10.48047/jocaaa.2025.34.05.47

Hybrid PAACDA Framework With Random Forests for Data Corruption Detection in Complex Systems

Mohammad Sayyed Pasha[1*], Nagengar Kumar Gupta[2], G.Srikar[2], B.Prabath[2], G.Preetham[2]

[1]Assistant Professor, [2]UG Student, [1,2]Department of Information Technology,
[1,2]Malla Reddy Engineering College and Management Sciences, Kistapur, Medchal, 501401, Telangana, India
*Corresponding author: mdsayeed.it@mrem.ac.in

Received: 02.03.2025 Revised: 15.04.2025 Accepted: 25.05.2025

Abstract

In the era of big data, maintaining data integrity is crucial, as even minor data corruption can significantly distort analytical outcomes in domains such as e-commerce behavior prediction, social network analysis, and intrusion detection systems. Manual inspection and traditional anomaly detection techniques often fail to identify subtle corruption in high-dimensional datasets and lack adaptability to unseen data, requiring frequent retraining. Studies reveal that nearly 10–15% of large-scale transactional and behavioral datasets are affected by noise or corruption, while conventional unsupervised methods such as Local Outlier Factor, Isolation Forest, and One-Class SVM typically achieve only 85–92% detection accuracy. To address these limitations, this research proposes the PAACDA (Proximity-based Adamic–Adar Corruption Detection Algorithm), a graph-theoretic approach that leverages the Adamic–Adar similarity index to identify anomalous data points based on local proximity deviations. A PAACDA index is computed for each instance using column-wise deviation measures and adaptive thresholding to flag corrupted data. Furthermore, a Hybrid PAACDA framework is introduced by integrating PAACDA-derived features with a Random Forest classifier, enabling predictive detection of corruption in future datasets. Experimental results demonstrate detection accuracies of 99–100%, providing a proactive, scalable, and highly reliable solution for data quality assurance in complex and dynamic systems.

Keywords: Data Corruption Detection, Data Integrity, PAACDA Algorithm, Adamic–Adar Similarity, Random Forest, Hybrid Machine Learning, Anomaly Detection, Big Data Analytics, Predictive Data Quality

1. INTRODUCTION

In today's digital economy, data has become an indispensable asset across diverse sectors such as finance, healthcare, e-commerce, cybersecurity, and social networking platforms. According to IBM, over 2.5 quintillion bytes of data are generated daily, with more than 80% of this data being unstructured or semi-structured, often vulnerable to inconsistency, corruption, or noise. As data-driven decision-making becomes increasingly prominent, the quality of data holds immense significance. Any discrepancies in the form of corrupted or outlier values can severely compromise analytical accuracy, leading to flawed predictions and erroneous business strategies.

Traditional data analysis methods have been extensively used to mine patterns, detect anomalies, and improve customer or system performance. However, a 2022 report by Gartner highlighted that nearly 27% of organizational datasets contain data quality issues, which can stem from system-level failures, integration mismatches, or malicious attacks. Most detection mechanisms, especially rule-based systems or standard statistical techniques, struggle to keep up with the growing complexity and volume of modern datasets, particularly when corruptions occur subtly within expected ranges.

10.48047/jocaaa.2025.34.05.47

Moreover, recent studies in data science underscore the inadequacy of legacy systems to perform precise outlier detection in dynamic environments. The anomaly detection methods such as K-Means, DBSCAN, One-Class SVM, and LOF, though effective to a degree, often fall short when applied to real-time or high-dimensional datasets. These approaches rely heavily on assumptions about the data's distribution or structure, making them unsuitable for evolving datasets common in e-commerce or online services. Therefore, the need for robust data corruption detection strategies is more vital than ever in the modern data landscape.

2. LITERATURE SURVEY

In [1], Burgdorf explored the effects of data corruption on Cyber-Physical Systems (CPS). The study developed a predictive model to assess how corrupted data influenced system reliability and performance. It employed simulation-based techniques to quantify operational disruptions. The approach integrated statistical analysis to identify critical failure points. The work provided a framework for mitigating risks in CPS. The model relied heavily on simulated data, limiting its applicability to real-world scenarios with unpredictable corruption patterns.

In [2], Chandola conducted a comprehensive review of anomaly detection techniques for identifying unusual patterns in data. The study categorized methods into statistical, machine learning, and hybrid approaches. It analyzed their suitability for various domains, including cybersecurity and finance. The work established a taxonomy for anomaly detection algorithms. The survey emphasized computational complexity and scalability. The broad scope lacked in-depth analysis of domain-specific challenges, reducing its practical utility for specialized applications.

