

Integration of Isolation Forest, Fellegi-Sunter, and DeepER for Anomaly Detection and Data Deduplication

Jamaludin^{1*}, Ahmad Musyafa², Taswanda Taryo³, Tri Istiana⁴

^{1,2,3}Program Studi Magister Teknik Informatika, Pasca Sarjana, Universitas Pamulang
Kota Tangerang Selatan, Banten 15310, Indonesia

⁴Badan Meteorologi Klimatologi dan Geofisika, BMKG Jakarta, Indonesia

jamaludin@bmkg.go.id¹, ahmad.musyafa@unpam.ac.id², taswandataryo@gmail.com³, tri.istiana@bmkg.go.id⁴

Article Info

Article history:

Received 2026-08-18

Revised 2026-09-07

Accepted 2026-10-19

Keyword:

*Anomaly Detection,
Deep Entity Resolution,
Fellegi-Sunter,
Isolation Forest,
Synoptic Data..*

ABSTRACT

Synoptic data is essential for climate analysis and aviation safety. However, sensor anomalies and record duplications at the Sudjarwo Tjondro Negoro Meteorological Station cause significant analytical biases. Current Quality Control (QC) procedures rely on manual checks and static thresholds, which lack adaptability to extreme weather. This research proposes an automated QC framework integrating artificial intelligence and statistics to address these limitations. The method applies Isolation Forest to detect multivariate outliers without data distribution assumptions. For deduplication, it utilizes the probabilistic Fellegi-Sunter (FS) method for exact record matching, alongside deep learning-based Deep Entity Resolution (DeepER) for resolving fuzzy duplicates. Using a one-year hourly synoptic dataset, the methodology encompasses preprocessing, anomaly detection, and systematic deduplication. Performance is evaluated through score distribution for anomalies and overlap analysis for deduplication. Results show Isolation Forest successfully isolates anomalies with a 2% contamination rate, capturing the highest deviation in Wind Direction (163.85° mean difference), validating suspected mechanical damage. In deduplication, FS efficiently identified 14,629 identical duplicates. DeepER complemented this by detecting 13,003 additional hidden duplicate candidates with minor noise missed by FS. Integrating these methods significantly reduces reliance on manual thresholding, providing a robust meteorological dataset to support aviation safety analyses.



This is an open access article under the [CC-BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.

I. INTRODUCTION

Synoptic data is surface meteorological observation data collected periodically (e.g., hourly or every 3 hours) at weather stations [1]. This data includes crucial parameters such as air temperature, pressure, humidity, wind speed, and rainfall, which play a vital role in weather analysis, numerical model assimilation, and climate change studies. The quality of synoptic data greatly determines the accuracy of weather forecasting and extreme weather prediction [2]. As the number of stations and data volume increase [3], the challenge of real-time Quality Control (QC) becomes increasingly complex [10].

Currently, meteorological data QC procedures in institutions like BMKG mostly rely on conventional methods

based on static threshold-based rules and manual checking. This approach has limitations, particularly in detecting complex multivariate anomalies, thus frequently resulting in high rates of false alarms and missed detections [4]. Furthermore, apart from instrumental anomalies, synoptic data quality is also affected by duplication issues or recording inconsistencies—for example, due to the integration of manual observation logs and Automatic Weather Stations (AWS). If not addressed through appropriate record linkage mechanisms, duplications can lead to analytical bias due to double-counting [5].

Previous studies have shown that the advancement of machine learning opens up opportunities for smarter and more automated QC improvements [6], [11]. The Isolation Forest algorithm has proven effective in identifying multivariate

outliers without requiring initial data distribution assumptions [7]. Meanwhile, for handling duplicate data, the probabilistic Fellegi-Sunter (FS) method has long been used as the foundation for record matching [8]. Furthermore, deep learning-based approaches like Deep Entity Resolution (DeepER) are capable of learning latent representations (embeddings) of data to capture semantic similarities automatically [9], [15].

