SvedbergOpen
DISSEMINATION OF KNOWLEDGE
International Journal of Artificial Intelligence and Machine Learning
Publisher's Home Page: https://www.svedbergopen.com/
Research Paper Open Access

Learning Under Uncertainty in Sparse and Noisy Data Environments

A. Yasodai[1], Deepika Sharma[2], Kapil Mundada[3], Pramod Rahate[4], Arivukkodi R[5], Uma Maheswari G[6], Dr. M. Sangeetha[7], Shubhashish Goswami[8]

[1] Associate Professor, Department of Electronics and Communication Engineering, Fatima Michael College of Engineering and Technology, Madurai, Tamil Nadu, India. Email: yasodaikannan@gmail.com
[2] Assistant Professor, Department of Computer Sciences, Noida International University, Greater Noida, Uttar Pradesh, India. Email: deepika.sharma@niu.edu.in
[3] Department of Instrumentation and Control Engineering, Vishwakarma Institute of Technology, Pune, Maharashtra 411037, India. Email: kapil.mundada@vit.edu
[4] Department of Mechanical Engineering, Bharati Vidyapeeth's College of Engineering, Lavale, Pune, Maharashtra, India. Email: pramodgrahate@gmail.com
[5] Assistant Professor, Department of Computer Science, Meenakshi College of Arts and Science, Meenakshi Academy of Higher Education and Research, India. Email: arivukodir@maher.ac.in
[6] Assistant Professor, Department of Mathematics, Meenakshi College of Arts and Science, Meenakshi Academy of Higher Education and Research, India. Email: umamahes@maher.ac.in
[7] Associate Professor, Department of Computer Science and Engineering, Panimalar Engineering College, India. Email: Sangeethameckanzi2224@gmail.com ORCID: 0000-0002-1973-6339
[8] Associate Professor, Department of Computer Science & Engineering, Dev Bhoomi Uttarakhand University, Dehradun, Uttarakhand, India. Email: socse.shubhashish@dbuu.ac.in ORCID: 0000-0002-6129-9822

Abstract
Learning from real-world data is challenging because sparse observations, stochastic noise, and uncertainty often reduce the reliability and predictive capability of conventional machine learning models. Existing approaches mainly focus on either noise mitigation or sparse data handling and often fail to simultaneously model aleatoric and epistemic uncertainties, thereby leading to unstable predictions in uncertain environments. To address this limitation, this research proposes an uncertainty-aware learning framework based on a Chaotic Dragonfly Algorithm-driven Efficient Deep Neural Network (CDA-EDNN) for accurate prediction in sparse and noisy data environments. Research utilizes the Uncertainty Sparse Noisy Signal Dataset, containing signal characteristics associated with noise, sparsity, uncertainty, and prediction quality. Data preprocessing is performed using Min-Max Normalization to scale feature values and reduce magnitude variations. Subsequently, Independent Component Analysis (ICA) is employed for feature extraction to separate latent source signals and enhance signal representation. The CDA optimizes network parameters, while the EDNN learns robust feature patterns under uncertainty conditions. Experimental results demonstrate that the proposed model achieves superior stability with an Area under the Precision-Recall Curve (AUPRC) of 0.96, an accuracy of 0.98, and an F1-Score of 0.97 compared to standalone machine learning and traditional interpolation methods. The model is implemented using Python 3.10, TensorFlow 2.x, NumPy, Pandas, and Matplotlib on a Graphics Processing Unit (GPU)-enabled computing platform. Overall, the proposed CDA-EDNN framework provides a scalable and reliable solution for prediction tasks involving sparse, noisy, and uncertain data.

Keywords: Uncertainty Learning, Sparse Data, Noisy Data, Sparsity, Deep Neural Network (EDNN), Chaotic Dragonfly Algorithm (CDA).

