|  |  |  |
| --- | --- | --- |
| National Poverty Severity and Depth Index Analysis using PySpark Linear Regression Method | | |
|  | | |
| Febro Herdyanto1 | | |
| 1)Teknik Informatika, Pelita Bangsa University, West Java, Indonesia | | |
|  |  |  |
| Article Info |  | Abstract |
| *Article history*  Received : diisi oleh editor  Revised : diisi oleh editor  Accepted : diisi oleh editor |  | *This research presents a thorough examination of poverty indices in Indonesia, employing PySpark for a comprehensive regional and sub-regional analysis. The primary objective is to unveil nuanced patterns and determinants influencing poverty, providing evidence-based insights for targeted policy interventions. The initial data preprocessing involves meticulous steps, including handling missing values and casting columns to ensure data quality. The exploration phase centers on Jawa Barat, where predictions are visually depicted through bar charts for both "Index Kedalaman Kemiskinan" and "Index Keparahan Kemiskinan." Subsequently, the analysis extends to the sub-regional level (kabupaten/kota), where distinct linear regression models are trained and visualized using line charts. The emphasis is on highlighting the top 5 regions displaying substantial reductions in poverty indices. The research findings contribute valuable insights into regional and sub-regional poverty variations, supporting evidence-based decision-making for effective poverty alleviation strategies in Indonesia. By leveraging machine learning techniques, this study demonstrates the potential for predictive models to offer actionable insights, enriching the ongoing discourse on poverty alleviation and providing a robust foundation for evidence-based policymaking.* |
|  |  |
| *Kata Kunci:*  *Linear Regression;*  *National Poverty;*  *PySpark Analysis;*  *Severity and Depth Index* |  |
| ***Corresponding Author:***  Febro Herdyanto,  Teknik Informatika  Pelita Bangsa Univerisity  Jl. Inspeksi Kalimalang Tegal Danas, Bekasi, West Java, Indonesia  febroherdyanto@mhs.pelitabangsa.ac.id | | |
| *This is an open access article under the* [*CC BY-NC*](https://creativecommons.org/licenses/by-nc/4.0/) *license.* | | |

1. **Introduction**

Poverty remains a persistent and complex challenge in Indonesia, necessitating rigorous attention and comprehensive efforts for alleviation. Despite ongoing initiatives and programs aimed at poverty reduction, the multifaceted nature of this issue demands in-depth analysis. Contributing factors include uneven economic growth, low education levels, regional disparities, and the effectiveness of public policies.

Unequal economic growth exacerbates economic disparities, leading to income inequality. Insufficient education impedes individuals' access to quality employment, while regional inequalities contribute to varying poverty levels across different areas. Assessing the impact of public policies on economic development is crucial in the context of poverty reduction. Effective public policies can significantly contribute to addressing poverty-related issues. Evaluating the effectiveness of implemented policies becomes a crucial step in identifying weaknesses and assessing their impact on poverty rates in Indonesia.

This research utilizes data from the Badan Pusat Statistik (BPS) Indonesia, spanning from 2015 to 2023. The primary focus is on the Depth and Severity Indices of Poverty, considered a robust foundation for understanding the dynamics and characteristics of poverty in Indonesia over recent years.

By employing Linear Regression and PySpark methods, this study aims to provide new insights into the relationships between economic growth, education levels, regional disparities, public policies, and poverty rates. A nuanced understanding of these factors is expected to lead to the formulation of targeted and effective policies for addressing the complex issue of poverty in Indonesia.

1. **Research Method**

The research methodology employed in this study is structured to comprehensively address key issues contributing to poverty in Indonesia. The analysis begins with a thorough examination of uneven economic growth, focusing on data collected from various regions and sectors between 2015 and 2023. Statistical tools, particularly Linear Regression, are applied to assess the contributions of both informal and formal sectors to economic growth, providing insights into their implications for income distribution and poverty.

Subsequently, the research delves into the impact of low education levels on poverty. Education-related data from the Badan Pusat Statistik (BPS) is scrutinized, emphasizing the evaluation of educational programs and regional accessibility. The analysis aims to discern the correlation between education levels and employment opportunities, thereby shedding light on the role of education in poverty alleviation.