In [3], Pang-Ning introduced foundational concepts of data mining for extracting patterns from large datasets. The book detailed clustering, classification, and association rule mining techniques. It emphasized preprocessing steps to ensure data quality. The work provided practical examples for implementing algorithms. It served as a reference for beginners and practitioners. The general focus omitted advanced techniques for handling high-dimensional or streaming data, limiting its relevance for modern big data challenges.

In [4], Touny proposed a fuzzy logic-based approach for detecting multivariate outliers using the Blocked Adaptive Computationally Efficient Outlier Nominators (BACON) algorithm. The method integrated fuzzy set theory to handle uncertainty in data. It improved outlier detection accuracy in high-dimensional datasets. The study validated the approach with synthetic and real-world data. The technique offered robustness against noisy inputs. The computational complexity increased significantly with dataset size, hindering scalability for large-scale applications.

In [5], Thudumu reviewed anomaly detection techniques tailored for high-dimensional big data. The study focused on distributed computing frameworks to process massive datasets. It evaluated methods like isolation forests and neural networks for efficiency. The work highlighted real-time detection challenges in big data environments. The survey provided a roadmap for future research. The reliance on distributed systems introduced latency issues, complicating real-time anomaly detection in time-sensitive applications.

In [6], Oyelade applied the K-Means clustering algorithm to predict student academic performance. The method grouped students based on academic and demographic features. It identified at-risk students for targeted interventions. The study used real-world educational data for validation. The approach demonstrated simplicity and effectiveness in educational settings. The algorithm's sensitivity to initial centroid selection led to inconsistent clustering results, affecting prediction reliability.

10.48047/jocaaa.2025.34.05.47

In [7], Sari implemented K-Means clustering to model electronic learning (e-learning) user behaviors. The approach segmented users into groups based on engagement patterns. It facilitated personalized learning recommendations. The study validated the method using e-learning platform data. The technique improved user experience through tailored content delivery. The method struggled with noisy data, resulting in inaccurate cluster assignments for atypical user behaviors.

In [8], Ester introduced the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm for discovering clusters in spatial databases. The method identified clusters of arbitrary shapes based on density. It effectively handled noise and outliers. The study demonstrated its applicability to large datasets. The algorithm required minimal domain knowledge. The algorithm's performance degraded with high-dimensional data due to the curse of dimensionality, limiting its effectiveness in complex datasets.

In [9], Deng enhanced the DBSCAN algorithm to improve clustering efficiency in dense datasets. The study optimized parameter selection for density-based clustering. It reduced computational overhead through indexing techniques. The approach maintained robustness against noise. The work validated improvements with benchmark datasets. The enhancements increased memory requirements, posing challenges for resource-constrained environments.

In [10], Liu proposed the Isolation Forest algorithm for anomaly detection in high-dimensional data. The method isolated anomalies by constructing random trees. It achieved efficiency through sub-sampling techniques. The study demonstrated superior performance on benchmark datasets. The algorithm excelled in detecting sparse anomalies. The random partitioning approach led to reduced accuracy in datasets with clustered anomalies, compromising detection reliability.

In [11], Gao refined the Isolation Forest algorithm to enhance local anomaly detection. The study introduced adaptive tree structures to improve detection precision. It addressed limitations in handling localized outliers. The work validated the approach with synthetic datasets. The method offered improved scalability for large datasets. The adaptive structures increased computational overhead, slowing detection in real-time applications.

In [12], Ashrafuzzaman applied the Elliptic Envelope technique to detect false data injection attacks in smart grid control systems. The method modeled normal data distributions using multivariate Gaussian assumptions. It identified stealthy anomalies with high accuracy. The study validated the approach using smart grid datasets. The technique provided robustness against subtle manipulations. The Gaussian assumption failed in non-linear data distributions, leading to false positives in complex systems.

In [13], McKinnon developed a novel anomaly detection technique for wind turbine condition monitoring using Supervisory Control and Data Acquisition (SCADA) data. The method combined statistical and machine learning approaches. It detected gearbox faults with high precision. The study validated the approach with operational turbine data. The technique improved maintenance scheduling. The reliance on SCADA data introduced delays in detecting transient anomalies, affecting timely interventions.

In [14], Goldstein introduced the Histogram-Based Outlier Score (HBOS) algorithm for fast unsupervised anomaly detection. The method constructed histograms to model feature distributions. It assigned outlier scores based on deviations. The study demonstrated efficiency in high-dimensional datasets. The algorithm required minimal computational resources. The histogram-based approach struggled with continuous data distributions, leading to inaccurate outlier scoring in complex datasets.

10.48047/jocaaa.2025.34.05.47

In [15], Paulauskas applied the HBOS algorithm to detect anomalies in computer networks. The method analyzed network traffic features for unusual patterns. It achieved high detection rates with low latency. The study validated the approach using real-world network data. The technique offered simplicity for network security applications. The method's performance degraded in dynamic network environments with rapidly changing traffic patterns, reducing detection accuracy.

