

# Research Progression from Data Preprocessing to Privacy-Preserving Information Systems

<sup>1</sup>Kailash Chandra Totla and <sup>2</sup>Sanjay Gour

<sup>1</sup> M L V Textile & Engineering College, Bhilwara, India Country

<sup>2</sup>Jaipur Engineering College and Research Centre, Jaipur, India

Email - <sup>1</sup>kctotla@gmail.com, <sup>2</sup>sanjay.since@gmail.com

**Abstract:** The fast progression of digital technologies has increased the necessity for protected information distribution and high-quality data administration. This review observes explore helps in privacy-preserving schemes, missing value approximation, anomaly detection along with the cloud-enabled information systems. It highlights the development of Visual Cryptography Scheme (VCS), DWT-SVD watermarking, Zernike Moment-based authentication along with the algorithms of Closest Fit approach for missing value repossession. The reviewed revisions validate significant enhancements in the image security, data integrity and preprocessing methods whereas addressing issues modelled by Rotation, Scaling, and Translation (RST) attacks. The evaluation also recognizes research gaps and highlights future occasions including artificial intelligence, blockchain, cloud computing and intelligent cybersecurity frameworks.

**Key Words:** Closest Fit, Missing Values, privacy, Preserving, Cloud, Cryptography, Computing.

## 1. INTRODUCTION:

The reviewed publications jointly characterize an advanced research journey straddling three key domains including privacy-preserving schemes, missing value approximation and irregularity discovery in data mining and cloud-enabled information systems. Though these studies resolve dissimilar research issues but actually these are strongly interrelated over the mutual objective of cultivating the quality, security and reliability of digital information schemes. The research advances on or after statistical data pre-processing methods to intelligent and smart security frameworks proficient of handling multifaceted multimedia and cyber-security issues.

The present study gives an amalgam of the basic technologies and their progression along with the correlation. These correlation also supports the new technology of the diverse field. This is beauty of the computer science technologies to produce mix approach.

## 2. REVIEW OF LITERATURE:

Research is simultaneously running correlated field; each one is assisting to one another. This study also gives the similar framework. Buck (1960) gives pioneer method to manage the missing values in the dataset [1], Chen L et al. (2005) suggest multiple imputation to handle the data [2], Sanjay Gaur et al. continue these studied with the help of various statistical and numerical approach, they produced better results in the domain, develop the closest fit approach [4,10,12,24,26], Pandya et al. works for detection of anomalies values and inliers detection and recovered missing value in data mining [3,22], Grzymala-Busse, J.W. (2004) studied data with missing attribute values and generalization of indiscernibility relation and rules induction with rough sets [5], Kim (1977) and Qin (2007) both working on the same and develop methods for imputations [6,8,9].

Kailash Totla et al (2016) gives a complete view of development of the ERP with critical success factors [7]. Sharma et al. (2013) and Sanjay G (2012,2014 and 2016) gives various algorithms to deal data-preparation in the efficient manner gradually [11, 20,21,23]. Zhang, S et al. (2003) discussed artificial intelligence utilization in the various existing approaches [12]. S D Degadwala et al. (2017-2018) discussed development in the privacy preserving system using Pseudo Zernike Moment with SURF & Affine Transformation on RST Attacks, also they discussed two way and 4-Share VCS Based Image Watermarking for Dual RST Attacks. Computational Vision and Bio Inspired Computing [14,15,16]. Solanki, N et al. (2018) discussed about comparative analysis of wavelet Families for Invisible Image embedding [18] and Seikh et al. (2019) discussed performance evaluation of cryptographic algorithm for emerging trends in expert applications and security [17]. Degadwala et al. (2016) carried out study of privacy preserving system based on progressive VCS and RST attacks [19]. Bhawesh et al. (2018): study on cloud technologies-research issues, challenges, engineering, platforms and applications.