Regional inequality, a critical concern in poverty management, is then addressed. Poverty distribution data across different regions is collected and analyzed to evaluate the extent to which regional disparities contribute to overall poverty rates. This analysis assists in identifying specific regions experiencing high levels of poverty, forming the basis for the design of targeted policies tailored to local conditions.

Furthermore, the study assesses the effectiveness of public policies as a key factor in poverty reduction. Data on implemented public policies, welfare programs, economic incentives, and educational aid are gathered and subjected to rigorous evaluation. The focus is on measuring the impact and sustainability of these policies, providing valuable insights into their role in addressing poverty-related challenges.

In terms of data collection, the study relies on comprehensive data from the Badan Pusat Statistik (BPS) for the period 2015-2023. The main variables of interest are the Depth and Severity Indices of Poverty, providing a nuanced understanding of the multifaceted nature of poverty in Indonesia.

The methodology involves integrating economic and social datasets, utilizing the strengths of PySpark for efficient data processing. The use of Linear Regression as the primary analytical tool facilitates the identification of relationships between key variables, such as economic growth, education levels, and regional disparities, and the Depth and Severity Indices of Poverty.

The research also incorporates a predictive element, employing model testing and experimentation using Linear Regression. The dataset is divided into training and testing sets to evaluate the model's predictive capabilities. Model evaluation metrics such as Mean Squared Error (MSE) or R-squared are utilized, and adjustments are made to enhance the model's predictive performance based on the evaluation results.

In conclusion, the proposed research methodology combines quantitative analysis, statistical tools, and predictive modeling to offer a comprehensive understanding of the complex dynamics of poverty in Indonesia. The structured approach ensures the reliability and relevance of the findings, contributing valuable insights for policymakers and stakeholders involved in poverty alleviation efforts.

1. **Result and Discussion**

Prior to delving into the intricate details of Data Preprocessing and Exploration, the initial phase of our analysis involved the careful curation and loading of the poverty indices dataset. This foundational step ensured a comprehensive understanding of the dataset's structure and laid the groundwork for subsequent insightful examinations

1. Data Preprocessing and Exploration

The initial phase of the analysis involved the preprocessing of the dataset using PySpark. The dataset, containing information on poverty indices, was loaded into a PySpark DataFrame. The schema of the dataset was examined to understand the structure and types of variables. To ensure data quality, missing values (null or NaN) in the 'data\_content' column were identified and subsequently removed from the dataset. The 'nama\_tahun' and 'data\_content' columns were cast to float types for further analysis.

|  |
| --- |
| df.printSchema() |

Figure 1. Code for print schema of dataset.

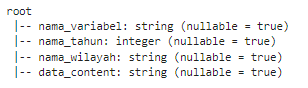


Figure 2. Output schema of dataset.

|  |
| --- |
| null\_count = df.filter(col("data\_content").isNull() | isnan("data\_content")).count()  print(f"Jumlah Null/NaN pada data\_content: {null\_count}") |

Figure 3. Code to display the number of null/NaN values in the data\_content column.



Figure 4. Output of display the number of null/NaN values in the data\_content column.

|  |
| --- |
| columns\_to\_cast = ["data\_content", "nama\_tahun"]  for col\_name in columns\_to\_cast:  df = df.withColumn(col\_name, df[col\_name].cast("float")) |

Figure 5. Convert data types of data\_content, nama\_tahun, and other variables to float.

1. Region-specific Analysis

The analysis focused on a specific region, in this case, "JAWA BARAT." The dataset was filtered to include only relevant entries for this region. The number of rows after filtering was verified to ensure sufficient data for model training. Subsequently, a Vector Assembler was employed to combine the 'nama\_tahun' column into a feature vector. A Linear Regression model was then trained to predict the 'data\_content' values.