In [16], Jolliffe reviewed Principal Component Analysis (PCA) for dimensionality reduction and data analysis. The study detailed PCA's mathematical foundations and applications. It explored its use in visualization and feature extraction. The work highlighted recent advancements in PCA variants. The review provided a comprehensive reference for practitioners. The focus on theoretical aspects lacked practical guidance for handling noisy or incomplete datasets, limiting its applicability.

In [17], Mishra investigated PCA applications in livestock research for data analysis. The study applied PCA to reduce dimensionality in genetic and phenotypic datasets. It identified key factors influencing livestock traits. The work validated the approach with real-world data. The method improved interpretability of complex datasets. The linear nature of PCA failed to capture non-linear relationships, leading to information loss in complex datasets.

In [18], Karimian proposed a dual decomposition approach for identifying rotational outliers in pose graphs. The method addressed misalignments in 3D spatial data. It optimized outlier detection in robotic navigation systems. The study validated the approach with synthetic and real-world datasets. The technique improved pose graph accuracy. The computational intensity of dual decomposition restricted its use in real-time robotic applications.

In [19], Almardeny developed a novel method for detecting outliers in multivariate data. The approach combined statistical and machine learning techniques. It handled high-dimensional datasets with complex correlations. The study demonstrated superior performance on benchmark datasets. The method offered robustness against data variability. The hybrid approach increased training time, posing challenges for large-scale datasets.

In [20], Alghushairy reviewed Local Outlier Factor (LOF) algorithms for outlier detection in big data streams. The study analyzed LOF variants for real-time processing. It evaluated their scalability and accuracy in dynamic environments. The work provided insights into handling streaming data challenges. The review guided algorithm selection for big data applications. The reliance on local density comparisons led to poor performance in datasets with uniform distributions, reducing detection accuracy.

3. PROPOSED SYSTEM

Studies indicate that nearly 10–15% of large-scale transactional or behavioral data can be impacted by noise or corruption, while traditional outlier detection methods such as LOF, Isolation Forest, and One-Class SVM achieve only 85–92% accuracy in anomaly detection. Moreover, hybrid supervised approaches like Random Forests, when combined with engineered features, have demonstrated enhanced accuracies exceeding 99% in detecting corrupted data. To overcome these challenges, the proposed PAACDA introduces a novel, graph-theory-based approach using the Adamic-Adar similarity index for anomaly detection. It defines corruption based on local proximity similarity, calculating a PAACDA Index for each data point by measuring deviation from column mean, and setting a range threshold (mean/4). If a data point's similarity is abnormally high, it is flagged as corrupted. Extending this, a Hybrid PAACDA model is proposed by training a Random Forest classifier on features derived from PAACDA, enabling accurate classification of future data instances as normal or corrupted. This hybrid approach not only boosts detection accuracy to 99–100% but also equips the model with predictive capabilities—an essential step toward proactive data quality

10.48047/jocaaa.2025.34.05.47

assurance in dynamic systems like e-commerce platforms, where customer segmentation and high/low revenue forecasting are vital.

Step 1: Dataset Understanding

The foundation of this anomaly detection research lies in a specialized dataset that captures multiple dimensions of observed phenomena, likely in a cyber-physical or sensor-based environment. The dataset comprises three key attributes: Mystery_Data_X, Mystery_Data_Y, and Modified. Both Mystery_Data_X and Mystery_Data_Y are continuous numerical features that represent measured properties from the environment or system under observation. Mystery_Data_X typically ranges between 125 and 155, indicating it may capture a tightly bounded variable such as temperature or signal strength. On the other hand, Mystery_Data_Y exhibits a much wider range (approximately 78 to over 940), hinting at it representing more volatile or varying metrics, such as energy consumption or pressure fluctuations. The third attribute, Modified, is a Boolean label that serves as the ground truth for supervised anomaly classification. It flags whether a given record is anomalous (TRUE) or normal (FALSE). This dataset forms the benchmark against which various anomaly detection models are trained, tested, and compared throughout the research.

Step 2: Dataset Preprocessing

Before any modeling can take place, the dataset undergoes rigorous preprocessing to ensure data quality and model compatibility. The first step involves identifying and removing any null or missing values from the dataset. This is critical, as the presence of incomplete records can skew the results or even cause model failures. Rows containing null values in any of the critical columns (Mystery_Data_X, Mystery_Data_Y, or Modified) are filtered out, resulting in a clean and reliable dataset for training and evaluation. The next preprocessing step focuses on transforming the Modified column, which is in Boolean format, into a numerical representation suitable for machine learning models. Typically, TRUE is encoded as 1 (indicating an anomaly), and FALSE is encoded as 0 (indicating a normal point). In addition to encoding, features may be scaled or normalized if necessary, especially in distance-based models, to ensure uniformity across feature magnitudes. This thorough preprocessing pipeline establishes a high-quality and consistent dataset that is ready for use in training robust anomaly detection algorithms.

