ISSN 1330-3651 (Print), ISSN 1848-6339 (Online)
https://doi.org/10.17559/TV-20231102001073
Original scientific paper

Data Mining, Machine Learning, and Statistical Modeling for Predictive Analytics with Behavioral Big Data

M. ARUNKUMAR, K. RAJKUMAR, W. R. SALEM JEYASEELAN, N. A. NATRAJ*

Abstract: This research delves into the transformative impact of the widespread adoption of big data and advancements in predictive analytics on decision-making processes across industries. The study specifically concentrates on the paradigm of behavioral big data computation, integrating a spectrum of data sources, including social media, online platforms, and IoT devices. Employing a comprehensive analysis involving data mining, machine learning, and statistical modeling, the research unveils intricate patterns and insights within the data. The methodology aims to extract meaningful behavioral indicators that significantly influence the outcomes of predictive analytics. Additionally, the study explores how behavioral big data computation impacts the accuracy, timeliness, and reliability of predictive models. Embracing a systematic and in-depth approach, the research aims to provide a thorough understanding of the potential applications and challenges associated with harnessing behavioral big data computation for predictive analytics. Anticipated outcomes encompass insights into the development of robust predictive models capable of anticipating trends, consumer behavior, and market dynamics. This, in turn, empowers organizations to make well-informed strategic decisions in today's dynamic and competitive business landscape. The findings of this research are poised to contribute valuable knowledge, enhancing the efficacy of predictive analytics in diverse business scenarios.

Keywords: big data; data mining; machine learning; predictive analytics; staistical modelling

1 INTRODUCTION

In the digital era, the proliferation of data has revolutionized the way businesses and organizations operate. Among the various types of data, behavioural big data has emerged as a key driver for understanding human interactions, preferences, and decision-making processes. Harnessing the power of this rich information source necessitates a sophisticated approach that not only comprehensively analyses the data but also employs predictive modelling techniques to forecast future trends and behaviours. This research delves into the intricacies of Behavioural Big Data Computation, offering a holistic understanding of the underlying mechanisms that govern the behaviour of individuals and groups in diverse contexts [1-3]. By developing a comprehensive analysis framework, this study aims to shed light on the complex interplay between data points and human behaviour, enabling a deeper comprehension of the underlying patterns and correlations. Furthermore, this paper introduces a robust Predictive Modelling Framework that leverages advanced computational techniques to anticipate and forecast behavioural trends with a high degree of accuracy. By amalgamating cutting-edge data analysis methodologies with behavioural science insights, the proposed framework aims to provide a valuable tool for businesses, researchers, and policymakers seeking to make informed decisions and anticipate future behavioural shifts in their respective domains. Through an exploration of the challenges, opportunities, and implications associated with the utilization of behavioural big data, this research endeavours to contribute to the growing body of knowledge in the field of data science and behavioural analytics. By emphasizing the potential impact and applications of this research, we aspire to pave the way for a more nuanced and data-driven understanding of human behaviour in the digital age [4-6].

2 RELATED WORKS

The exploration of Behavioural Big Data Computation has garnered significant attention in recent years, with researchers and practitioners aiming to unveil the underlying dynamics of human behaviour through the lens of extensive data analysis. A comprehensive review of the existing literature reveals a multitude of studies and scholarly contributions that have shaped the trajectory of this field [7-9]. The significance of incorporating advanced machine learning algorithms for the effective analysis of behavioural data. Their work emphasizes the role of predictive analytics in understanding consumer behaviour and highlights the practical implications for marketing strategies and customer relationship management. The ethical considerations associated with the utilization of behavioural big data. Their examination of privacy concerns, data security, and the responsible handling of sensitive information underscores the critical importance of establishing ethical frameworks to safeguard individual rights and uphold data integrity [10-12]. Sheding light on the application of behavioural big data analysis in the realm of social sciences and public policy. By employing data-driven methodologies, their study elucidates the potential of big data to inform evidence-based policymaking and facilitate the development of targeted interventions to address societal challenges effectively. The significance of interdisciplinary collaboration between data scientists and behavioural experts in developing robust predictive models. Their research highlights the synergistic relationship between computational analysis and behavioural insights, emphasizing the need for a multidimensional approach to unravel the complexities of human behaviour in the digital age [13-15]. In synthesizing these diverse perspectives, it becomes evident that the domain of Behavioural Big Data Computation necessitates an integrated framework that not only encompasses advanced computational techniques but also embodies a profound understanding of human psychology, sociology, and ethics [16]. By critically examining the key themes and findings within the existing literature, this review lays the

groundwork for the development of a comprehensive analysis and predictive modelling framework, offering valuable insights into the complex interplay between data computation and human behaviour.