1. Introduction

The problem of data sparsity is an ongoing issue in similarity-based algorithms especially within the current age of technology when only few observations are common. Within the context of sparse data, estimation of similarities is not effective as well as reliable for making predictions [1]. Current applications, including social media and electronic commerce, constantly create huge amounts of data streams in mini-batch form, leading to huge and sparse data [2]. In light of the development of the digital economy, the collection of data takes place at a much faster pace, thus requiring effective management, processing, and analysis of such sparse data. In addition to this, the occurrence of redundancies in web data makes it hard for users to search for information [3]. The fact that noise probability distributions are randomly occurring, as well as that they lack a sparse representation, makes them extremely hard to model using the sparse domain method [4]. As such, sparse regularization techniques are extensively used to remove noise from signals. Uncertainty-sensitive machine learning methods are extended to different applications, where they surpass traditional methods in small data set cases, especially for meteorology and climatology, materials science, biomedical engineering, and economics. Time series generators generally consider noisy inputs to be uniformly distributed in time, thus obtaining regular time series [5, 6]. In any case, irregularities can be observed within uncertain systems because of problems with sensors, communication latencies, and inaccurate measurements. Irregularities cause uncertainty in the sampling phase, thus reducing the quality of the state space representation. In this regard, models become very sensitive to observational errors, negatively impacting performance during the inference and forecast phases. Empirical Mode Decomposition (EMD) is a method that enables addressing these problems and improves the representation under observation noise and sparsity [7, 8].

Research Aim: The objective of this research is the designs of an uncertainty-based learning model that can predict accurately even in cases where the data environment is sparse and noisy. In particular, the research seeks to utilize the new Chaotic Dragonfly Algorithm-based Efficient Deep Neural Network (CDA-EDNN).

Research Organization: Research is divided into five parts. Section 1 introduces the research intro with background on learning from sparse and noisy data. Section 2 highlights previous work related to uncertainty learning and robust machine learning models. The proposed methodology is presented in Section 3 focuses with the CDA-EDNN model. Experimental results and performance evaluation based on metrics such as MSE, and R2 values are discussed in Section 4. Finally, conclusions are discussed in section 5.

2. Related work

The prediction of a pivotal part in Intelligent Transportation Systems (ITS) for reducing the level of noise was investigated in [9] using the Bayesian Neural Network (BNN) method for noise prediction. The outcome revealed that there is improvement while having different levels of data availability, such as having less available data (90%, 50%, and 10%), together with aleatoric uncertainty learning, baseline conditions, and consistent gains under the limitations of noisy data. The application of a machine learning model in predicting spare demand in end-of-life products was analyzed in [10]. Using the Decay-Function-Blended Machine Learning (DFB-ML) in predicting maintenance activities and end-of-life stock requirements resulted in a Root Mean Square Error (RMSE) value of 741.39, which indicates high predictive performance even for low and irregular demands, and there is a limitation in dropout modeling and sparse data availability. Addressing the issues of data sparsity in processing data scarcity and missing values patterns was considered by [11]. It indicates that Graph Attention Network (GAT) can efficiently calculate data weight, reaching up to a level of 0.892 in predicting noise data, and there are constraints regarding the complexity of data and missing patterns. Finding out the Ordinary Differential Equations (ODEs) from the data via sparse data regression using the least amount of noise in the dataset was considered in [12]. Using the Sparse Identification of Nonlinear Dynamics (SINDy) in discovering interpretable models, the preliminary hard-thresholding process of SINDy by default discarded the actual dynamic terms of the system. This model is difficult when dealing with high-dimensional systems with multicollinearity in the design matrix. A model for recovering time-series dynamics from sparse data by using uncertainty learning for data augmentation was discussed in [13]. Applying the Gaussian model in predicting the noisy data results in highly accurate predictions. However, due to the sparsity of the data and error propagation, the method fails to accurately predict long-term time-series data. The improvement in trajectory reconstruction based on sparse data and noisy Global Positioning System (GPS) data for improving noise prediction was developed in [14]. With the help of the ProChunkFormer model, an effective model for predicting noise data level, which shows high noise prediction performance. Computational constraints lead to the setting of 64 as the maximum size of the GPS data. A stochastic method for training a machine learning algorithm that

mitigates any effects caused by noisy data was conducted in [15] using Metropolis-Hastings Noise Detection (MH-ND). It is revealed that approximately 28% of the total amount of data (which amounts to around 737 data points) is identified as noise, while about 72% are signals that can be used to train the ML algorithm.
Machine Learning techniques for noisy and sparse data for the accurate prediction of noise in sparse data were considered in [16]. The Multivariate Adaptive Regression Splines (MARS) method is employed for noise level prediction, achieving a Mean Squared Error (MSE) of 0% for both noise-free data and data with 10% noise, regardless of the training dataset size. However, its performance is constrained by the limited availability of scarce data. Noise level prediction in positive and unlabeled learning in large machine learning models by Uncertainty-Aware Neighbor Calibration (UANC) was examined [17], employing the UANC model for the uncertainty calculation. It achieved a noise level prediction accuracy of 0.96; however, the unlabeled data present in the hidden label noise, is limited.