|  |
| --- |
| # Ensure there is training data  if jumlah\_baris\_setelah\_filtering > 0:  # Use VectorAssembler to combine columns into a feature vector  assembler = VectorAssembler(inputCols=["nama\_tahun"], outputCol="features")  df\_assembled = assembler.transform(df\_wilayah)  # Linear Regression Model  lr = LinearRegression(featuresCol="features", labelCol="data\_content")  model = lr.fit(df\_assembled)  # Predictions  predictions = model.transform(df\_assembled)  # Fetch data for plotting  actual\_data = df\_wilayah.select("nama\_tahun", "data\_content").toPandas()  predicted\_data = predictions.select("nama\_tahun", "prediction").toPandas()  # Plotting using a bar chart  plt.figure(figsize=(12, 6))  # Chart for Poverty Depth Index  plt.subplot(1, 2, 1)  plt.bar(actual\_data["nama\_tahun"], actual\_data["data\_content"], label="Actual")  plt.bar(predicted\_data["nama\_tahun"], predicted\_data["prediction"], label="Predicted", alpha=0.7)  plt.title(f"Index Kedalaman Kemiskinan - {wilayah}")  plt.xlabel("Tahun")  plt.ylabel("Index Value")  plt.legend()  # Chart for Poverty Severity Index  plt.subplot(1, 2, 2)  plt.bar(actual\_data["nama\_tahun"], actual\_data["data\_content"], label="Actual")  plt.bar(predicted\_data["nama\_tahun"], predicted\_data["prediction"], label="Predicted", alpha=0.7)  plt.title(f"Index Keparahan Kemiskinan - {wilayah}")  plt.xlabel("Tahun")  plt.ylabel("Index Value")  plt.legend()  plt.tight\_layout()  plt.show()  else:  print(f"Tidak ada data yang cukup untuk melatih model untuk {wilayah}.") |

Figure 6. Data Filtering, Feature Vector Creation, and Model Training for Region-specific

1. Model Evaluation

The trained model's performance was evaluated using standard metrics, including Root Mean Squared Error (RMSE) and R-squared (R2). These metrics provide insights into the accuracy and goodness of fit of the regression model.

|  |
| --- |
| # Model Evaluation  evaluation = model.evaluate(df\_assembled)  # Fetch evaluation metrics  rmse = evaluation.rootMeanSquaredError  r2 = evaluation.r2  # Display evaluation metrics  print(f"Root Mean Squared Error (RMSE): {rmse}")  print(f"R-squared (R2): {r2}") |

Figure 7. Code to check Model Evaluation



Figure 8. Output of Model Evaluation

1. Visualization of Predictions

The predictions of the model were visualized using bar charts for both the "Index Kedalaman Kemiskinan" and "Index Keparahan Kemiskinan." The actual and predicted values were plotted side by side, providing a clear comparison over the years. This visualization aids in understanding the model's ability to capture the underlying patterns in the data.

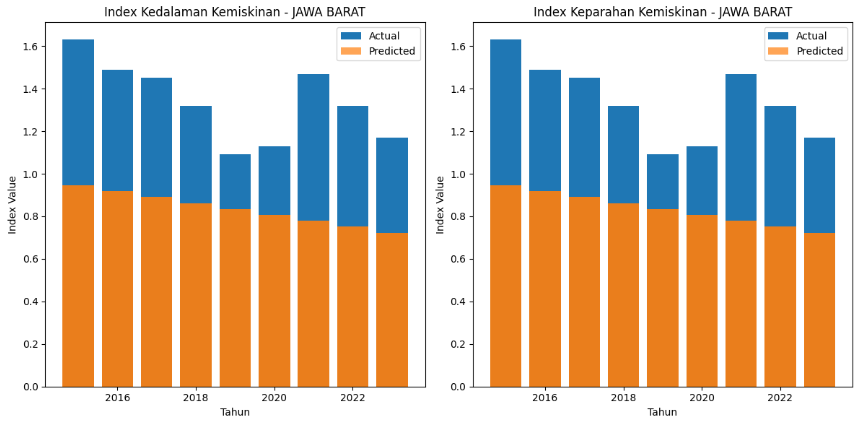


Figure 9. Visualization of Region-specific analysis

1. Regional Disparities

To address regional disparities, the analysis extended to focus on provinces. Separate models were trained for both "Index Kedalaman Kemiskinan" and "Index Keparahan Kemiskinan" at the provincial level. Line charts were created using Plotly Express, showcasing the actual values over time and overlaying them with predicted values.