Research Gap: Despite the individual advancements in anomaly detection and record linkage, there is a distinct lack of research integrating physical anomaly filtering and multi-tier deduplication into a single, cohesive, unsupervised pipeline for unlabelled meteorological datasets.

Contributions: The methodological novelty lies in the sequential pipeline: data is first filtered from anomalous physical behaviors using iForest, preventing extreme weather variations from being falsely flagged as duplicates. Subsequently, deduplication is executed in two tiers: FS handles exact statistical matches, while DeepER resolves remaining fuzzy duplicates in the latent space. This integration provides a comprehensive solution for unlabelled historical logs, supporting robust meteorological data for aviation safety analyses at the Sudjarwo Tjondro Negoro Meteorological Station.

II. METHOD

This section outlines the systematic approach used to construct the automated Quality Control (QC) framework for meteorological synoptic data. The analytical workflow is divided into data preprocessing, unsupervised anomaly detection, and hybrid deduplication.

A. Data Requirements and Preprocessing

The study utilizes a historical hourly synoptic dataset over one full operational year (January 1, 2024 – December 31, 2024) from the Sudjarwo Tjondro Negoro Meteorological Station, Serui-Papua. Based on the hourly frequency, the dataset is expected to contain approximately 8,784 observation records. The analyzed multivariate parameters include 13 variables. The preprocessing step filters administrative errors by:

- 1) Standardizing time formats to UTC;
- 2) Converting built-in machine diagnostic error codes (e.g., -999) into NaN format; and
- 3) Removing impossible physical violations. Prior to anomaly detection, all continuous variables were standardized using feature scaling to ensure that variables with larger numerical ranges (e.g., pressure) do not dominate the Isolation Forest path length calculations.

B. Anomaly Detection using Isolation Forest

The administratively clean data is inputted into the Isolation Forest (iForest) algorithm. iForest is an unsupervised anomaly detection algorithm based on an ensemble of decision trees that directly isolates rare and distinct anomalies without needing normal distribution

assumptions [7]. The anomaly score is calculated based on the path length, formulated as:

$$c(n) = 2H(n-1) - \frac{2(n-1)}{n} \quad (1)$$

Where $H(i) \approx \ln(i) + \gamma$, and γ is Euler's constant. The anomaly score is defined as:

$$s(x, n) = 2^{-\frac{E[h(x)]}{c(n)}} \quad (2)$$

If the value of $s(x, n)$ approaches 1, the point is highly likely to be an anomaly. The contamination parameter in this study is heuristically set at 0.02 (2%). This threshold was selected based on preliminary domain knowledge regarding the expected historical sensor failure rate at the station, acting as a conservative upper bound to differentiate valid extreme weather variations from field technical malfunctions without aggressively pruning valid severe weather events

C. Hybrid Deduplication (Fellegi-Sunter and DeepER)

To overcome data redundancy caused by the integration of manual logging and AWS, a parallel two-tier approach is implemented:

1) *Adapted Fellegi-Sunter (Vector Similarity)*: The classical FS method relies on likelihood ratios of matching attributes [12]. However, due to the lack of ground truth labels for Expectation-Maximization (EM) estimation [5], this study adaptively implements FS into a Vector Similarity Matching approach. Daily records are zero-padded, and their similarity is calculated directly using Cosine Similarity with a strict threshold of 0.95.

2) *DeepER (Autoencoder LSTM Embedding)*: DeepER addresses "hidden duplicates with minor noise", operationally defined as records describing the exact same physical weather event but containing slight discrepancies due to logging delays (e.g., timestamps differing by 2 minutes) or floating-point rounding errors. Adapted for unsupervised data, the Siamese-style Autoencoder LSTM model is trained through 100 epochs to minimize Mean Squared Error (MSE) reconstruction loss. Daily embeddings are extracted from the encoder layer and evaluated using Cosine Similarity with the identical 0.95 threshold [14]. Although the raw padded vector and the 8-dimensional latent embedding operate in different geometric spaces, the identical 0.95 threshold is established as a heuristic control to force both models to operate at the same extreme confidence level, limiting error tolerance to 5%.