3. Methodology

An uncertainty-aware learning model is developed to achieve accurate predictions in sparse and noisy data environments. Data are acquired from a Kaggle dataset containing stochastic signal observations and variables. Preprocessing includes Min-Max Normalization, and after that, features are extracted through ICA analysis based on signals' and statistics' properties. The CDA-EDNN technique can be effective in modeling uncertainties and hence improve prediction performance (Figure 1).

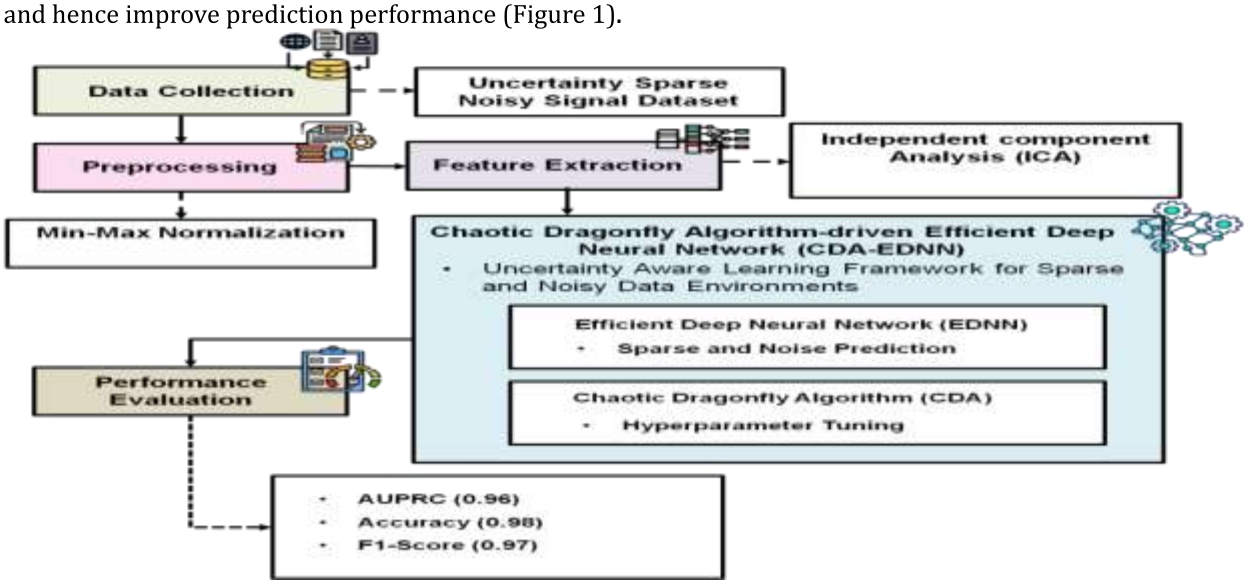

Figure 1: Methodology Workflow for Uncertainty Learning in Sparse and Noisy Data Environments

3.1 Data Collection

The dataset employed in this research is the Uncertainty Sparse Noisy Signal, which can be accessed via the Kaggle dataset from (https://www.kaggle.com/datasets/zara2099/uncertainty-sparse-noisy-signal-dataset/data). The dataset consists of a total of 4000 rows of data points and 16 columns, including attributes like Noise Level, Sparsity Ratio, Signal Variance, Sampling Frequency, Entropy Measure, Outlier Ratio, Interpolation Error, Adaptive Threshold, Aleatoric Uncertainty, Epistemic Uncertainty, MSE Value, and R2 Score. The dataset is split into training data 80% and testing data 20%. This is able to help build an appropriate model for uncertain learning-aware conditions. The dataset helps provide proper experimentation, since the inclusion of both artificial signals and signals makes it possible to experiment on aspects such as sparsity, noise, and uncertainty.

3.2 Data Preprocessing using Min-Max Normalization

Preprocessing involving Min-Max scaling is adopted in this research to address the issues associated with signal data that is either sparse data or noisy data before building a prediction model. This involves scaling the input variables in the range of [0, 1], min-max scaling to minimize biases resulting from differences in variable magnitude. It ensures the noise and sparsity environments, which are mathematically expressed in Equation (1).

$$Z_{new} = \frac{(Z - Z_{min})}{(Z_{max} - Z_{min})} \qquad (1)$$

Where Z refers to the value of the feature before normalization, the Sparsity Ratio. Z_{min} and Z_{max} represent the minimum and maximum values of that noise or sparsity feature, respectively, while Z_{new} denotes the new value of the normalized sparsity feature on a scale of [0,1].