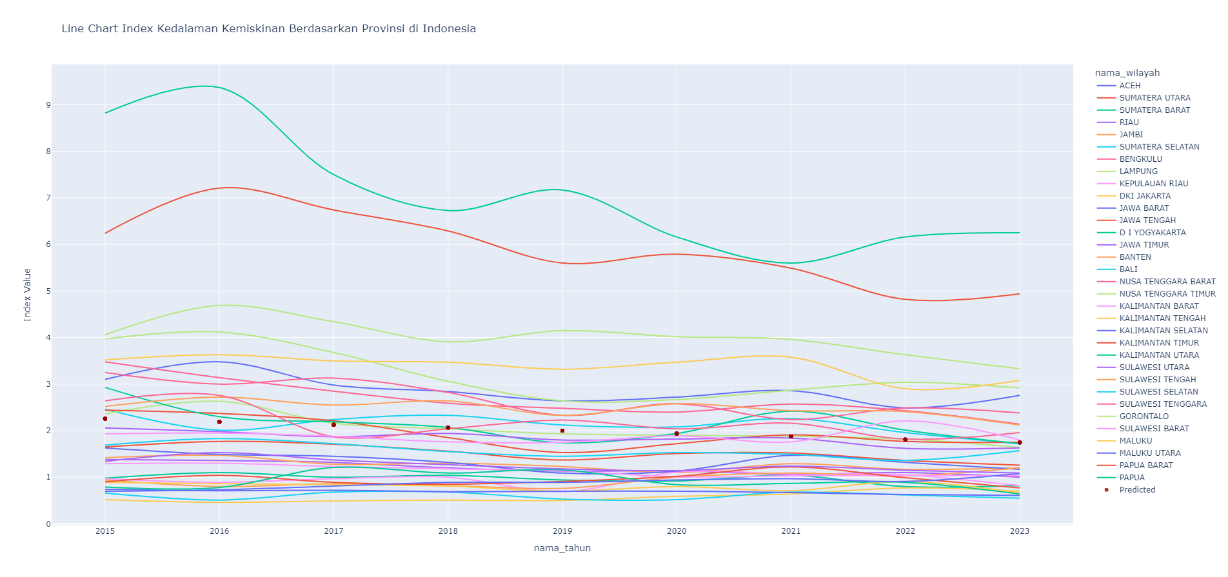


Figure 10. Provincial Analysis with Predictive Modeling - Index of Poverty Depth

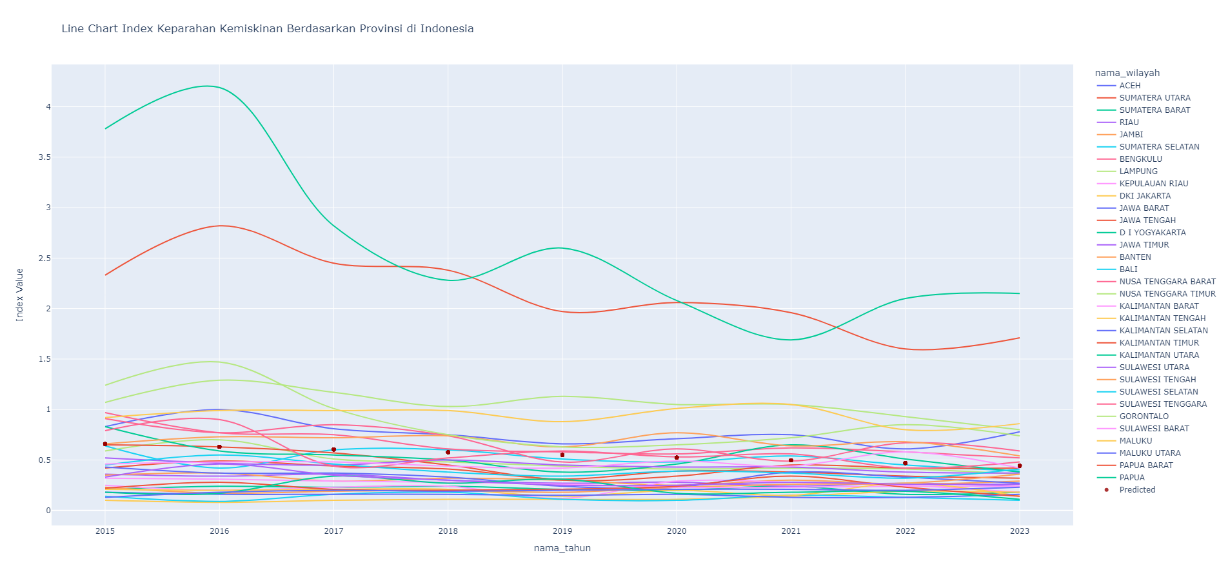


Figure 11. Provincial Analysis with Predictive Modeling - Index of Poverty Severity

