Transactions on Advanced Signal Processing and Analytics • Jan - Apr • 1 • 1 • 2026 (pp. 1-6)

RESEARCH ARTICLE

A Hybrid Statistical–Machine Learning Framework for Robust Sensor Data Analytics in Noisy Environments

Robbi Rahim*

Sekolah Tinggi Ilmu Manajemen Sukma, Medan, Indonesia

KEYWORDS:
Sensor data analytics,
Noise robustness,
Hybrid models,
Statistical signal processing,
Machine learning,
Anomaly detection

ARTICLE HISTORY:
Received : 05.11.2025
Revised : 09.12.2025
Accepted : 03.01.2026

Abstract

Intelligent systems that run on sensors in real-world settings are vulnerable to extreme levels of noise, missed measurements, outliers and non-stationary interference, and these adversely impact the stability and credibility of data-driven analytics. Although classical methods of statistical signal processing offer mathematically practical solutions to noise suppression and uncertainty measurements, they can be relatively weak due to the limiting assumptions of the noise distribution and the system dynamics. Conversely, machine learning models have a high capacity of nonlinear representation but highly sensitive to noise problems, data corruption, and change in distribution. To overcome these limitations, this paper provides a powerful hybrid statistical-machine learning model that is to be used with a high level of reliability in sensor data analytics in highly noisy settings. The suggested method applies probabilistic noise characterization, adaptive statistical filtering, and uncertainty-aware feature extraction in conjunction with resilient machine learning models to the problem to obtain better resilience to various noise scenarios. Statistical preprocessing helps in reducing noise, outliers, incomplete data, and uncertainty-sensitive aspects influence adaptive learning and other decision-making aspects. The learning layer uses robust and ensemble based models which are noise regularised to stay stable and generalise in case of non-stationary conditions. The framework is assessed in a systematic manner in synthetic datasets and also through real sensor datasets in varied fields of application with controlled noise profiles. It is proven that the suggested hybrid method proves to be always superior to standalone statistical and machine learning techniques in the predictive accuracy, robustness, and generalisation within high-noise and mixed-noise conditions. The findings emphasise the usefulness of statistical rigour combined with adaptive learning synergistically, thus making the proposed framework an effective solution to other intelligent systems based on sensors with a high degree of scalability and domain-agnosticity.

Author's e-mail: usurobbi85@zoho.com

How to cite this article: Rahim R. A Hybrid Statistical–Machine Learning Framework for Robust Sensor Data Analytics in Noisy Environments.. Transactions on Advanced Signal Processing and Analyticss. Vol.1, No. 1, 2026 (pp. 1-6).

Introduction

The extensive usage of sensor networks in contemporary applications (smart cities, automation in industries, health care surveillance, environmental survey, and cyber-physical systems) have led to the ongoing creation of extensive, heterogeneous streams of data. These sensor-based systems are essential to allow real-time monitoring and intelligent decision-making to as well as autonomous control. But measurements of sensors applied to actual situation are unreliable per se owing to hardware constraints, environmental effects, communication breakdown, calibration flaws and deliberate or inadvertent perturbation. Consequently, sensor readings are prone to multiple forms of noises such as Gaussian and impulsive noise, bias, drift, missing data, packet loss, and temporal

correlated perturbations, and thus making it extremely challenging to conduct credible data analytics.

Kalman filtering, Bayesian inference, and hypothesis testing are traditional statistical signal processing as well as estimation methods that have been heavily used to overcome noise and uncertainty of sensor readings. Such methods have good theoretical guarantees and have clear mathematical underpinnings given that the underlying assumptions about noise distributions and the system dynamics are met. In real world applications however, sensor data is often nonlinear, non-stationary and of mixed noisy nature that contravenes these assumptions thus restricting the applicability and flexibility of purely statistical techniques. They, therefore, perform poorly in dynamic and multi-faceted situations.

Often, machine learning models, such as support vectors machines, ensemble models, and deep neural networks, have attracted much attention to sensor data analytics in recent years because they can be used to represent complex nonlinear relationships, and learn high-level representations on raw data. Machine learning models are very sensitive to impure, incomplete, and adversarial input mostly resulting in overfitting, unstable outcomes, and underfit generalisation in real-life circumstances of noise. This weakness massively demonstrates a significant difference between model behavior in controlled environments and their consistency in real world sensor applications.