3.3 Feature Extraction using Independent Component Analysis (ICA)

Feature Extraction is carried out through ICA to improve representation learning in the presence of sparse and noisy datasets. ICA separates mixed and latent source signals in an uncertain environment, thereby enhancing their interpretation while minimizing redundancies. It enhanced the robustness of sparsity and noise to achieve accurate uncertainty learning-aware predictions, and the observed signals are presented in Equations (2 and 3).

$$y(u) = Bt(u) \quad (2)$$

$$t(u) = Iy(u) \quad (3)$$

y(u) represents the measured signal vector obtained from various properties of the dataset, like noise signal strength, and B is the mixing matrix that contains the uncertainty learning-related information about noise signal transformation. t(u) denotes the vector containing source signals, which include Aleatoric Uncertainty and Epistemic Uncertainty as its two components. In Equation (3), I refers to the separation matrix used to recover the latent sources for accurate noise data.

3.4 Uncertainty Aware Learning Model in Sparse and Noisy Data Environments using Chaotic Dragonfly Algorithm-driven Efficient Deep Neural Network (CDA-EDNN)

The proposed Uncertainty-Aware Learning model incorporates the CDA-EDNN method to improve the prediction performance in scenarios involving sparsity and noise. The proposed learning model is intended to be capable of explicitly accounting for both types of uncertainties, with increased robustness in scenarios of missing and corruptive signals. CDA ensures stable network parameter optimization based on improved exploration and exploitation strategies. The EDNN component learns deep features necessary for the correct mapping of the underlying signal patterns.

Efficient Deep Neural Network (EDNN) for Sparse and Noise Prediction: The EDNN is developed with the increasing accuracy of prediction in scenarios where there is sparse, noisy, and ambiguous data. The model extracts robust features from the input signal while minimizing the effect of stochastic noise and ambiguity in the data. In this way, EDNN enhances generalization capabilities by concentrating on features that are important and ignoring features that are irrelevant, as expressed in Equation (4).

$$||M^1 - M^2||_L = \left(\sum_{l=2}^{m^1} \sum_{k=1}^{t_j^1} \sum_{l=1}^{t_{j-1}^1} \left|X_j^1[k,l] - X_i^2[k,l]\right|^O\right)^{1/O} \quad (4)$$

M^1 and M^2 represent the two noise neural networks, in which $X_j^1[k,l]$ and $X_i^2[k,l]$ correspond to the two noise weights of connections from noise neuron l to neuron k in layer j for both noise networks. O specifies the type of noise norm to be applied. $\Sigma_{l=2}^{m^1}$ represents the layer index, where summation starts from the noise second layer up to the last layer of the network, $\Sigma_{k=1}^{t_j^1}$ denotes the noise neurons in the current layer j, where t_j^1 is the number of noise neurons in layer j, and $\Sigma_{k=1}^{t_{j-1}^1}$ refers to the noise neurons in the previous layer j−1, where t_{j-1}^1 is the number of noise neurons in layer $j - 1$. This formula controls the EDNN optimization process to prevent any unnecessary alterations in the noise weight values, as expressed in Equation (5).

$$||M^1 - M^2||_1 = (|-0.5| + |1| + |1| + |-0.5| + |0.5| + |0.5| + 0 + 0) = 4 \quad (5)$$

The expression $||M^1 - M^2||_1$ specifies the tolerable difference in the structure of the original neural network. M^1 is the perturbed one, M^2 is the defined tolerance value, regulating the stability of the network. Here, $(|-0.5| + |1| + |1| + |-0.5| + |0.5| + |0.5| + 0 + 0)$ denotes the distance between both networks, and thus guarantees that updates in the weights are restricted within the optimization process. 4 refers to the mentioned condition, which enables the adaptation of the EDNN topology to learn from noisy and sparse datasets, as expressed in Equation (6).

$$||M^1 - M^2||_\infty = \max\{|-0.5|, |1|, |1|, |-0.5|, |0.5|, |0.5|, 0, 0\} = 1 \quad (6)$$

$||M^1 - M^2||_\infty$ indicate two deep neural network architectures with the same structure, but∞take the maximum value only (worst-noise case difference measure). $\max\{|-0.5|, |1|, |1|, |-0.5|, |0.5|, |0.5|, 0, 0\}$ represents the numerical values that indicate the differences between the respective maximum noise weight parameters of the two networks, and 1 ensures that all values become positive.