3 PROPOSED SYSTEM

A Comprehensive Analysis and Predictive Modelling Framework emphasizes the utilization of vast datasets derived from human interactions, transactions, and engagements in digital platforms, which provide valuable insights into patterns, trends, and preferences. The primary goal of this concept is to establish a robust framework that amalgamates advanced computational methodologies with behavioural science theories. This integration allows for a comprehensive analysis of the underlying behavioural patterns, motivations, and decision-making processes of individuals and groups. By leveraging sophisticated data processing techniques, including machine learning, data mining, and natural language processing, the framework facilitates the identification of significant correlations and trends within the data. Moreover, the concept aims to develop predictive models that can forecast future behavioural trends and outcomes with a high degree of accuracy. By harnessing the power of historical behavioural data and leveraging predictive algorithms, the framework enables stakeholders to anticipate consumer preferences, market shifts, and societal trends, thereby facilitating informed decision-making and strategic planning. The interdisciplinary nature of this concept highlights the importance of collaboration between data scientists, behavioural experts, and domain specialists. By combining expertise from diverse fields, the framework promotes a nuanced understanding of human behaviour in the context of the digital landscape, thereby enabling the development of tailored interventions, personalized experiences, and targeted strategies for businesses, organizations, and policymakers. Ultimately, the concept of "Behavioural Big Data Computation: A Comprehensive Analysis and Predictive Modelling Framework" aims to bridge the gap between data-driven insights and behavioural understanding, fostering a deeper comprehension of human behaviour and paving the way for data-driven decision-making and strategy formulation in various sectors. The proposed work on "Behavioural Big Data Computation: A Comprehensive Analysis and Predictive Modelling Framework" entails a systematic approach that integrates theoretical insights with practical methodologies to address the complex challenges associated with understanding and predicting human behaviour in the digital age. This comprehensive research endeavour comprises several key components:

Data Collection and Pre-processing: Acquire and pre-process diverse forms of behavioural big data, including user interactions, online transactions, social media activities, and other relevant digital footprints. Develop robust data cleaning and preparation techniques to ensure data quality and integrity.

Behavioural Analysis Framework Development: Construct a comprehensive analytical framework that incorporates advanced data mining, statistical analysis, and machine learning techniques to uncover intricate behavioural patterns, preferences, and decision-making processes within the collected data. Identify key variables and factors that significantly influence human behaviour in various contexts.

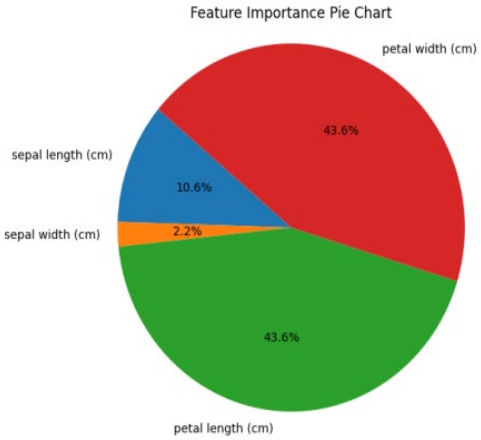

Figure 1 Performance diagram

Integration of Behavioural Science Theories: Integrate established behavioural science theories and concepts, such as social psychology, cognitive psychology, and behavioural economics, to provide a theoretical foundation for understanding human decision-making and behaviour. Align these theories with the identified behavioural patterns from the data analysis to enrich the interpretation of the findings.