D. Overlap Analysis Evaluation

Because ground truth labels are absent, the performance is not evaluated through cross-punishing metrics, but through Overlap Analysis using the Jaccard Similarity Index to map the consensus between the two deduplication methods. It is formulated as:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (3)$$

Where A is the set of duplicates detected by Fellegi-Sunter, and B is the set of duplicates detected by DeepER.

III. RESULTS AND DISCUSSION

This chapter presents the comprehensive results and detailed discussions of the proposed automated Quality Control (QC) framework applied to the historical synoptic dataset. The findings are systematically organized to reflect the sequential pipeline of the methodology. It begins by examining the outcomes of the unsupervised physical anomaly detection using the Isolation Forest algorithm, highlighting its capability to identify multivariate sensor malfunctions and contextual weather extremes. Subsequently, the chapter details the performance of the two-tier hybrid deduplication process. It explores the computational efficiency of the adapted Fellegi-Sunter method as the first-tier filter for exact redundancies, followed by the effectiveness of the DeepER Autoencoder LSTM model in resolving complex, fuzzy duplicates in a latent space. Finally, to address the absence of absolute ground truth labels, the chapter concludes with a rigorous Overlap Analysis, demonstrating empirically how both deduplication methods complementarily integrate to produce a clean, reliable meteorological dataset.

A. Physical Anomaly Detection Results (Isolation Forest)

The Isolation Forest effectively classified daily observations into normal (blue) and anomaly (red) groups. Evaluations of key parameters yielded the following findings:

- 1) Wind Direction: Showed the most significant deviation. Mean difference analysis ranked wind direction 1st with a gap of 163.85 degrees. Due to the absence of absolute maintenance logs, this finding was qualitatively evaluated as suspected mechanical damage (stuck wind vane) [16]. However, it must be explicitly noted that wind direction is circular data (0° – 360°). The unsupervised calculation in Isolation Forest currently treats this linearly, meaning values like 359° and 1° mathematically appear as extreme deviations. Therefore, this anomaly indicates a combination of suspected physical errors and the algorithmic limitation of linear representation, highlighting the need for circular regression transformations in future studies.
- 2) Wind Speed: Normal data concentrated at 0–6 m/s. Extreme wind gusts up to 21 m/s were isolated, as well as contextual anomalies within the 6–8 m/s range that disrupted combined patterns.
- 3) Visibility: Normal scores were between 8–10. Anomalies were successfully captured at low visibilities (0–4), correlating with temporary physical events like heavy rain or fog.
- 4) Air Pressure (QFF): Highly homogeneous between 1005–1012 mb. One extreme outlier was identified at ~1100 mb, completely isolated from the major distribution, verifying a severe logging error.

- 5) 3D Visualization: Plotting the top three problematic features (Wind Direction, Current Weather Code, and Relative Humidity) demonstrated that anomalies are multivariate occurrences appearing on the distribution boundaries or isolated entirely from the dense main clusters.

B. First-Tier Deduplication Results (Adapted Fellegi-Sunter)

In the first pipeline stage, the adapted probabilistic approach based on Vector Similarity was applied to the zero-padded raw vectors. By setting the similarity threshold strictly at 0.95, this classical method successfully identified 14,629 date pairs as exact duplicate data. The histogram distribution showed a strong concentration at the 0.90–0.95 peak, proving this method's computational efficiency in sweeping dimensionally identical operational duplicates.

C. Second-Tier Deduplication Results (DeepER)

In the deep semantic resolution stage, the DeepER Autoencoder LSTM model processed the data into an 8-dimensional latent space. Utilizing the identical 0.95 threshold, DeepER detected a substantially higher volume of duplication candidates, identifying 25,365 pairs. The generated heatmap visually confirmed strong similarity blocks extending beyond the main diagonal, validating the model's ability to cluster fuzzy data characteristics regardless of minor format variations.