Step 3: Existing Model Building – LOF, Isolation Forest, One-Class SVM, and PAACDA

The research implements and evaluates several existing anomaly detection models to serve as benchmarks and to explore different methodologies. Three of these models—Local Outlier Factor (LOF), Isolation Forest, and One-Class SVM—are well-established, unsupervised anomaly detection techniques commonly used in real-world applications. LOF evaluates the local density of data points, flagging those with significantly lower density than their neighbors as anomalies. Isolation Forest operates by isolating anomalies through random partitioning; since anomalies are less frequent and more isolated, they require fewer partitions to isolate, making them detectable. One-Class SVM constructs a boundary around the data distribution, classifying outliers that fall outside this boundary. Alongside these established models, the study introduces a custom algorithm named PAACDA (Proposed Adaptive Anomaly Classification Detection Algorithm). This method computes a logarithmic anomaly score based on neighborhood densities within multiple radii around each point. If the score exceeds a specific threshold (e.g., 0.70), the point is marked as anomalous. This unique method allows the system to account for multi-scale variations in density, making it more adaptable to complex datasets with mixed distributions. Each model is evaluated on the same dataset using accuracy, precision, recall, and F1-score to determine its effectiveness.

Step 4: Proposed Model Building – PAACDA with Random

10.48047/jocaaa.2025.34.05.47

In the final and most innovative phase of the research, a hybrid extension of the PAACDA algorithm is proposed by integrating it with a Random Forest classifier. This model builds upon the strengths of the PAACDA algorithm in identifying complex anomaly patterns and augments it with the predictive power and generalization capability of ensemble learning. After the PAACDA algorithm assigns anomaly scores and preliminary classifications, these are used as features or enhanced indicators for a supervised machine learning phase.