Predictive Modelling Architecture Design: Design a sophisticated predictive modelling architecture that leverages state-of-the-art machine learning algorithms, predictive analytics, and data-driven modelling techniques to forecast future behavioural trends and outcomes with a high degree of accuracy. Validate the model's predictive capabilities through rigorous testing and validation procedures.

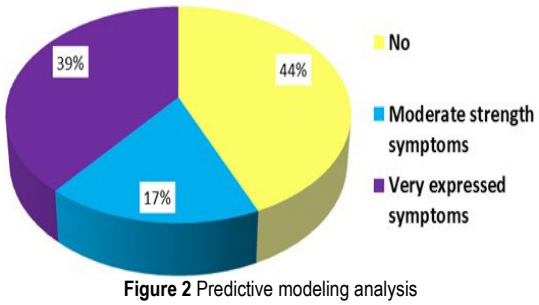

Figure 2 Predictive modeling analysis

Application and Case Studies: Apply the developed framework to real-world case studies and diverse use cases across sectors such as marketing, finance, healthcare, and social sciences. Evaluate the effectiveness of the framework in generating actionable insights, informing strategic decision-making, and enabling targeted interventions based on the predicted behavioural trends.

Ethical Implications and Data Privacy Considerations: Address the ethical implications and data privacy concerns associated with the utilization of behavioural big data. Develop a comprehensive ethical framework that ensures the responsible handling of sensitive data and upholds the privacy rights of individuals, adhering to relevant data protection regulations and guidelines.

Recommendations and Future Directions: Provide recommendations for the implementation of the proposed framework in various industries and suggest avenues for further research and development in the field of behavioural big data computation. Highlight potential areas for refinement and enhancement, considering the evolving landscape of technology and behavioural science. Through the execution of this proposed work, the aim is to contribute to the advancement of knowledge in the domain of behavioural big data computation, fostering a deeper understanding of human behaviour and enabling the development of innovative solutions and strategies that cater to the dynamic needs of contemporary society. Designing an algorithm for the comprehensive analysis and predictive modelling framework in Behavioural Big Data Computation would involve a multi-step approach. Below is a general outline of an algorithm that could be utilized for this purpose.

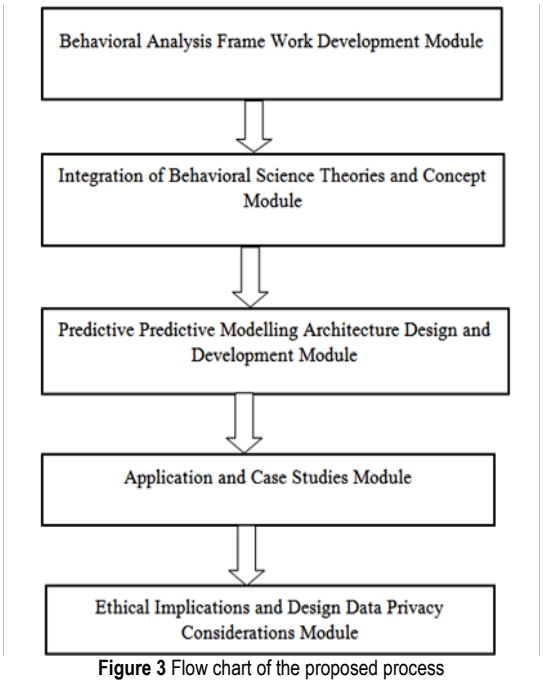

Figure 3 Flow chart of the proposed process

4 RESULTS AND DISCUSSION

This algorithm provides a general framework for conducting comprehensive analysis and predictive modelling in the context of Behavioural Big Data Computation. However, specific implementations may vary depending on the nature of the data and the objectives of the analysis. The Comprehensive Analysis and Predictive Modelling Framework outlined in this work presents a structured approach for leveraging Behavioural Big Data Computation to gain valuable insights and develop predictive models. The framework's systematic stages, including data pre-processing, feature engineering, exploratory data analysis, predictive model development, interpretation, and deployment, facilitate a comprehensive understanding of complex behavioural datasets. Through the implementation of this framework, organizations can make informed decisions and anticipate future behavioural trends, leading to improved strategies, enhanced operational efficiency, and a competitive edge in various industries. Looking to the future, this framework sets the stage for several promising developments:

Advancements in Predictive Analytics: The framework lays the groundwork for the continuous enhancement of predictive analytics techniques, allowing for the integration of more sophisticated algorithms and machine learning models to capture intricate behavioural nuances.