D. Overlap Analysis in Hybrid Integration

Due to the absence of absolute human-made ground truth labels, forcing a True Label assumption between the two methods would be methodologically invalid. Therefore, the evaluation was transformed into Overlap Analysis to assess how the two methods integrate as a pipeline.

TABLE 1.
OVERLAP ANALYSIS AND SIMILARITY MAPPING RESULT

Overlap Evaluation Category	Number Of Pairs
Total Observation Data Evaluation	114,481
Consensus (Intersection)	12,362
Exclusive Tier 1 (FS Only)	2,267
Exclusive Tier 2 (DeepER Only)	13,003
Similarity Index	0.4474

Based on Table 1, the hybrid deduplication framework proves its validity:

- 1) Classical Strength (Consensus of 12,362 pairs): 12,362 records were captured by FS and fully agreed upon by DeepER. This represents the core of operational duplicate data that are identical in both text and numerical formats. FS serves perfectly as a fast, first-tier cleaner.
- 2) Statistical Rigidity Limitation (Exclusive FS 2,267 pairs): FS only found 2,267 exclusive pairs. This proves that almost 95% of the maximum capabilities of the classical statistical method are successfully encompassed and understood by the Deep Learning neural network.

- 3) Semantic Resolution Effectiveness (Exclusive DeepER 13,003 pairs): Through similarity mapping in the latent space, DeepER successfully unraveled 13,003 hidden duplicate candidates (fuzzy duplicates) that were rejected by the rigid FS linear boundaries. Minor decimal variations (e.g., 25.6°C vs. 25.8°C recorded at 07:00 and 07:02 UTC) were successfully generalized by the machine as integration synchronization issues, rather than separate climate events.
- 4) Jaccard Similarity Index (0.4474): This moderate figure academically validates that FS and DeepER do not redundantly overlap, but map different error dimensions in a complementary manner. It must be acknowledged that without labeled ground truth, traditional entity resolution metrics such as Precision, Recall, and F1-Score cannot be objectively computed. Therefore, the 13,003 exclusive findings by DeepER represent "highly probable duplicate candidates" that require human-in-the-loop expert verification prior to automated deletion. Nevertheless, compared to traditional threshold-based QC methods [4], [5], this hybrid integration successfully automates the initial screening of fuzzy duplicates without relying on rigid manual rules, confirming previous studies on deep learning capabilities in meteorological data curation [13], [14].

IV. CONCLUSION

Based on the experimentation and comparative evaluation within the automated Quality Control framework on synoptic data at Sudjarwo Tjondro Negoro Station, the following key conclusions are drawn :

- 1) Error Logic Separation and Physical Anomaly Detection: The framework successfully established firm functional boundaries. Preprocessing effectively swept system errors (-999 converted to Missing Values). Furthermore, Isolation Forest successfully captured Physical Anomalies with a 2% contamination rate. The Wind Direction parameter contributed the largest extreme deviation (163.85 degrees), empirically validating the machine's success in detecting instrument mechanical damage in the field.
- 2) Overlap Analysis Validation: Evaluation of 114,481 pairs using overlap mapping (rather than dropping comparisons) proved the synergy of integration. Both methods reached a strong consensus, designating 12,362 pairs as absolute duplicates, supported by a Jaccard Similarity Index of 0.4474, demonstrating that the methods map complementary problem-solving spaces.
- 3) Methodological Integration Success: Pipeline integration is proven to be vastly superior to a single approach. Tier 1, powered by adapted Fellegi-Sunter, is highly efficient as an initial filter for exact identical duplicates. Tier 2, powered by the DeepER Autoencoder, acted as an advanced safety net, successfully extracting 13,003 hidden duplicate candidates (with minor noise due to system synchronization delays) that linear statistical

methods consistently fail to catch. Because of the absence of absolute ground truth, these findings represent highly probable candidates requiring expert verification before automated deletion.