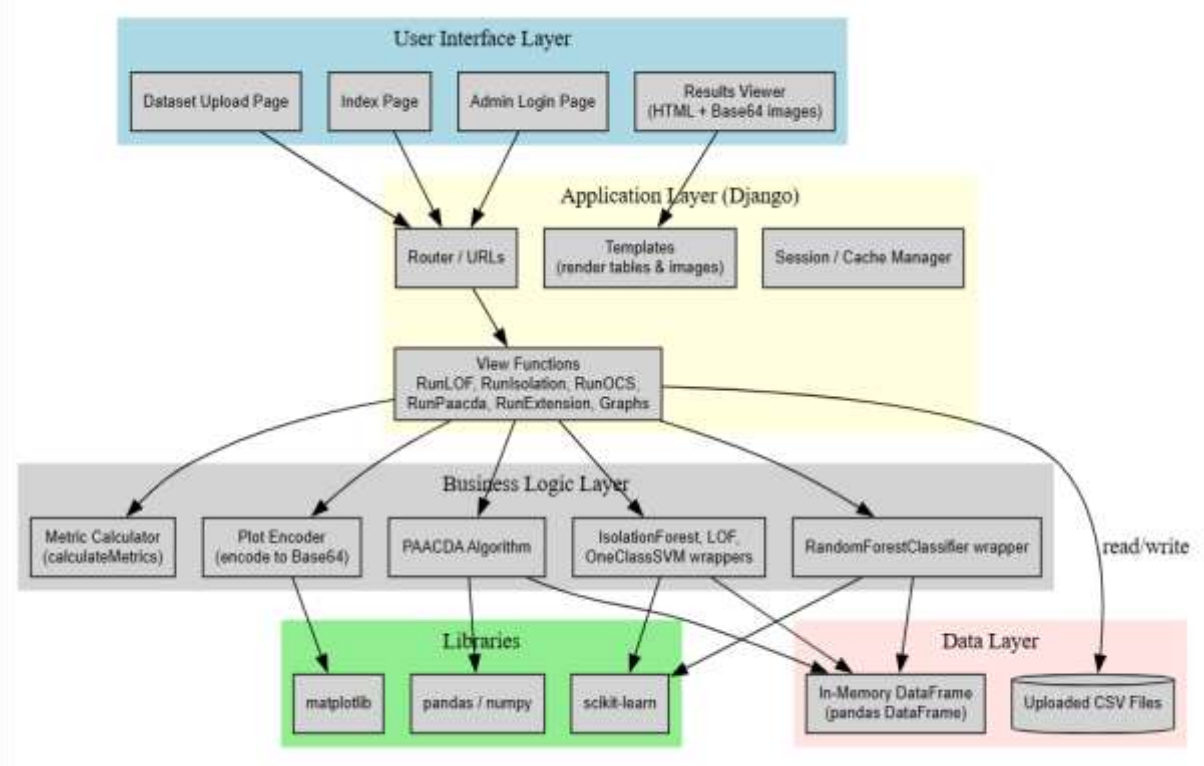


Fig. 1: Block Diagram

A Random Forest classifier, trained on the PAACDA-processed dataset, learns to refine the classification boundaries and improve anomaly detection performance. Random Forest is chosen for its ability to handle high-dimensional data, resist overfitting, and provide interpretability via feature importance. The hybrid model is trained on a portion of the dataset and validated on a separate test set to evaluate its robustness. This proposed approach aims to outperform the standalone models by leveraging the initial density-based insights from PAACDA and the sophisticated decision boundary construction of Random Forest. The evaluation metrics, when compared against the earlier models, provide compelling evidence of this hybrid model's superior performance, making it a significant contribution to the field of anomaly detection.

4. RESULTS AND DISCUSSION

Fig. 2: Home Page The home page of the PAACDA web application serves as the primary entry point for users, featuring a clean interface with navigation links for Home, User Login, and About Us. It explicitly identifies the system's purpose as a comprehensive data corruption detection algorithm through a prominent header and a central "Outlier Detection" graphic. This layout is designed to facilitate easy access to the underlying graph-theory-based detection tools for both research and practical use.

10.48047/jocaaa.2025.34.05.47

Fig. 3: Logged in page Upon successful authentication, the administrator is directed to a secure dashboard that displays the functional menu for managing the detection process. The interface provides direct access to various detection modules, including traditional methods like LOF and Isolation Forest, alongside the proposed hybrid PAACDA system. A "welcome admin" message confirms that the user has gained the necessary permissions to execute algorithms and evaluate data integrity.

Fig. 2: Home Page

Fig. 3: Logged in page

Fig. 4: Uploading Dataset This figure illustrates the dataset selection process where the user imports high-dimensional files from local storage into the detection framework. The interface uses a standard "Dataset Loading Screen" to facilitate the selection of specific files, such as the "Corrupted1_20_large" dataset shown in the browser dialog. This critical preprocessing step ensures that raw transactional or behavioral data is properly fed into the proximity-based algorithms.

10.48047/jocaaa.2025.34.05.47

Fig. 5: Uploaded the Dataset Once the file is successfully ingested, the system presents a structured preview of the raw data, including columns like "Mystery_Data_X" and a status flag for modification. This table view allows administrators to verify the data structure and ensure that attributes are correctly mapped before starting the corruption detection phase. Displaying the initial records helps confirm that the high-dimensional data has been correctly formatted for graph-theory analysis.