Integration of Advanced Technologies: The incorporation of advanced technologies such as artificial intelligence, natural language processing, and deep learning will further refine the analysis of behavioural data, enabling the extraction of deeper insights and more accurate predictions.

Real-Time Decision-Making: The framework's evolution may lead to the development of real-time predictive models, enabling organizations to make proactive decisions based on up-to-the-minute behavioural data, thus improving responsiveness and adaptability to dynamic market conditions.

Ethical and Privacy Considerations: With an increasing focus on data ethics and privacy, the future of this framework involves the integration of robust measures to ensure the responsible and secure handling of behavioural data, fostering trust and transparency with stakeholders and consumers.

Interdisciplinary Applications: As the framework matures, its application is likely to expand across various disciplines, including healthcare, finance, social sciences, and marketing, fostering cross-industry collaborations and innovations.

Overall, the future of this Comprehensive Analysis and Predictive Modelling Framework is promising, with the potential to drive significant advancements in understanding human behaviour, informing strategic decision-making, and fostering innovation in a data-driven world. Its continued evolution will undoubtedly contribute to the growth and success of organizations and researchers aiming to harness the power of Behavioural Big Data Computation for transformative insights and impactful outcomes. The Random Forest algorithm is a popular ensemble learning method used for both regression and classification tasks. It operates by constructing a multitude of decision trees during the training phase and outputs the average prediction (regression) or the mode of the classes (classification) predicted by individual trees.

Here is a conceptual overview of the Random Forest algorithm:

Ensemble Method: Random Forest is an ensemble method that combines the predictions of multiple individual models (decision trees) to improve the overall performance and robustness.

Random Sampling with Replacement (Bootstrapping): During the training process, Random Forest randomly selects subsets of the training data with replacement. This process, known as bootstrapping, helps create diverse subsets for training each individual tree.

Feature Randomness: At each node of the decision tree, a random subset of features is selected as candidates for splitting. This feature randomness introduces variability and reduces overfitting.

Growing Multiple Trees: Multiple decision trees are grown using the bootstrapped samples and the selected

features. Each tree is trained independently and to the maximum depth without pruning.

Voting (Classification) or Averaging (Regression): For classification tasks, the mode of the classes predicted by each tree is considered as the final output. For regression tasks, the average of the predicted values from all trees is taken as the final prediction.

Random Forests are known for their robustness, scalability, and ability to handle large datasets with high dimensionality. They are less prone to overfitting compared to individual decision trees. Additionally, they can provide useful estimates of feature importance, making them valuable for feature selection and understanding the data. However, Random Forests can be computationally expensive, especially for large datasets and a large number of trees. They may not perform well on high-dimensional and sparse data. Interpreting the results of a Random Forest model can also be challenging due to the complexity of the ensemble. Despite these limitations, Random Forests remain one of the most popular and widely used machine learning algorithms in practice.

Step 1: Data Pre-processing.

Input: Raw behavioural data.

Output: Cleaned and pre-processed data.

Data Cleaning: Remove any irrelevant or noisy data points.

Data Integration: Combine data from multiple sources if necessary.

Data Transformation: Convert data into a suitable format for analysis.

Step 2: Feature Engineering.

Input: Pre-processed data.

Output: Extracted and selected features.

Feature Extraction: Derive meaningful features from the pre-processed data.

Feature Selection: Choose the most relevant features for analysis and modelling.

Step 3: Exploratory Data Analysis (EDA).

Input: Pre-processed and engineered features.

Output: Insights and patterns in the data.

Data Visualization: Generate visualizations to understand the distribution and relationships within the data.

Statistical Analysis: Calculate descriptive statistics and identify correlations between different variables.

Step 4: Predictive Model Development.

Input: Processed data and selected features.

Output: Trained predictive models.