### 3. DATA DRE-PROCESSING STUDY:

The initial works by Gaur and Dulawat attentive on the statistical features of data mining. The research studies entitle as A Perception of Statistical Inference in Data Mining, Univariate Analysis for Data Preparation in Context of Missing Values, and Improved Closest Fit Techniques to Handle Missing Attribute Values established and well-recognised a hypothetical foundation for the treatment of incomplete and imperfect datasets. These research investigations highlighted that inappropriate handling of missing values acquaint with bias, decreases prediction accuracy and destructively affects the knowledge discovery. The statistical inference along with the descriptive analysis were consequently recognized as important preprocessing stages previously to applying the data mining and machine learning algorithms.

Structuring upon this groundwork, Closest Fit-based algorithms developed the central and essential research progression theme in the domain. The research paper A Closest Fit Approach to Missing Attribute Values in Data Mining make known to the similarity-based approximation, representing significant development over conventional mean, median and regression-enabled imputation methods. This research methodology was further advanced by refining through the segmented Closest Fit Approach to Handle Missing Data, which enhanced approximation by separating datasets into homogeneous clusters beforehand to the imputation. After that novel concept, Closest Fit Approach through Linear Interpolation to Recover Missing Values combined interpolation approaches, encompassing the applicability of the algorithm to utilise and make continuous numerical datasets. Communally, these studies specify a strong and vibrant development from elementary statistical approximation toward adaptive and context-aware missing value recovery procedures and mechanisms.

### 4. ANOMALY DETECTION:

Similarly, a parallel contour of research addressed anomaly detection, identifying such type of missing information and anomalous explanations recurrently coexist in real-world datasets. The publications like Detection of Anomalous Value in Data Mining, Closest Fit Approach for Pattern Designing to Recover Anomalous Values, Inliers Detection and Recovered Missing Value in Data Mining, and Handling of Missing and Anomalous Data through Inference in Data Mining long-drawn-out the Closest Fit methodology elsewhere modest imputation. These studies united statistical inference along with the anomaly detection, representing those recognizing irregular observations before approximating missing values considerably recovers complete data quality. Consequently, anomaly detection research study directly matches with the missing value estimation, founding a wide-ranging data preprocessing background.

### 5. SECURITY AND PRIVACY PRESERVING:

While data mining research focused on data quality, alternative significant research study direction engrossed on privacy-preserving image security. Degadwala and Gaur introduced this province through Two-Way Privacy Preserving System Using Combine Approach: QR-Code and Visual Cryptography Scheme (VCS). This research work shared QR codes along with the Visual Cryptography to make available secure image communication without exposing and revealing confidential information in separate shares. The incorporation of QR codes enhanced authentication even though maintaining computational easiness.

Succeeding studies significantly heightened this framework by integrating transform-domain watermarking techniques in the domain. An Efficient Privacy Preserving System Using VCS, Block DWT-SVD and Modified Zernike Moment on RST Attacks introduced a hybrid security architecture that combined Visual Cryptography with Discrete Wavelet Transform (DWT), Singular Value Decomposition (SVD), and Modified Zernike Moments. Compared with conventional VCS methods, this method established more robustness in contradiction of Rotation, Scaling and Translation (RST) attacks whereas conserving quality of image. The utilization of transform-domain watermarking exemplifies the development from modest encryption procedures in the direction of intelligent multimedia authentication schemes.

Further developments were familiarized over Privacy Preserving System Using Pseudo Zernike Moment with SURF and Affine Transformation on RST Attacks, in this study the invariant image descriptors and feature-based harmonization enhanced watermark retrieval under severe geometric misrepresentations. Likewise, 4-Share VCS Based Image Watermarking for Dual RST Attacks augmented system safety and security by issuing watermark information amid manifold shares, dropping vulnerability to unauthorized rebuilding. The next research study A Study of Privacy Preserving System Based on Progressive VCS and RST Attacks prolonged this work by empowering advanced image reconstruction which representing and depicting the continuous progression in visual cryptographic methods.

An imperative subsidiary contribution derives from A Comparative Analysis of Wavelet Families for Invisible Image Embedding. Rather than suggesting a new-fangled watermarking algorithm, this research study methodically assessed diverse wavelet relations to regulate their inspiration on watermark strength and imperceptibility. Findings of this study deliver practical direction for selecting appropriate wavelet transforms in following watermarking schemes, by this means of solidification the efficiency of the hybrid VCS-DWT-SVD frameworks.