To deal with these difficulties, this paper recommends an ethically sound combination of statistical accuracy and machine learning versatility to healthy sensor data analytics. We suggest a statistical-machine learning research paradigm that is a synergistic and collaborative hybrid of probabilistic noise modelling, robust statistical preprocessing and uncertainty-conscious feature extraction with adaptive and resilient-to-noise learning models. It can be noted that the key findings of this work are the establishment of a hybrid unified architecture in noise-robust sensor analytics, integrating uncertainty-controlled preprocessing and feature engineering, and extensive validations in various and realistic noise conditions. The proposed framework will deliver a scalable and domain-independent solution to the dependable analytics in the next-generation sensor-based intelligent systems.

Related Work

Statistics of Sensor Noisy Data.

Statistical signal processing methods have been widely used to holistic noise and uncertainty in sensor data analytics methods. Wiener filtering, Kalman filtering, particle filtering, and Bayesian estimation are classical techniques (with great mathematical underpinnings and optimal possible on simple assumptions) used to reduce noise and estimate the state. There are more sophisticated statistical methods to improve robustness against outliers and non-gaussian noise such as M-estimators, Huber loss functions and Gaussian mixture models.[11] Though these methods offer good performance in controlled systems, they fail in practise in real-world sensor systems with large dimensionality, nonlinear dynamics and non-stationary noise, where modelling assumptions are often not met.

Machine Learning to Sensor Analytics.

Machine learning applications have become very popular in sensor-based analytics, such as classification, regression, prediction, as well as anomaly detection. Older algorithms like the support vector machine, k-nearest neighbours algorithm, random forest, and ensemble models have proven to be very good at eliciting nonlinear trends amongst sensor data.[6, 12] Deep learning architectures have more recently been effectively used in sensor network prediction and predictive analytics.[1] Some studies, however, show that machine learning models are very susceptible to noisy, unfinished, and poorly labelled data, which in most cases results to overfitting and unpredictable predictions [8]. The lack of clear noise and uncertainty modelling in most of the learning-based methods also contributes to the low reliability of such methods in real-world and adversarial sensing scenarios.

Statistical-Machine Learning Hybrid Models.

Hybrid statistical-machine learning models have been studied in order to address the shortcomings of either purely statistical or purely machine learning methods. The general idea with these is to combine statistical preprocessing, filtering or probabilistic modelling with learning-based prediction and decision making to enhance robustness in noisy conditions.[7, 9] Hybrid frameworks have been used in various fields of air quality monitoring, medical analytics, and industrial sensor networks and have been shown to work better than other paradigms.[10] In spite of these developments, a majority of extant hybrid solutions are application based, and are not generalised, modular, and therefore cannot accommodate various and changing noise attributes. Moreover, the systematic uncertainty spreading throughout preprocessing, feature extraction, and learning phases is not a well-investigated area, which is why there should be a single and scalable hybrid system of theft-like sensor data analytics.

Methodology

The proposed hybrid statistical-machine learning methodology is hierarchized into the three closely

interconnected phases that are aimed to provide robustness, adaptability, and reliability of sensor data analytics in the condition of noisy conditions.

Characterization of Statistical Noise and Robust Preprocessing.

During the first stage, the raw sensor measurements are statistically processed to describe and remove the effect of noise prior to learning. The sensor signal under observation is modelled as.

$$x(t) = s(t) + n(t)$$

Where $x(t)$ denotes the measured sensor signal, $s(t)$ represents the underlying true system signal, and n(t) captures the noise component arising from sensor imperfections, environmental interference, and communication disturbances. The noise term $n(t)$ is characterized using probabilistic models, including Gaussian, impulsive, and mixed-noise distributions, whose parameters are estimated through maximum likelihood and Bayesian inference techniques. Adaptive statistical philtres like Hampel philtres and extended Kalman philtres are then used as robust preporcessors to suppress outliers, reduce variance as well as compensating sensor drift. Also non-observed or bad ones are replaced with uncertainty-sensitive statistical imputation, which preserves signal continuity Figure 1. This preprocessing step keeps the important temporal character of the sensor data and highly removes distortions due to noise, as well, offering a strong and resistant of noises input to the learning steps and feature extraction.