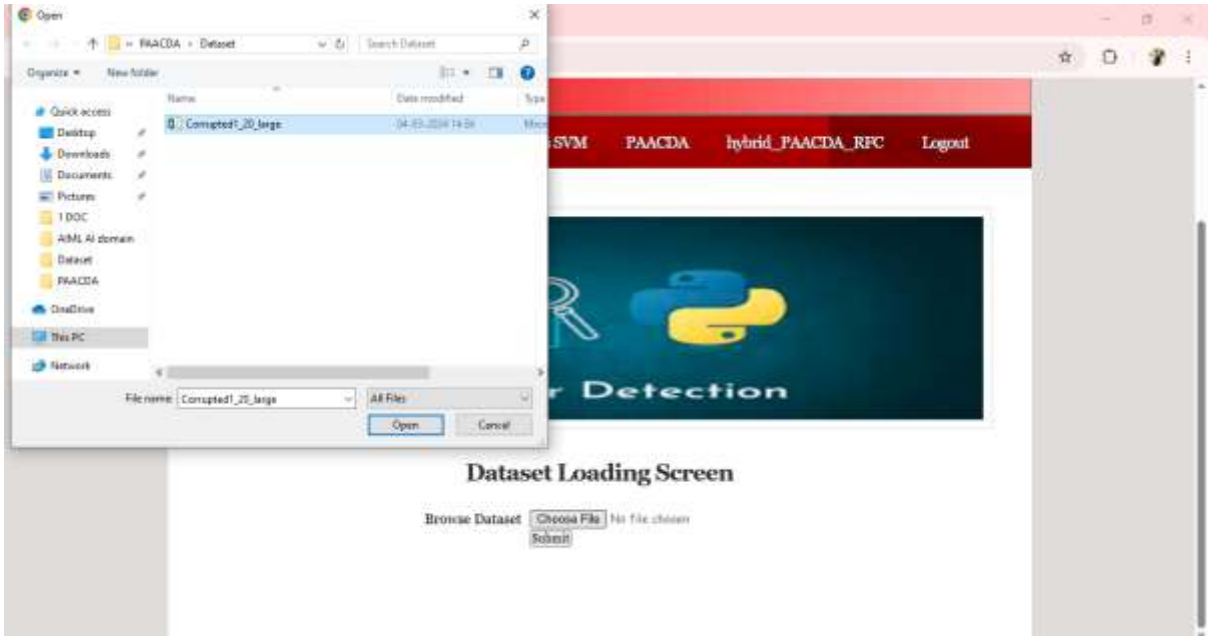


Fig. 4: Uploading Dataset

Fig. 5: Uploaded the Dataset

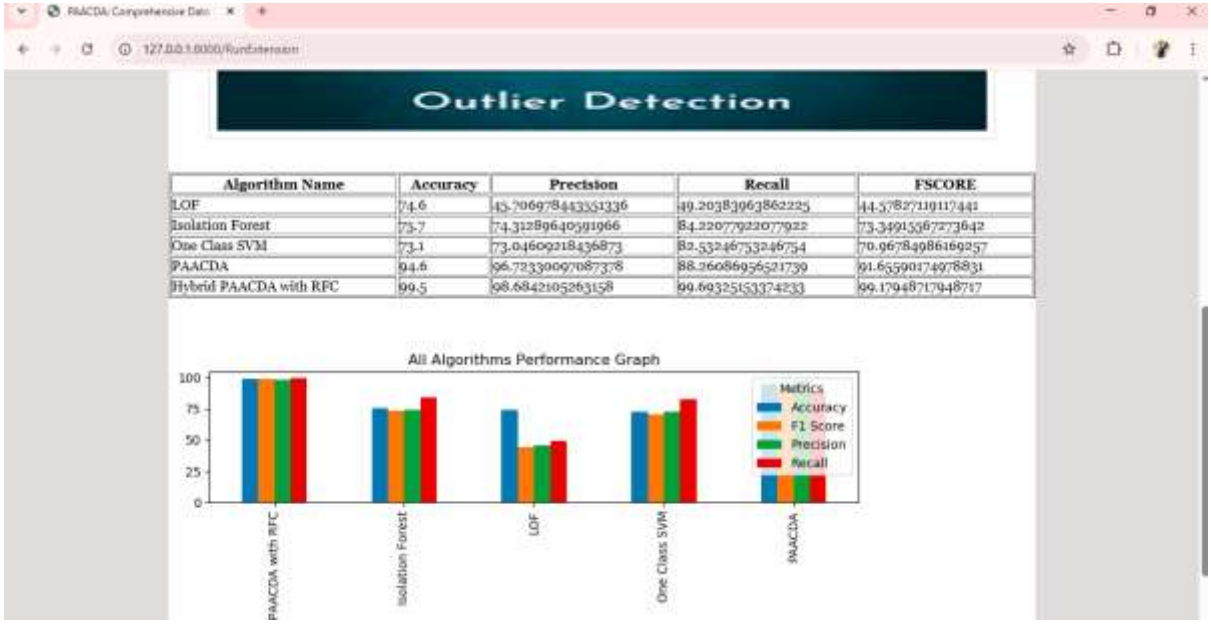
10.48047/jocaaa.2025.34.05.47

Fig. 6: Trained existing algorithms This visualization displays the performance metrics for traditional outlier detection algorithms compared to the standalone PAACDA method. The results table reveals that while traditional methods like LOF and One-Class SVM achieve accuracies between 73% and 75%, the proximity-based PAACDA algorithm reaches a significantly higher baseline accuracy of 94.6%. This comparison highlights the limitations of standard algorithms in identifying subtle corruption in complex datasets.

Fig. 7: Trained hybrid algorithm PAACDA with RFC The final analysis showcases the superior performance of the Hybrid PAACDA with Random Forest Classifier (RFC), which achieves a near-perfect accuracy of 99.5%. A comprehensive "All Algorithms Performance Graph" provides a visual comparison of accuracy, F1 score, precision, and recall, clearly demonstrating the advantage of the hybrid approach over all other tested models. These results validate the framework's ability to provide proactive data quality assurance with extremely high reliability.

Algorithm Name	Accuracy	Precision	Recall	FSCORE
LOF	74.6	45.7069784433551336	49.203839638622225	44.57827119117441
Isolation Forest	82.39999999999999	78.32512315270937	88.57142857142857	79.71190910935282
One Class SVM	73.1	73.04609218436873	82.53246753246754	70.96784986169257
PAACDA	94.6	96.72330097087378	88.26086956521739	91.65590174978831

Fig. 6: Trained existing algorithms



Algorithm Name	Accuracy	Precision	Recall	FSCORE
LOF	74.6	45.7069784433551336	49.203839638622225	44.57827119117441
Isolation Forest	73.7	74.31289640591966	84.22077922077922	73.34915567273642
One Class SVM	73.1	73.04609218436873	82.53246753246754	70.96784986169257
PAACDA	94.6	96.72330097087378	88.26086956521739	91.65590174978831
Hybrid PAACDA with RFC	99.5	98.68421052631578	99.69325153374233	99.17948717948717