## 6. CRYPTOGRAPHY:

The studied literature also validates expressive and meaningful incorporation amid security research and cryptographic performance examination. A Study on Performance Evaluation of Cryptographic Algorithm delivers proportional insights towards encryption algorithms grounded on the execution time, computational complexity along with the security strength. These discoveries sustenance the development of competent privacy-preserving schemes by recognizing cryptographic approaches appropriate for resource-constrained atmospheres like as cloud and IoT platforms as future prospectives.

## 7. CLOUD AND ERP:

Elsewhere safety, security and data pre-processing, the studied publications examine that the enterprise and cloud computing surroundings is the bog domain where protected information supervision becomes progressively significant. A Study of Cloud Technologies Research Issues, Challenges, Engineering, Platforms and Applications focused on the issues together with scalability, privacy, virtualization and resource allocation. These issues in the form of challenges thoroughly align with the purposes of privacy-preserving image safety and secure data administration, signifying that cloud computing offers a suitable deployment atmosphere for numerous of the projected security frameworks. Likewise, An Insight of Critical Success Factors for ERP Model swings the focus in the direction of organizational information arrangements. Even though this work fluctuates from the image processing, security and data mining, it strengthens the standing of data integrity, system reliability, user acceptance and secure information distribution. Subsequently, ERP research accompaniments the wider objective of emerging responsible information systems.

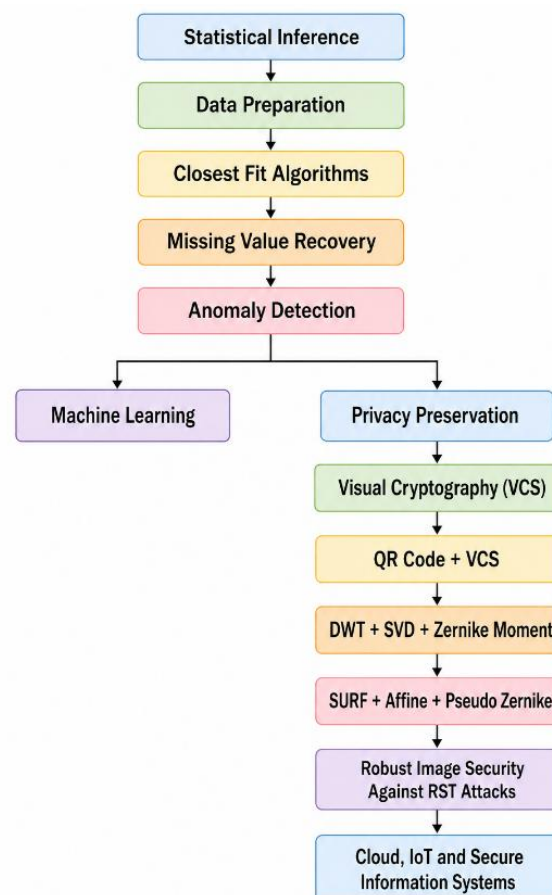


Figure 1: The progression and correlations amid technologies

The reviewed and studied publications disclose an intelligible research development. Initial statistical studies recognized with reliable preprocessing methodologies for imperfect datasets. These methodologies progressed into progressive Closest Fit algorithms proficient of concurrently resolving missing and anomalous data and values. In corresponding, privacy-preserving research study advanced from traditional Visual Cryptography to cultured hybrid outlines and frameworks integrating DWT, SVD, SURF, QR codes and Zernike Moments for strong multimedia verification. The cloud computing and ERP studies offer the positioning and application background inside which these algorithms function. From a methodological viewpoint, the correlation amid these research studies validates a gradual changeover

from statistical techniques to hybrid computational intelligence methods. The statistical inference provisions data quality improvement; Closest Fit algorithms will play enhances role for machine learning performance over intelligent imputation. The visual cryptography confirms privacy, confidentiality along with the transform-domain watermarking which guarantees genuineness. The cloud computing empowers scalable employment of various technologies. Jointly, we observed that these contributions establish a wide-ranging framework for protected, intelligent and trustworthy information processing.