1. Top 5 Regions with Significant Reduction

An additional analysis highlighted the top 5 regions with the most significant reduction in poverty indices. Line charts were generated for both depth and severity indices, displaying the actual and predicted values. This detailed analysis of specific regions provides valuable insights for policymakers and stakeholders

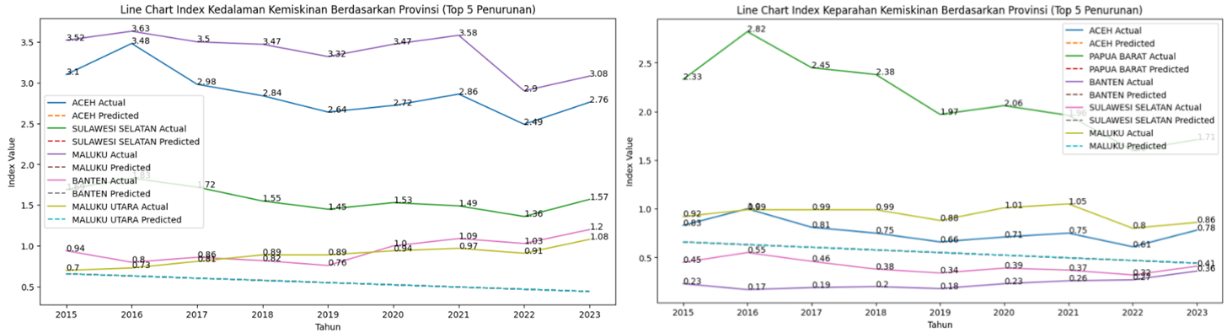


Figure 12. Top 5 Regions Analysis - Index of Poverty Depth and Severity

1. Sub-Regional Analysis

The analysis delved further into sub-regional data, specifically focusing on kabupaten/kota (districts/cities). Linear Regression models were trained separately for both "Index Kedalaman Kemiskinan" and "Index Keparahan Kemiskinan" at this level. Visualizations, including line charts, illustrated the predicted and actual values, emphasizing the top 5 regions with substantial reductions.

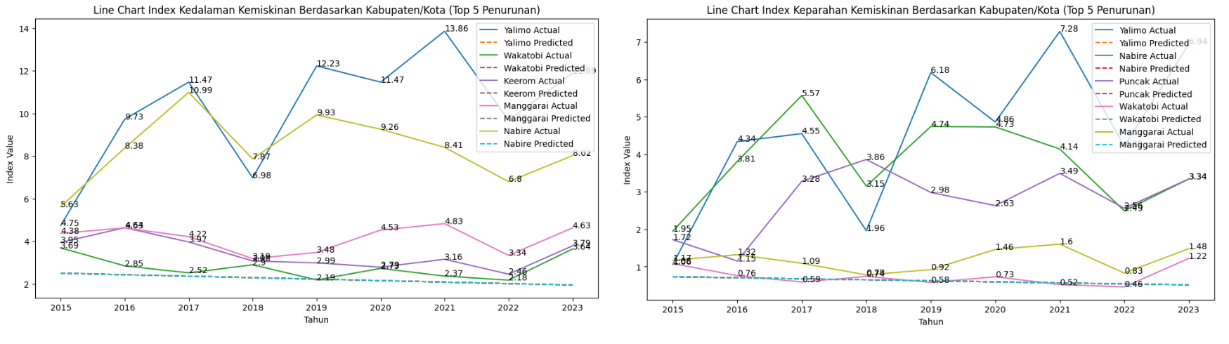


Figure 13. Top 5 Sub-Regions Analysis - Index of Poverty Depth and Severity

1. Insights and Implications

The results from this comprehensive analysis offer insights into the factors influencing poverty indices in Indonesia. Regional and sub-regional variations have been explored, providing a nuanced understanding of the poverty landscape. The predictive models demonstrate the potential for leveraging machine learning techniques to gain actionable insights for targeted policy interventions. These findings contribute to the ongoing discourse on poverty alleviation strategies in Indonesia and provide a foundation for evidence-based decision-making.

