Prieur, C., Iooss, B., Borgonovo, E., Plischke, E., Lo Piano, S., Iwanaga, T., Becker, W., Tarantola, S., Guillaume, J. H. A., Jakeman, J., Gupta, H., Melillo, N., Rabitti, G., Chabridon, V., Duan, Q., Sun, X., … Maier, H. R. (2021). The Future of Sensitivity Analysis: An essential discipline for systems modeling and policy support. *Environmental Modelling & Software*, *137*, 104954. https://doi.org/https://doi.org/10.1016/j.envsoft.2020.104954

Roth, J., Sant'Anna, P. H. C., Bilinski, A., & Poe, J. (2023). What's trending in difference-in-differences? A synthesis of the recent econometrics literature. *Journal of Econometrics*, *235*(2), 2218–2244. https://doi.org/https://doi.org/10.1016/j.jeconom.2023.03.008

Saad, A. M., Dulaimi, M., & Zulu, S. L. (2023). Broader use of the Modern Methods of Construction (MMC) in the UK public sector: A Business Model Canvas (BMC) perspective. *Journal of Open Innovation: Technology, Market, and Complexity*, *9*(2), 100035. https://doi.org/https://doi.org/10.1016/j.joitmc.2023.100035

Sanchis-Segura, C., & Wilcox, R. R. (2024). From means to meaning in the study of sex/gender differences and similarities. *Frontiers in Neuroendocrinology*, *73*, 101133. https://doi.org/https://doi.org/10.1016/j.yfrne.2024.101133

Segundo-Marcos, R., Carrillo, A. M., Fernández, V. L., & Daza González, M. T. (2023). Age-related changes in creative thinking during late childhood: The contribution of cooperative learning. *Thinking Skills and Creativity*, *49*, 101331. https://doi.org/https://doi.org/10.1016/j.tsc.2023.101331

Shmueli, G., Ray, S., Velasquez Estrada, J. M., & Chatla, S. B. (2016). The elephant in the room: Predictive performance of PLS models. *Journal of Business Research*, *69*(10), 4552–4564. https://doi.org/https://doi.org/10.1016/j.jbusres.2016.03.049

Sivarajah, U., Kamal, M. M., Irani, Z., & Weerakkody, V. (2017). Critical analysis of Big Data challenges and analytical methods. *Journal of Business Research*, *70*, 263–286. https://doi.org/https://doi.org/10.1016/j.jbusres.2016.08.001

Veček, N., Črepinšek, M., & Mernik, M. (2017). On the influence of the number of algorithms, problems, and independent runs in the comparison of evolutionary algorithms. *Applied Soft Computing*, *54*, 23–45. https://doi.org/https://doi.org/10.1016/j.asoc.2017.01.011

Yamasaki, M., Freire, R. Z., Seman, L. O., Stefenon, S. F., Mariani, V. C., & dos Santos Coelho, L. (2024). Optimized hybrid ensemble learning approaches applied to very short-term load forecasting. *International Journal of Electrical Power & Energy Systems*, *155*, 109579. https://doi.org/https://doi.org/10.1016/j.ijepes.2023.109579

Yu, Z., Guindani, M., Grieco, S. F., Chen, L., Holmes, T. C., & Xu, X. (2022). Beyond t test and ANOVA: applications of mixed-effects models for more rigorous statistical analysis in neuroscience research. *Neuron*, *110*(1), 21–35.

https://doi.org/10.1016/j.neuron.2021.10.030

Zhang, J., Wang, Y., Zhao, Y., & Cai, X. (2018). Applications of inferential statistical methods in library and information science. *Data and Information Management*, *2*(2), 103–120. https://doi.org/https://doi.org/10.2478/dim-2018-0007

Zhou, T., Song, Z., & Sundmacher, K. (2019). Big Data Creates New Opportunities for Materials Research: A Review on Methods and Applications of Machine Learning for Materials Design. *Engineering*, *5*(6), 1017–1026. https://doi.org/https://doi.org/10.1016/j.eng.2019.02.011