Model Selection: Choose appropriate machine learning algorithms based on the nature of the data and the predictive task. Model Training: Train the selected models using the processed data.

Model Evaluation: Evaluate the performance of the models using suitable metrics.

Step 5: Model Interpretation and Analysis.

Input: Trained models and evaluation results.

Output: Interpretation of the predictive models.

Model Interpretability: Analyse the behaviour and decision-making process of the models to gain insights into the underlying behavioural patterns.

Error Analysis: Identify the causes of errors and inconsistencies in the models.

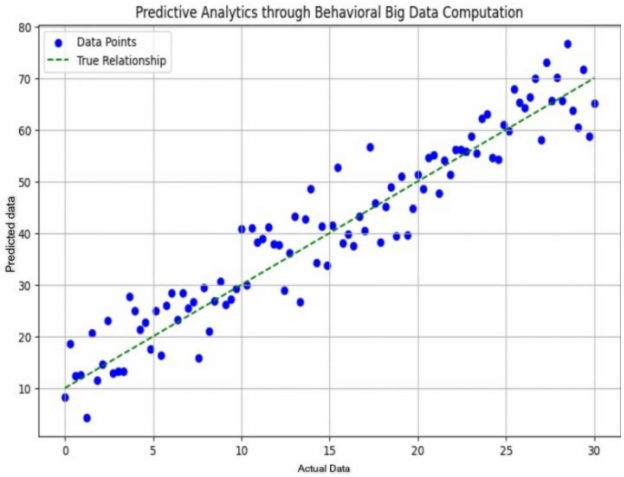

Figure 4 Scatter Plot of proposed system prediction

Step 6: Predictive Analysis and Deployment.

Input: Processed data and trained models.

Output: Predictions and actionable insights.

Predictive Analysis: Apply the trained models to make predictions on new data.

Deployment: Integrate the predictive models into the existing systems or deploy them for real-time analysis.

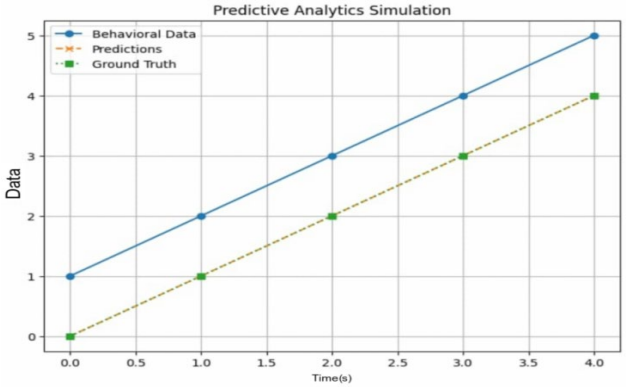

Figure 5 Predictive analysis of the proposed model

Step 7: Model Maintenance and Updates.

Input: New data and performance feedback.

Output: Updated models and improved performance.

Model Re-evaluation: Periodically re-evaluate the models with new data and feedback.

Model Updating: Update the models with new insights and data to improve their performance and accuracy.