1. **Conclusion**

This research sheds light on the intricate dynamics of poverty in Indonesia, offering a comprehensive understanding of contributing factors and potential solutions. Unequal economic growth, low education levels, regional disparities, and policy effectiveness were identified as crucial elements influencing poverty rates.

The analysis, spanning from 2015 to 2023, utilized data from the Badan Pusat Statistik (BPS), focusing on the Depth and Severity Indices of Poverty. Employing Linear Regression and PySpark methods, the study unraveled insights into the relationships between economic growth, education, regional variations, and public policies with poverty rates.

The research methodology, incorporating statistical tools and predictive modeling, facilitated a nuanced exploration of economic disparities and educational impacts on poverty. It further highlighted regional inequalities and critically assessed the effectiveness of public policies.

Key findings included the exacerbation of income inequality due to uneven economic growth, the role of education in shaping employment opportunities, and the need for targeted policies addressing regional disparities. The predictive models, especially at regional and sub-regional levels, provided actionable insights for policymakers.

Model evaluations using metrics like Root Mean Squared Error (RMSE) and R-squared demonstrated the reliability of the Linear Regression approach. Visualizations, including bar and line charts, enhanced the interpretation of predictions and regional disparities, emphasizing the top 5 regions with substantial poverty reductions.

In essence, the research contributes to the ongoing discourse on poverty alleviation in Indonesia, offering evidence-based insights for policymakers and stakeholders. The findings underscore the importance of tailored policies that consider regional nuances and the multifaceted nature of poverty, paving the way for effective strategies in the country's pursuit of sustainable development and poverty reduction.

**References**

(one blank single space line, 10 pt)

References must be written in the order they appear in the text, using Arabic numerals in square brackets, as shown in the example. References of at least 25 journal articles in the last 5 years, 80% of which must be sourced from international reputable journals and try the reference has a URL/Link. References must use a Reference Management application such as Mendeley or similar application and use IEEE style and 9pt font size.

[1] R. E. Caraka *et al.*, “Impact of COVID-19 large scale restriction on environment and economy in Indonesia,” *Glob. J. Environ. Sci. Manag.*, vol. 6, no. Special Issue (Covid-19), pp. 65–84, 2020.

[2] M. Machida *et al.*, “Incorrect use of face masks during the current COVID-19 pandemic among the general public in Japan,” *Int. J. Environ. Res. Public Health*, vol. 17, no. 18, p. 6484, 2020.

[3] W. Bourequat and H. Mourad, “Sentiment Analysis Approach for Analyzing iPhone Release using Support Vector Machine,” *Int. J. Adv. Data Inf. Syst.*, vol. 2, no. 1, pp. 36–44, 2021, doi: 10.25008/ijadis.v2i1.1216.

[4] P. Chen, R. Tong, B. Yu, and Y. Wang, “Reliable shortest path finding in stochastic time-dependent road network with spatial-temporal link correlations: A case study from Beijing,” *Expert Syst. Appl.*, vol. 147, 2020, doi: 10.1016/j.eswa.2020.113192.

[5] D.-H. Oh, J. Ahn, S. Lee, and H. Choi, “Measuring technical inefficiency and CO2 shadow price of Korean fossil-fuel generation companies using deterministic and stochastic approaches,” *Energy Environ.*, p. 0958305X20932547, 2020.

[6] Y. J. Han *et al.*, “Treatment of severe acute respiratory syndrome (SARS), Middle East respiratory syndrome (MERS), and coronavirus disease 2019 (COVID-19): a systematic review of in vitro, in vivo, and clinical trials,” *Theranostics*, vol. 11, no. 3, p. 1207, 2021.

[7] A. Hocine, M. S. Guellil, E. Dogan, S. Ghouali, and N. Kouaissah, “A fuzzy goal programming with interval target model and its application to the decision problem of renewable energy planning,” *Environ. Ecol. Stat.*, vol. 27, no. 3, pp. 527–547, 2020.

[8] Z. Zhai, J. Fernán Martínez, N. Lucas Martínez, and V. Hernández Díaz, “Applying case-based reasoning and a learning-based adaptation strategy to irrigation scheduling in grape farming,” *Comput. Electron. Agric.*, vol. 178, p. 105741, Nov. 2020, doi: 10.1016/j.compag.2020.105741.