Hyperparameter Tuning with Chaotic Dragonfly Algorithm (CDA): CDA is used for optimizing the suggested CDA-EDNN model to make accurate predictions amidst environments with sparse and noise-filled datasets. This is achieved through improved exploration and exploitation using dragonfly swarm-based behaviors, including separation, alignment, cohesion, attraction, and distraction, with the help of dynamic and static swarming mechanisms. Chaotic mapping, particularly logistic chaotic mapping, brings about some level of randomness at the beginning stage that enhances the rate of convergence while at the same time ensuring that it does not converge on local optimal solutions. Through this combination, it enhances feature selection and makes the model stable, as expressed in Equations (7 and 8).

$$T_j = \sum_{l=1}^{M}(Y - Y_l) \quad (7)$$

$$B_j = \frac{\sum_{l=1}^{M} Y_l}{M} \quad (8)$$

Where T_j stands for the separation term, which is used for calculating repulsion in the swarm agents based on the feature set (for example, Noise Level and Outlier Ratio), and where $\Sigma_{l=1}^{M}$ stands for the summation of the entire neighbors up to M in the noise search space, while Y_l represents the difference between the current noise candidate solution and its noise neighbors with behavior. Y denotes the current agent position for balanced convergence in sparse data and noisy data environments, as expressed in Equation (8).B_j represents the alignment operator that signifies the noise behavior based on the overall directional movement of the swarm agents, as expressed in Equation (9).

$$D_j = \frac{\sum_{l=1}^{M} Y_l}{M} - Y \quad (9)$$

D_j is defined as the cohesive noise behavior, the mean value of the positions of all neighbors, denoted by $\Sigma_{l=1}^{M}$, where it stands for the summation of all neighbors lup to M in the noise search space. Y represents the current agent position for balanced convergence in sparse data and noisy data environments. Y_l denotes the difference between the current noise candidate solution and its noise neighbors with behavior, which is mathematically expressed in Equation (10).

$$G_j = Y^{best} - Y \quad (10)$$

G_j refers to the force acting towards the best noise solution, while Y^{best} is the global optimum found during the noise process of optimization, while Y stands for the current noise solution coordinates, as expressed in Equation (11).

$$F_j = Y^{worst} + Y \quad (11)$$

Where F_j is the level of evasion or noise avoidance associated with the worst noise solution, Y^{worst} is the worst noise solution in the population set, while Y is the current noise solution location being used to evade bad regions, as expressed in Equation (12).

$$\Delta Y_u + 1 = \left(tT_j + bB_j + dD_j + gG_j + fF_j\right) + \omega\Delta Y_u \quad (12)$$

ΔY_u stands for the velocity change of the swarm agent, tT_j denotes Noise Separation, bB_j refers to the Noise alignment, and dD_j represents the Noise cohesion. gG_j denotes the Noise attraction, and fF_j refers to the Noise distraction weight parameters. $\omega\Delta Y_u$ is the inertia part that ensures the continuation of the earlier direction of movement, which is expressed in Equations (13 and 14).

$$Y_u + 1 = Y_u + \Delta Y_u + 1 \quad (13)$$

$$y_u + 1 = \gamma y_u(1 - Y_u) 0 < \gamma \leq 4 \quad (14)$$

Where Y_u refers to the new noise position of the solution in the iterative process $(u + 1)$, y_u refers to the current noise position, while ΔY_u refers to the noise velocity increment that is calculated, and Equation (14) γ denotes the parameter that controls the noise strength of the chaos.

Hyperparameters of proposed CDA-EDNN: The proposed CDA-EDNN model for prediction on sparse and noisy data employs hyperparameter settings that allow it to be uncertainty aware while demonstrating high prediction performance. The model learning rate of 0.001 with the Adam optimization technique and mean square loss, and the other settings include the use of ReLU activation function, dropout value of 0.3. The model architecture comprises three hidden layers with a number of nodes [128, 64, 32], and contains 150 epochs of training with a batch size of 32.

4. Result and Discussion

The findings from the experiment clearly reveal that the newly developed uncertainty learning-based CDA-EDNN model performs significantly better in terms of accuracy when it comes to predicting data in noisy environments. The algorithm is capable of maintaining low error rates despite different amounts of stochastic noise and data scarcity. Apart from this, the newly proposed model considers AUPRC, Accuracy, and F1-score. The use of Chaos Optimization (CO) and adaptive learning ensures more accurate predictions.