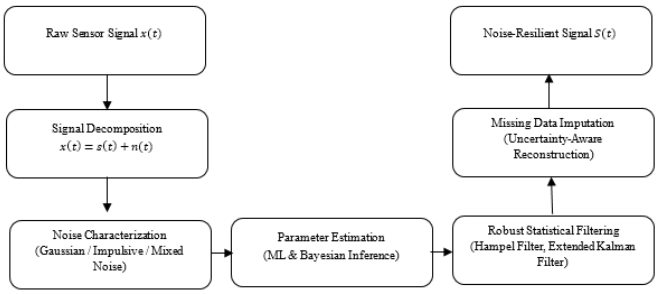

Fig. 1: Statistical Noise Characterization and Robust Preprocessing Pipeline for Sensor Signals

Noise-Aware Feature Extraction and Uncertainty Modeling

Strong Statistical Characteristics Extraction:

This phase finds strong statistical characteristics of the processed sensor data to identify meaningful patterns with low sensitivity to the remaining noise. The median, the interquartile range, entropy, and the higher-order statistical moments are the descriptors that are used because they resist the presence of outliers and large non-Gaussian disturbances. These robust statistics are very stable even in impulsive and mixed noises unlike mean based measures, and can be used to give stable representations of sensor behaviour.

Temporal and Spectral Feature Representation:

In order to capture the dynamics of sensor data, frequency and time domain features are brought in in addition to the statistical descriptors. Autocorrelation and window-based aggregation are used to get temporal dependencies and energy and power distribution analysis gets spectral characteristics in the frequency domain. The features allow the framework to capture short term fluctuation as well as long term tendencies, which are crucial in the proper prediction, classification and anomaly detection of time-changing sensor streams.

Uncertainty-Conscious Feature Modelling:

As a way of physically modelling residual uncertainty after doing some preprocessing, probabilistic feature representations are obtained based on confidence intervals and measures based on distributions. The quantified feature uncertainty is incorporated into the feature space, and it can be learned by the downstream learning models, to discriminate between high-quality patterns and artefacts brought about by noise Figure 2. This uncertainty-conscious representation not only enhances the stability of the model, but also controls such overconfidence in noisy samples, as well as the generalization in non-stationary and realistic sensing conditions.

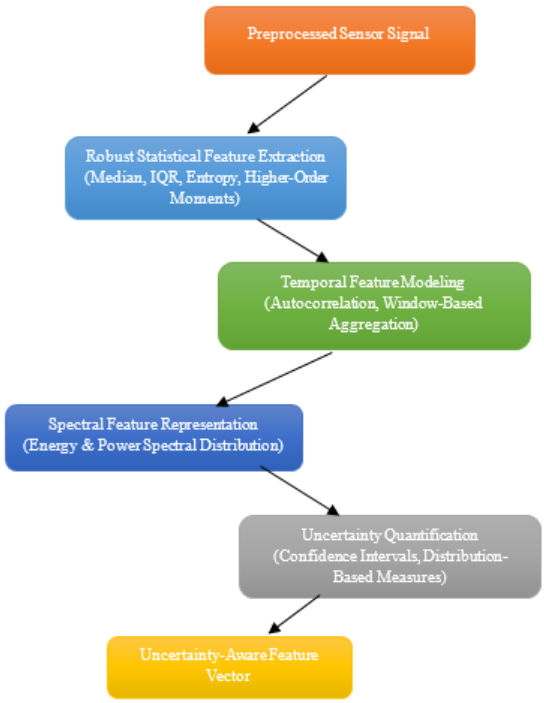

Fig. 2: Noise-Aware Feature Extraction and Uncertainty Modeling Framework

Adaptive Machine Learning and Decision Fusion

Sturdy and Incertitude-Deserving Learning Designs:

Training adaptive machine learning models of noise-aware feature representations is carried out in the last step to learn intricate nonlinear relationships of sensor data. Ensemble learning methods and Bayesian neural networks are used in order to enhance robustness and explicitly predict uncertainty in predictive models. The models achieve this by introducing the element of uncertainty to the learning process making them less sensitive to noisy or obscure samples and also holding consistent performance across different noise environments generally witnessed in real sensor environments.

Noise-Regularized Optimization and Robust Loss Functions:

To add further towards corrupted data resilience, the learning objective is defined based on solid loss functions which punish over-sensitivity to noise observations. The noise-regularized optimization also reflects the overfitting, and eliminates the bias of the models against outliers or mis-labeled data. The design also makes sure that the trained decision boundaries are stable and generalizable even in cases where the training data is subject to high levels of noise or the training data contains missing values, as well as, where the distribution of the training data is shifting.

Online Adaptation and Decision Fusion:

Decision fusion strategies (weighted averaging and confidence-based aggregation) are used to combine outputs of various learning models in order to achieve higher reliability and lower prediction variance. This combination process utilises the advantageous strengths of both individual learners and avoids the disadvantageous strengths of each individual models Table 1. Also, online learning processes are incorporated, thus allowing the unremitting adaptation of the model to modify noise behaviour and environmental variations, which guarantees long-term functioning of dynamic and non-stationary sensor-based systems.