## 8. CORRELATION MATRIX OF THE REVIEWED PAPERS:

From the correlation table it is observed that data preprocessing concept is too much important for every technology as to prepare the ground. Further the security and cryptography get improved with pre-processed data it would be numeric as well as images. The proper managed dataset and database are greatly and efficiently useful for cloud facilities and ERP systems.

| Research Theme               | Key Papers                                                                     | Major Contribution                           | Relationship with Other Studies                        |
|------------------------------|--------------------------------------------------------------------------------|----------------------------------------------|--------------------------------------------------------|
| Statistical Data Preparation | Gaur & Dulawat (Statistical Inference, Univariate Analysis)                    | Foundation for preprocessing                 | Supports missing value estimation algorithms           |
| Missing Value Estimation     | Closest Fit, Improved Closest Fit, Segmented Closest Fit, Linear Interpolation | Intelligent imputation techniques            | Basis for anomaly recovery and machine learning        |
| Anomaly Detection            | Detection of Anomalous Value, Inliers Detection, Pattern Designing             | Detection and correction of abnormal records | Complements missing value estimation                   |
| Visual Cryptography          | QR-Code & VCS, Progressive VCS                                                 | Secure image sharing                         | Foundation for hybrid privacy-preserving systems       |
| Hybrid Image Security        | VCS + DWT-SVD + Zernike, SURF + Pseudo Zernike, 4-Share VCS                    | Robust watermarking against RST attacks      | Extends traditional VCS with transform-domain security |
| Image Watermarking           | Wavelet Family Analysis                                                        | Selection of optimal wavelet transform       | Supports DWT-based watermark embedding                 |
| Cryptography                 | Performance Evaluation of Cryptographic Algorithms                             | Encryption efficiency comparison             | Guides secure privacy-preserving implementations       |
| Cloud & ERP                  | Cloud Technologies Survey, ERP Critical Success Factors                        | Secure deployment environments               | Application domain for data security and privacy       |

Table 1: Correlation of technologies for synchronizations

## 9. CONCLUSION:

The reviewed research studies cooperatively develop the privacy preservation, secure image validation on the basis of quality data, The quality data utilised in the enhancement of the innovative cryptographic and data mining methods. Their combination delivers a robust foundation for developing smart, intelligent, secure and scalable information arrangements. Forthcoming research should emphasis on AI-ML driven, blockchain-enabled and cloud-IoT-based privacy-preserving resolutions.

## REFERENCES:

1. Buck, S.F. (1960): A method of estimation of missing values in multivariate data suitable for use with an electronic computer, J. Royal Statistical Society, Series B, 2, 302-306.
2. Chen, L., Drane, M.T., Valois, R.F., and Drane, J.W. (2005): Multiple imputation for missing ordinal data, Journal of Modern Applied Statistical Methods, 4(1), 288-299.
3. D D Pandya and Sanjay Gaur, (2018): Detection of Anomalous value in Data Mining, Kalpa Publications in Engineering, 2, 1-6.
4. Gaur, Sanjay and Dulawat, M.S. (2010): Univariate Analysis for Data Preparation in context of Missing Values, Journal of Computer and Mathematical Sciences, 1(5), 628-635.
5. Grzymala-Busse, J.W. (2004): Data with missing attribute values: Generalization of in-discernibility relation and rules induction, Transactions of Rough Sets, Lecture Notes in Computer Science Journal Subline, Springer-Verlag, 1,8-95.
6. Kim, J.O., and Curry, J. (1977): The treatment of missing data in multivariate analysis, Social Methods and Research, 6, 215-240.
7. Kailash Totla, Manju Mandot and Sanjay Gaur (2016): An Insight of Critical Success Factors for ERP Model, International Journal of Emerging Research in Management & Technology, 5(2), 90-92.