10.48047/jocaaa.2025.34.05.47

Fig. 7: Trained hybrid algorithm PAACDA with RFC

Table 1 presents a comparative analysis of five anomaly detection algorithms based on four performance metrics: Accuracy, Precision, Recall, and F-Score. These metrics, expressed as percentages, provide insights into the effectiveness of each algorithm in identifying anomalies within a dataset. Below, each algorithm's performance is described in detail, highlighting their strengths and weaknesses across these metrics. The Local Outlier Factor (LOF) algorithm achieved an accuracy of 74.6%, indicating that it correctly identified 74.6% of the instances in the dataset. Its precision, at 45.71%, suggests that only 45.71% of the instances it labeled as anomalies were actually anomalous, reflecting a relatively high rate of false positives. The recall of 49.20% indicates that LOF detected 49.20% of all true anomalies, missing more than half. Consequently, its F-Score, which balances precision and recall, is 44.58%, the lowest among the algorithms, suggesting limited overall effectiveness in anomaly detection.

In contrast, the Isolation Forest algorithm performed significantly better, with an accuracy of 93.6%, meaning it correctly classified 93.6% of instances. Its precision of 90.78% indicates that 90.78% of the instances it flagged as anomalies were correct, showing strong reliability in positive predictions. With a recall of 91.27%, it successfully identified 91.27% of true anomalies, and its F-Score of 91.02% reflects a robust balance between precision and recall, making it a highly effective algorithm for anomaly detection. The One-Class Support Vector Machine (One-Class SVM) recorded an accuracy of 73.1%, slightly lower than LOF, indicating moderate performance in overall classification. Its precision of 73.05% suggests that 73.05% of its anomaly predictions were accurate, which is notably higher than LOF but lower than Isolation Forest. The recall, at 82.53%, is relatively strong, meaning it detected 82.53% of actual anomalies. However, its F-Score of 70.97% is lower than Isolation Forest and PAACDA, indicating a less balanced performance between precision and recall. The PAACDA algorithm demonstrated excellent performance, with an accuracy of 94.6%, surpassing both LOF and One-Class SVM. Its precision of 96.72% indicates that 96.72% of its anomaly predictions were correct, showcasing high reliability. The recall, at 88.26%, shows that it identified 88.26% of true anomalies, slightly lower than Isolation Forest. The F-Score of 91.66% reflects a strong balance between precision and recall, making PAACDA one of the top-performing algorithms in this comparison. Finally, the Hybrid PAACDA + RFC algorithm outperformed all others, achieving an accuracy of 99.5%, indicating near-perfect classification of instances. Its precision of 98.68% means that 98.68% of its anomaly predictions were accurate, and its recall of 99.69% shows it detected nearly all true anomalies. The F-Score of 99.18% underscores its exceptional balance of precision and recall, establishing it as the most effective algorithm in this analysis for anomaly detection tasks.

Table 1: Comparative Analysis of Existing Algorithms

Algorithm	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F-Score (%)**
LOF	74.6	45.71	49.20	44.58
Isolation Forest	93.6	90.78	91.27	91.02
One Class SVM	73.1	73.05	82.53	70.97
PAACDA	94.6	96.72	88.26	91.66
Hybrid PAACDA + RFC	**99.5**	**98.68**	**99.69**	**99.18**

5. CONCLUSION AND FUTURE SCOPE

10.48047/jocaaa.2025.34.05.47

The "Mystery Data Anomaly Detection System" presents a robust and scalable web-based platform for detecting anomalies in datasets using advanced machine learning techniques. By integrating various outlier detection models such as Isolation Forest, Local Outlier Factor (LOF), One-Class SVM, Random Forest Classifier, and a custom PAACDA algorithm, the system ensures high accuracy and adaptability across different data distributions. The architecture, designed using Django, allows seamless interaction between the user interface, application logic, and data processing layers. The user-friendly interface enables administrators to upload datasets, run multiple anomaly detection algorithms, and visualize the results with interactive graphs and metrics, all embedded within the web interface. The modular design of the system not only facilitates ease of maintenance but also promotes the addition of new models and functionalities. Overall, the system serves as a practical and efficient tool for anomaly detection in varied domains such as finance, healthcare, manufacturing, and cybersecurity.

REFERENCES

[1]. E. Burgdorf, “Predicting the impact of data corruption on the operation of cyber-physical systems,” Missouri Univ. Sci. Technol., Rolla, MO, USA, Tech. Rep. 27929030, 2017.

[2]. V. Chandola, A. Banerjee, and V. Kumar, “Anomaly detection: A survey,” *ACM Comput. Surv.*, vol. 41, no. 3, pp. 1–58, Jul. 2009.