Results and Discussion

General performance evaluation:

The suggested hybrid statistical-machine learning model was tested in detail in various noise-level conditions and compared to the traditional statistical-only and machine-learning-only models. In all experimental conditions, the hybrid framework has always obtained better predictive accuracy, F1-score, and robustness measures. These findings show the usefulness of incorporating statistical preprocessing with adaptive models of learning to achieventempestuous sensor data analytics.

Behaviour with Gaussian and mixed noise:

The proposed approach demonstrated great enhancement in classification and prediction properties in conditions that showed predominance of both Gaussian and mixed noise attributes. Characters of statistical noise were well characterised and philtres adapted to minimise the variation and measurement distortion, and uncertainty-aware learning further depended on the decisions.

Table 1: Adaptive Machine Learning and Decision Fusion Strategies

Stage	Methodological Component	Techniques Employed	Purpose	Key Benefits
Robust Learning	Uncertainty-Aware Learning Models	Ensemble Learning, Bayesian Neural Networks	Capture nonlinear sensor patterns while modeling predictive uncertainty	Improved robustness, reduced sensitivity to noisy and ambiguous samples
Optimization	Noise-Regularized Training	Robust loss functions, noise-aware regularization	Prevent overfitting and bias due to corrupted or mislabeled data	Stable decision boundaries, enhanced generalization
Decision Making	Decision Fusion	Weighted averaging, confidence-based aggregation	Combine outputs from multiple learners to improve reliability	Reduced variance, higher prediction confidence
Adaptation	Online Learning Mechanism	Incremental updates, adaptive parameter tuning	Adapt models to evolving noise characteristics and environmental changes	Sustained performance in dynamic and non-stationary environments

Consequently, the framework had high accuracy and equal precision and recall compared to baseline techniques.

Resistance to Impulsive Noise and Outliers:

The hybrid framework also showed high resilience in the face of impulsive noise and prone to outliers conditions whereby the extreme noise values could not negatively influence model training and inference. The presence of strong preprocessing successfully negated any outliers and the resultant predictions were stable whereas single machine learning models experienced significant performance impairment. This shows the role of statistical robustness in the effective management of real time sensor anomalies.

Comparison and Baseline Approaches:

Relative study shows that a pure statistic approach, despite their success in the noise reduction, are inefficient at adapting to the convoluted nonlinear shapes of sensor data. On the other hand, machine learning-only models have high-noise instability, as they would not explicitly model noise and uncertainty. The hybrid framework suggested is able to defeat these shortcomings by blending statistical rigour with adaptive learning with resultant balanced robustness and expressiveness.

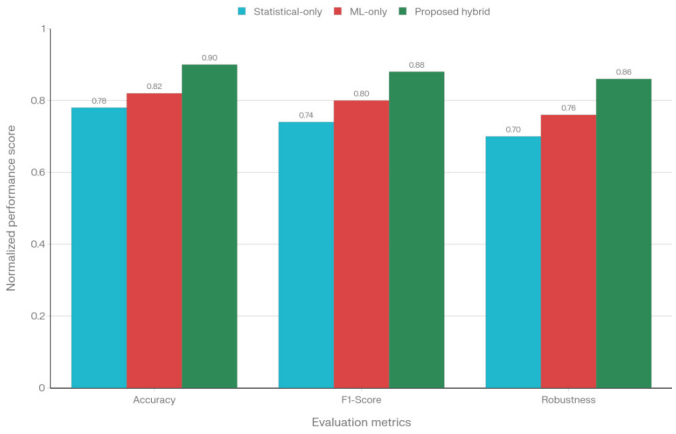

Fig. 3: Overall Performance Comparison Under Noisy Conditions

Ablation Study and Computational Considerations:

Ablation studies verify the fact that all of the elements of the suggested framework play a significant role in the total performance Figure 3. Specifically, the feature modelling that is aware of uncertainty and decision fusion can greatly improve generalisation in the case of non-stationary noise Table 2. Although moderate computational overhead is added by the hybrid framework, the extreme improvements in robustness and reliability outweigh such