ACKNOWLEDGMENT

The author expresses gratitude to the Sudjarwo Tjondro Negoro Meteorological Station, Serui, Papua (BMKG), for providing the 2024 operational database, which served as the primary foundation for the smooth execution of this computational thesis research.

REFERENCES

- [1] WMO-No. 488 (Official), "WMO," 2022.
- [2] X. Sun, X. Zhong, X. Xu, and D. Chen, "A data-to-forecast machine learning system for global weather," no. December 2024, 2025.
- [3] H. Sun et al., "A Machine Learning-Based Quality Control Algorithm for Heavy Rainfall Using Multi-Source Data," pp. 1–24, 2025.
- [4] Y. Sha, D. J. Gagne, G. West, and R. Stull, "Deep-learning-based precipitation observation quality control," *J. Atmos. Ocean. Technol.*, vol. 38, no. 5, pp. 1075–1091, 2021, doi: 10.1175/JTECH-D-20-0081.1.
- [5] O. Azeroual, M. Jha, A. Nikiforova, K. Sha, M. Alsmirat, and S. Jha, "A Record Linkage-Based Data Deduplication Framework with DataCleaner Extension," *Multimodal Technol. Interact.*, vol. 6, no. 4, 2022, doi: 10.3390/mti6040027.
- [6] C. Bodnar et al., "A foundation model for the Earth system," *Nature*, vol. 641, no. April, 2025, doi: 10.1038/s41586-025-09005-y.
- [7] T. Zemichal and T. G. Dietterich, "Anomaly Detection in the Presence of Missing Values for Weather Data Quality Control," 2019.
- [8] X. Li, H. Xu, and S. Grannis, "The Data-Adaptive Fellegi-Sunter Model for Probabilistic Record Linkage: Algorithm Development and Validation," vol. 24, pp. 1–15, 2022, doi: 10.2196/33775.
- [9] S. Mudgal et al., "Deep Learning for Entity Matching," pp. 19–34, 2018, doi: 10.1145/3183713.3196926.
- [10] E. Beele, M. Reyniers, R. Aerts, and B. Somers, *Quality control and correction method for air temperature data from a citizen science weather station network in Leuven, Belgium*, 2022.
- [11] F. Wang, Y. Jiang, R. Zhang, A. Wei, J. Xie, and X. Pang, "A Survey of Deep Anomaly Detection in Multivariate Time Series: Taxonomy, Applications, and Directions," *Sensors*, vol. 25, no. 1, pp. 1–27, 2025, doi: 10.3390/s25010190.
- [12] M. Ebraheem, S. Thirumuruganathan, S. Joty, M. Ouzzani, and N. Tang, "Distributed Representations of Tuples for Entity Resolution," vol. 11, no. 11, 2018.
- [13] T. Zaarour and A. Zahran, *Using Distributed Ledgers To Build Knowledge Graphs For Decentralized Computing Ecosystems*, vol. 1, no. 1. Association for Computing Machinery. doi: 10.1145/3627673.3679644.
- [14] C. Sun and D. Shen, "Neurocomputing Towards deep entity resolution via soft schema matching," *Neurocomputing*, vol. 471, pp. 107–117, 2022, doi: 10.1016/j.neucom.2021.10.106.
- [15] D. Jin, B. Sisman, H. Wei, X. L. Dong, and D. Koutra, "Deep Transfer Learning for Multi-source Entity Linkage via Domain Adaptation," *Proc. VLDB Endow.*, vol. 15, no. 3, pp. 465–477, 2021, doi: 10.14778/3494124.3494131.
- [16] A. Bâra, A. Gabriel, V. Simona, and V. Oprea, "Anomaly Detection in Weather Phenomena : News and Numerical Data - Driven Insights into the Climate Change in Romania ' s Historical Regions," *Int. J. Comput. Intell. Syst.*, vol. 2, no. 6, pp. 1–26, 2024, doi: 10.1007/s44196-024-00536-2.