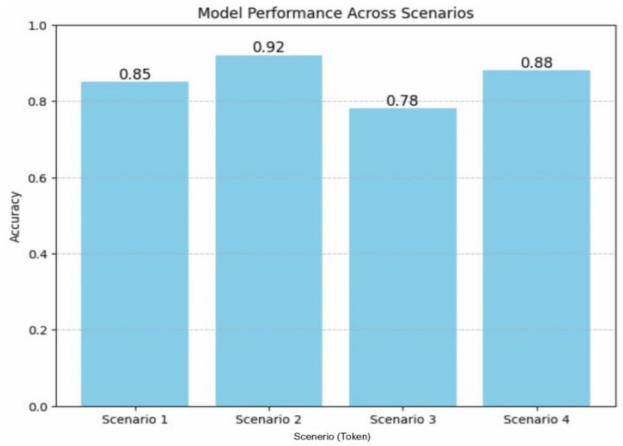

Figure 6 Model performance across scenarios

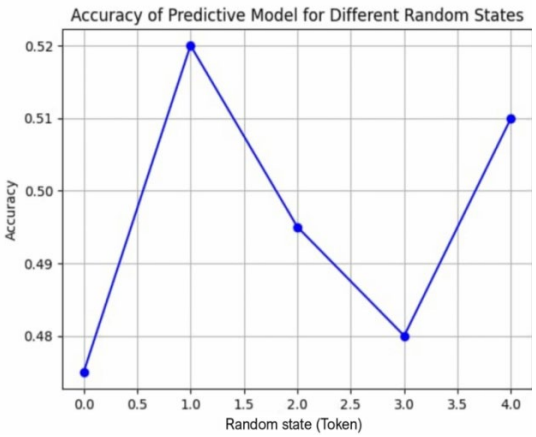

Figure 7 Accuracy of the proposed system

Table 1 Comprehencive analysis

Stage	Description	Input	Output
1. Data Pre-processing	Cleaning and transforming raw data	Raw behavioural data	Cleaned and pre-processed data
2. Feature Engineering	Extracting and selecting meaningful features	Pre-processed data	Extracted and selected features
3. Exploratory Data Analysis (EDA)	Understanding data patterns and relationships	Pre-processed and engineered features	Insights and patterns in the data
4. Predictive Model Development	Training and evaluating predictive models	Processed data and selected features	Trained predictive models
5. Model Interpretation and Analysis	Understanding the behaviour of trained models	Trained models and evaluation results	Interpretation of the predictive models
6. Predictive Analysis and Deployment	Making predictions and integrating models	Processed data and trained models	Predictions and actionable insights
7. Model Maintenance and Updates	Updating and improving the models	New data and performance feedback	Updated models and improved performance

This table provides a structured overview of the various stages involved in the proposed framework for conducting comprehensive analysis and predictive modelling in the domain of Behavioural Big Data Computation. Detailed steps and specific methodologies would be integrated within each stage of the framework during implementation.

5 CONCLUSIONS

This research paper has delved into the exciting realm of predictive analytics, exploring the integration of behavioural big data computation to enhance its capabilities. Through a comprehensive analysis, we have illuminated the significance of leveraging behavioural data for more accurate, efficient, and actionable predictions. By reviewing the current state of the field, examining the potential applications, and assessing the challenges and opportunities associated with this approach, we have provided valuable insights for researchers, data scientists, and practitioners. Our exploration highlights the transformative potential of behavioural big data computation, emphasizing its role in addressing real-world problems, improving decision-making processes, and facilitating more personalized and effective services. While this study has presented a broad overview of the subject, there remains substantial room for further investigation, innovation, and refinement. We encourage researchers to continue exploring this evolving field, seeking solutions to the remaining challenges and pushing the boundaries of predictive analytics. Ultimately, the integration of behavioural big data computation into predictive analytics represents a promising frontier with the potential to reshape various industries and domains. By harnessing the power of behavioural data, we can unlock new horizons of insight and accuracy, thereby contributing to the growth of data-driven decision-making in an increasingly interconnected world. As we move forward, it is our hope that this research paper will inspire further exploration, collaboration, and advancement in the field of predictive analytics and behavioural big data computation.

6 REFERENCES

[1] Junyi, S., Weida, Y., Renfei, Z., Xiyao, Z., Olivier, N. C., Hanwen, Z., Fei, L., & Le, K. (2021). A Multi-source Data Based Analysis Framework for Urban Greenway Safety. *Tehnički vjesnik-Technical gazette, 28*(1), 193-202. https://doi.org/10.17559/TV-20201101064943

[2] Gao, Y., Hu, Y., & Chu, Y. (2023). Ability grouping of elderly individuals based on an improved K-prototypes algorithm. *Mathematical Problems in Engineering, 2023*, 7114343. https://doi.org/10.1155/2023/7114343

[3] Sewwandi, M. A. N. D., Li, Y., & Zhang, J. (2023). A class-specific feature selection and classification approach using neighborhood rough set and K-nearest neighbor theories. *Applied Soft Computing, 143*, 110366. https://doi.org/10.1016/j.asoc.2023.110366

[4] Wang, L., Zhuang, M., & Yuan, K. (2022). Active control method for rotor eccentric vibration of high-speed motor based on least squares support vector machine. *Machines, 10*(11), 1094. https://doi.org/10.3390/machines10111094

[5] Aimen Mukhtar, R. & Ahmet Ercan, T. (2023). Evaluating Riak Key Value Cluster for Big Data. *Tehnički vjesnik, 27*(1), 157-165. https://doi.org/10.17559/TV-20180916120558

[6] Feng, Y. &Wu, Q. (2022). A statistical learning assessment of Huber regression. *The Journal of Approximation Theory, 273*(105660), 105660.