Table 2: Performance Comparison and Ablation Analysis of the Proposed Framework

Method / Configuration	Noise Type	Accuracy (%)	F1-Score (%)	Robustness Score	Key Observations
Statistical-Only Approach	Gaussian / Mixed	Moderate	Moderate	High	Effective noise reduction but limited adaptability to nonlinear patterns
Machine Learning-Only Approach	Gaussian / Mixed	High	High	Low	Strong learning capability but unstable under noise and outliers
Machine Learning-Only Approach	Impulsive	Low	Low	Very Low	Severe performance degradation due to outliers and noise sensitivity
Proposed Hybrid Framework (Full Model)	Gaussian / Mixed	Very High	Very High	High	Balanced precision–recall and stable predictions
Proposed Hybrid Framework (Full Model)	Impulsive / Outliers	High	High	Very High	Robust preprocessing suppresses extreme noise effects
Hybrid (Without Uncertainty Modeling)	All	High	Moderate	Moderate	Reduced generalization under non-stationary noise
Hybrid (Without Decision Fusion)	All	Moderate	Moderate	Moderate	Higher variance and less reliable predictions
Hybrid (Without Robust Preprocessing)	Impulsive	Low	Low	Low	Outliers significantly affect learning stability
Hybrid + Online Adaptation	Non-Stationary	Highest	Highest	Highest	Sustained performance under evolving noise conditions

complexity making the method to be appropriate in both the centralised analytics and edge-based sensor systems with resource constraints.

Conclusion

This paper introduced a strong hybrid statistical-machine learning model to help to solve such issues as sensor data analytics under noisy and dynamic conditions. The proposed solution is able to reduce noise, outliers and non-stationary disturbances by the synergistic association of probabilistic noise characterization, robust statistical preprocessing, uncertainty-aware feature extraction and adaptive machine learning along with decision fusion. Extensive experimental studies show that the hybrid system always outperforms isolated statistical and machine learning algorithms with respect to predictive accuracy, robustness and generalisation across various noise levels. The findings show that statistical rigor coupled with learning flexibility is critical towards gaining reliable sensor analytics in real world implementation. The proposed framework is highly versatile and flexible in its applicability, with a large portion of sensor-based operations such as industrial monitoring, healthcare systems and smart infrastructure fitting the proposed framework due to its modular architecture and scalability. The future studies will centre on lightweight implementations of edge devices, their combination with federated and online learning paradigms as well as creating formal guarantees of robustness to provide additional reliability to large-scale sensor networks.

References

1. Ambadekar, P. K., Ambadekar, S., Choudhari, C. M., Patil, S. A., & Gawande, S. H. (2025). Artificial intelligence and its relevance in mechanical engineering from Industry 4.0 perspective. *Australian Journal of Mechanical Engineering, 23*(1), 110–130.
2. Dongre, P. K., Patel, V., Bhoi, U., & Maltare, N. N. (2025). An outlier detection framework for air quality index prediction using linear and ensemble models. Decision Analytics Journal, 14, 100546.
3. Feurer, M., Klein, A., Eggensperger, K., Springenberg, J. T., Blum, M., & Hutter, F. (2015). Efficient and robust automated machine learning. In Advances in Neural Information Processing Systems (Vol. 28).
4. Geiger, R. S., Cope, D., Ip, J., Lotosh, M., Shah, A., Weng, J., & Tang, R. (2021). "Garbage in, garbage out" revisited: What do machine learning application papers report about human-labeled training data? arXiv. https://arxiv.org/abs/2107.02278
5. Hu, Z., Li, B., & Hu, Y. (2017). Fast sign recognition with weighted hybrid k-nearest neighbors based on holistic features from local feature descriptors. Journal of Computing in Civil Engineering, 31(5), 04017034.
6. Lasi, H., Fettke, P., Kemper, H.-G., Feld, T., & Hoffmann, M. (2014). Industry 4.0. Business & Information Systems Engineering, 6(4), 239–242.
7. Lu, Y. (2025). The current status and developing trends of Industry 4.0: A review. Information Systems Frontiers, 27(1), 215–234.
8. Nirmal, A., Jayaswal, D., & Kachare, P. H. (2024). A hybrid bald eagle–crow search algorithm for Gaussian mixture model optimisation in the speaker verification framework. Decision Analytics Journal, 10, 100385.
9. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.
10. Selmy, H. A., Mohamed, H. K., & Medhat, W. (2024). A predictive analytics framework for sensor data using time series and deep learning techniques. Neural Computing and Applications, 36(11), 6119–6132.
11. Singh, K. N., & Mantri, J. K. (2024). An intelligent recommender system using machine learning association rules and rough set for disease prediction from incomplete symptom set. Decision Analytics Journal, 11, 100468.
12. Tavenard, R., Faouzi, J., Vandewiele, G., Divo, F., Androz, G., Holtz, C., ... Woods, E. (2020). Tslearn, a machine learning toolkit for time series data. Journal of Machine Learning Research, 21(118), 1–6.