8. Qin, Y.S. (2007): Semi-parametric optimization for missing data imputation, *Applied Intelligence*, 27(1),79-88.
9. Rubin, D.B., Inference and missing data, *Biometrika*, 63, 581-592.
10. Sanjay Gaur and DD Pandya (2018): Closest fit approach for pattern designing to recovered anomalous values in data mining, 2018 Second world conference on smart trends in system, security and sustainability (worksS4), 308-312, IEEE.
11. Sharma, Swati and Gaur, Sanjay (2013): Contiguous Agile Approach to Manage Odd Size Missing Block in Data Mining”, *International Journal of Advanced Research In Computer Science*, 4(11), 214-217.
12. S. Gaur and M.S. Dulawat (2010): A perception of statistical inference in data mining, *International Journal of Computer Science and Communication*, 1(2), 653-658.
13. Zhang, S., Zhang, C. and Young, Q. (2003): Data preparation for data mining, *Applied Artificial Intelligence*, 17, 375-381.
14. Sheshang D. Degadwala, Sanjay Gaur (2017): Privacy Preserving System Using Pseudo Zernike Moment with SURF & Affine Transformation on RST Attacks, *International Journal of Computer Science and Information Security (IJCSIS)*, 15(4), 147-152.
15. S. D. Degadwala and S. Gaur (2017): Two-way privacy preserving system using combine approach: QR-code & VCS, 2017 Innovations in Power and Advanced Computing Technologies (i-PACT), Vellore, India, 2017, pp. 1-5, doi: 10.1109/IPACT.2017.8245120.
16. Degadwala, S.D., Gaur, S. (2018): 4-Share VCS Based Image Watermarking for Dual RST Attacks. *Computational Vision and Bio Inspired Computing, Lecture Notes in Computational Vision and Biomechanics*, vol 28. Springer, Cham. [https://doi.org/10.1007/978-3-319-71767-8\\_77](https://doi.org/10.1007/978-3-319-71767-8_77)
17. Sheikh, M.F.A., Gaur, S., Desai, H., Sharma, S.K. (2019): A Study on Performance Evaluation of Cryptographic Algorithm. *Emerging Trends in Expert Applications and Security. Advances in Intelligent Systems and Computing*, vol 841. Springer, Singapore. [https://doi.org/10.1007/978-981-13-2285-3\\_44](https://doi.org/10.1007/978-981-13-2285-3_44)
18. Solanki, N., Khandelwal, S., Gaur, S., Gautam, D. (2018): A Comparative Analysis of Wavelet Families for Invisible Image Embedding, *Proceedings of ICE-TEAS 2018. Emerging Trends in Expert Applications and Security. Advances in Intelligent Systems and Computing*, vol 841. 219-227, Springer, Singapore. [https://doi.org/10.1007/978-981-13-2285-3\\_27](https://doi.org/10.1007/978-981-13-2285-3_27)
19. S. D. Degadwala and S. Gaur (2016): A study of privacy preserving system based on progressive VCS and RST attacks, 2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication (ICGTSPICC), Jalgaon, India, 2016, pp. 138-142, doi: 10.1109/ICGTSPICC.2016.7955285.
20. Sanjay Gaur (2014): Estimation of Missing Value at Extremes in Data Mining, *International Journal Advance Foundation and Research in Computer*, 1(3), 13-19.
21. Sanjay Gaur (2012) Closest fit approach to handle odd size missing block values, *International Journal Mathematical Achieve*, 3 (7), 2594-2597.
22. Pandya, D. D., & Gaur, D. S. Inliers Detection and Recovered Missing value in Data Mining. *International Journal of Emerging Technology and Advanced Engineering*, 8, 1-6.
23. Sanjay, G. (2016). Handling of missing and anomalous data through Inference in Data Mining. *International Transactions in Mathematical Sciences and Computers*, 9(1-2), 1-7.
24. Gaur Sanjay and Dulawat, M. S. (2011): Improved closest fit techniques to handle missing attribute values, *Journal of Computer and Mathematical Sciences*, 2(2), 384-390.
25. Bhawesh Kumawat, Sanjay Choudhary and Sanjay Gaur (2018): A Study of Cloud Technologies-Research Issues, Challenges, Engineering, Platforms and Applications: A Survey, *International Journal of Multimedia, Image Processing and Pattern Recognition, Eureka Journals 2018* 1(1), 1-10. ISSN: 2581-625X (online)
26. Gaur Sanjay and Dulawat, M. S. (2011): A closest fit approach to missing attributes values in data mining”, *International Journal of Advance in Science and Technology*, 2(4), 18-23. ISSN 2229-5216.