[9] A. M. Catai, C. M. Pastre, M. F. de Godoy, E. da Silva, A. C. de Medeiros Takahashi, and L. C. M. Vanderlei, “Heart rate variability: are you using it properly? Standardisation checklist of procedures,” *Brazilian J. Phys. Ther.*, vol. 24, no. 2, pp. 91–102, 2020.

[10] B. M. F. Viana, L. T. Pereira, C. F. M. Toledo, S. R. dos Santos, and S. M. D. M. Maia, “Feasible–Infeasible Two-Population Genetic Algorithm to evolve dungeon levels with dependencies in barrier mechanics,” *Appl. Soft Comput.*, vol. 119, p. 108586, 2022.

[11] S. R. Jayashree, G. Dias, J. J. Andrew, S. Saha, F. Maurel, and S. Ferrari, “Multimodal Web Page Segmentation Using Self-organized Multi-objective Clustering,” *ACM Trans. Inf. Syst.*, vol. 40, no. 3, pp. 1–49, 2022.

[12] A. Mughaid *et al.*, “An intelligent cybersecurity system for detecting fake news in social media websites,” *Soft Comput.*, pp. 1–15, 2022.

[13] O. Nahli, A. Sanna, M. Bandini, and F. Boschetti, “Commerce Numérique: traffic signals for the crossroads between cultures,” *Int. J. Inf. Sci. Technol.*, vol. 5, no. 3, pp. 36–45, 2022.

[14] M. Kumar, J. Aggarwal, A. Rani, T. Stephan, A. Shankar, and S. Mirjalili, “Secure video communication using firefly optimization and visual cryptography,” *Artif. Intell. Rev.*, vol. 55, no. 4, pp. 2997–3017, 2022.

[15] M. K. Gupta *et al.*, “Detection and localization for watermarking technique using LSB encryption for DICOM Image,” *J. Discret. Math. Sci. Cryptogr.*, pp. 1–12, 2022.

[16] M. Salehi, G. De Borger, D. Hughes, and B. Crispo, “NemesisGuard: Mitigating interrupt latency side channel attacks with static binary rewriting,” *Comput. Networks*, p. 108744, 2022.

[17] G. Palumbo *et al.*, “A Structured Approach for the Design and Manufacturing of Titanium Cranial Prostheses via Sheet Metal Forming,” *Metals (Basel).*, vol. 12, no. 2, p. 293, 2022.

[18] J.-S. Park, J.-H. Kim, Y.-C. Jeong, H.-T. Kim, S.-K. Kim, and J.-M. Lee, “Effect of Corrugated Sheet Diameter on Structural Behavior under Cryogenic Temperature and Hydrodynamic Load,” *Metals (Basel).*, vol. 12, no. 3, p. 521, 2022.

[19] P. Stachelek, L. MacKenzie, D. Parker, and R. Pal, “Circularly polarised luminescence laser scanning confocal microscopy to study live cell chiral molecular interactions,” *Nat. Commun.*, vol. 13, no. 1, pp. 1–8, 2022.

[20] D. Mezad-Koursh *et al.*, “Response to: Comment on:‘Peripapillary hyperreflective ovoid mass-like structures—a novel entity as frequent cause of pseudopapilloedema in children,’” *Eye*, p. 1, 2022.

[21] P. Chen, L. Xu, and Z. Li, “Modeling methods of self-centering energy dissipation braces using OpenSees,” *J. Constr. Steel Res.*, vol. 193, p. 107267, 2022.

[22] D. Stanford, Z. Yang, and S. Yao, “Subleading weingartens,” *J. High Energy Phys.*, vol. 2022, no. 2, pp. 1–50, 2022.

[23] B. Joksimoski *et al.*, “Technological solutions for sign language recognition: a scoping review of research trends, challenges, and opportunities,” *IEEE Access*, 2022.

[24] J. K. Lê and T. Schmid, “The practice of innovating research methods,” *Organ. Res. Methods*, vol. 25, no. 2, pp. 308–336, 2022.

[25] J. Miller and R. Ulrich, “Optimizing Research Output: How Can Psychological Research Methods Be Improved?,” *Annu. Rev. Psychol.*, vol. 73, pp. 691–718, 2022.

[26] S. Hosseinpour, A. Gaudin, and O. A. Peters, “A critical analysis of research methods and experimental models to study biocompatibility of endodontic materials,” *Int. Endod. J.*, 2022.