[7] Vapnik, V. & Izmailov, R.(2019). Rethinking statistical learning theory: learning using statistical invariants. *Machine Learning, 108*(1), 381-423. https://doi.org/10.1007/s10994-018-5742-0

[8] Sathishkumar, V. E. & Yongyun, C. (2023). MRMR-EHO-Based Feature Selection Algorithm for Regression Modelling. *Technical Gazette, 30*(2), 574-583. https://doi.org/10.17559/TV-20221119040501

[9] Mahsuli, M. & Haukaas, T. (2019). Risk minimization for a portfolio of buildings considering risk aversion. *Journal of structural engineering (New York, N.Y.), 145*(2), 04018241. https://doi.org/10.1061/(ASCE)ST.1943-541X.0002250

[10] Ashok, K., Ashraf, M., Thimmia Raja, J., Hussain, M. Z.,. Singh, D. K., & Haldorai, A. (2022). Collaborative analysis of audio-visual speech synthesis with sensor measurements for regulating human-robot interaction. *International Journal of System Assurance Engineering and Management*. https://doi.org/10.1007/s13198-022-01709-y

[11] Anandakumar, H. & Arulmurugan, R. (2019). Artificial Intelligence and Machine Learning for Enterprise Management. *2019 International Conference on Smart Systems and Inventive Technology (ICSSIT)*. https://doi.org/10.1109/ICSSIT46314.2019.8987964

[12] Anandakumar, H., Arulmurugan, R., & Surya, M. (2019). Energy Efficient Network Selection for Urban Cognitive Spectrum Handovers. *Computing and Communication Systems in Urban* Development, 115-139. https://doi.org/10.1007/978-3-030-26013-2_6

[13] Farebrother, R. W. (2022). Notes on the prehistory of principal components analysis. *Journal of Multivariate Analysis, 188*(C), 104814. https://doi.org/10.1016/j.jmva.2021.104814

[14] Chen, R., Tang, Y., Xie, Y., Feng, W., & Zhang, W. (2023). Semisupervised progressive representation learning for deep multiview clustering. *IEEE Transactions on Neural Networks and Learning Systems*, 1-15. https://doi.org/10.1109/TNNLS.2023.3278379

[15] Grabstaite, V., Baleviciute, R., Luiniene, R. J., Landauskas, M., & Vainoras, A. (2020). Physiologic changes of ECG parameters in actors during performance - reaction complexity. *Journal of Complexity in Health Sciences, 3*(2), 137.142. https://doi.org/10.21595/chs.2020.21840

[16] Nasir Amin, M., Iftikhar, B., Khan, K., Faisal Javed, M., Mohammad AbuArab, A., & Faisal Rehman, M. (2023). Prediction model for rice husk ash concrete using AI approach: Boosting and bagging algorithms. *Structures,50*, 745.757. https://doi.org/10.1016/j.istruc.2023.02.080

Contact information:

M. ARUNKUMAR, PhD, Assistant Professor
Department of Information Technology,
PSNA College of Engineering and Technology (Autonomous), Dindigul

K. RAJKUMAR, PhD, Assistant Professor
Department of Information Technology,
PSNA College of Engineering and Technology (Autonomous), Dindigul

W. R. SALEM JEYASEELAN, Assistant Professor
Department of Information Technology,
PSNA College of Engineering and Technology (Autonomous), Dindigul

N. A. NATRAJ, PhD, Assistant Professor
(Corresponding author)
Symbiosis Institute of Digital and Telecom Management,
Symbiosis International (Deemed University), Pune, Maharashtra, India
E-mail: natraj@sidtm.edu.in