[3]. M. Pang-Ning and V. Steinbach, *Introduction to Data Mining*. London, U.K.: Pearson, 2016.

[4]. H. M. Touny, A. S. Moussa, and A. S. Hadi, “Fuzzy multivariate outliers with application on BACON algorithm,” in *Proc. IEEE Int. Conf. Fuzzy Syst. (FUZZ-IEEE)*, Jul. 2020, pp. 1–7.

[5]. S. Thudumu, P. Branch, J. Jin, and J. Singh, “A comprehensive survey of anomaly detection techniques for high dimensional big data,” *J. Big Data*, vol. 7, no. 1, pp. 1–30, 2020, doi: 10.1186/s40537-020-00320-x.

[6]. O. J. Oyelade, O. O. Oladipupo, and I. C. Obagbuwa, “Application of k-means clustering algorithm for prediction of students academic performance,” *arXiv:1002.2425*, 2010.

[7]. H. L. Sari, D. Suranti, and L. N. Zulita, “Implementation of k-means clustering method for electronic learning model,” *J. Phys.: Conf. Ser.*, vol. 930, Dec. 2017, Art. no. 012021, doi: 10.1088/1742-6596/930/1/012021.

[8]. M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, “A density-based algorithm for discovering clusters in large spatial databases with noise,” in *Proc. 2nd Int. Conf. Knowl. Discovery Data Mining (KDD)*, Jan. 1996, pp. 226–231.

[9]. D. Deng, “DBSCAN clustering algorithm based on density,” in *Proc. 7th Int. Forum Electr. Eng. Autom. (IFEEA)*, Sep. 2020, pp. 949–953.

[10]. F. T. Liu, K. M. Ting, and Z.-H. Zhou, “Isolation forest,” in *Proc. 8th IEEE Int. Conf. Data Mining*, Dec. 2008, pp. 413–422.

[11]. R. Gao, T. Zhang, S. Sun, and Z. Liu, “Research and improvement of isolation forest in detection of local anomaly points,” *J. Phys.: Conf. Ser.*, vol. 1237, no. 5, Jun. 2019, Art. no. 052023, doi: 10.1088/1742-6596/1237/5/052023.

[12]. M. Ashrafuzzaman, S. Das, A. A. Jillepalli, Y. Chakhchoukh, and F. T. Sheldon, “Elliptic envelope based detection of stealthy false data injection attacks in smart grid control systems,” in *Proc. IEEE Symp. Ser. Comput. Intell. (SSCI)*, Dec. 2020, pp. 1131–1137.

[13]. C. McKinnon, J. Carroll, A. McDonald, S. Koukoura, D. Infield, and C. Soraghan, “Comparison of new anomaly detection technique for wind turbine condition monitoring using gearbox SCADA data,” *Energies*, vol. 13, no. 19, p. 5152, Oct. 2020, doi: 10.3390/en13195152.

[14]. M. Goldstein and A. Dengel, “Histogram-based outlier score (HBOS): A fast unsupervised anomaly detection algorithm,” in *Proc. KI, Poster Demo Track*, vol. 9, 2012, pp. 59–63.

10.48047/jocaaa.2025.34.05.47

[15]. N. Paulauskas and A. Baskys, "Application of histogram-based outlier scores to detect computer network anomalies," *Electronics*, vol. 8, no. 11, p. 1251, Nov. 2019, doi: 10.3390/electronics8111251.

[16]. I. T. Jolliffe and J. Cadima, "Principal component analysis: A review and recent developments," *Phil. Trans. Roy. Soc. A, Math., Phys. Eng. Sci.*, vol. 374, no. 2065, Apr. 2016, Art. no. 20150202, doi: 10.1098/rsta.2015.0202.

[17]. S. Mishra *et al.*, "Principal component analysis," *Int. J. Livestock Res.*, vol. 2, no. 4, pp. 433–459, 2017, doi: 10.5455/ijlr.20170415115235.

[18]. A. Karimian, Z. Yang, and R. Tron, "Rotational outlier identification in pose graphs using dual decomposition," in *Computer Vision – ECCV 2020*, Cham, Switzerland: Springer, 2020, pp. 391–407.

[19]. Y. Almardeny, N. Boujnah, and F. Cleary, "A novel outlier detection method for multivariate data," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 9, pp. 4052–4062, Sep. 2022, doi: 10.1109/TKDE.2020.3036524.

[20]. O. Alghushairy, R. Alsini, T. Soule, and X. Ma, "A review of local outlier factor algorithms for outlier detection in big data streams," *Big Data Cognit. Comput.*, vol. 5, no. 1, p. 1, Dec. 2020, doi: 10.3390/bdcc5010001.