Experimental Setup: The experimental evaluation to verify the efficiency of the CDA-EDNN model on the sparsity and noise datasets is conducted using a platform designed for performing deep learning efficiently. This platform consists of the following hardware: Intel Core i7 10th-generation processor, NVIDIA GTX 1660 Ti Graphics Processing Unit with CUDA capability, 16 GB of DDR4 memory, and 512 GB solid-state drive with 64-bit capability. As far as the software platform is concerned, it includes Windows 10/11 Operating System, Python 3.10, TensorFlow 2.x, NumPy, Pandas, Matplotlib, and CDA optimization library.

Exploratory Data Analysis: The analysis framework of the CDA-EDNN model is depicted in Figure 2 with an aim of achieving reliable predictions through proper modeling of uncertainties in the learning process. In Figure 2(a), an analysis of the dependencies among signals, statistical properties, and uncertainties is carried out to identify valuable features. The separation of the input signal into its fundamental constituents can be achieved using Figure 2(b), which allows for the suppression of noise and identification of patterns in the random data. The confidence level of predictions is evaluated in Figure 2(c), which classifies data points based on certainty, medium certainty, and uncertainty. Such an evaluation is an understanding of the accuracy and reliability of the prediction made by the models. Figure 2(d) explains how sample data points are classified based on sparsity, outlier presence, and interpolation error to segregate areas having different degrees of uncertainties. Together, all such evaluations help in providing reliable representations, uncertainty evaluations, and robust learning in noisy scenarios.

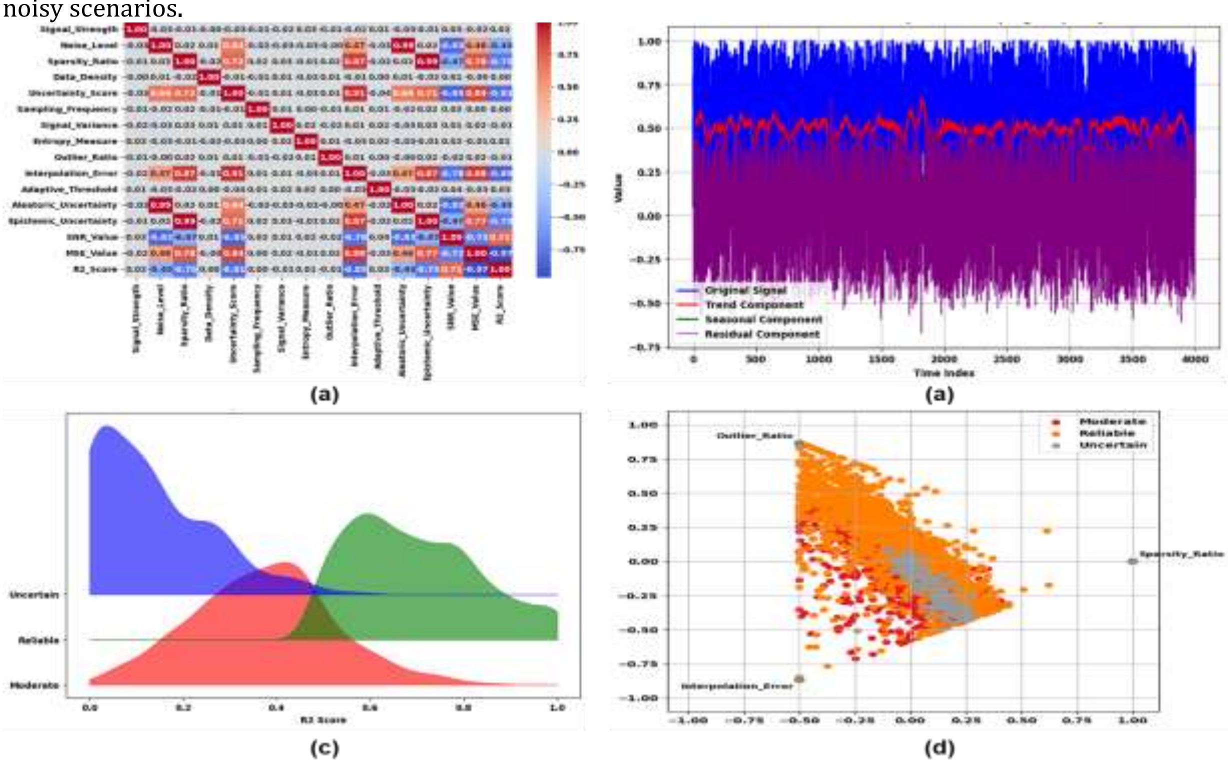

Figure 2: Data analysis of (a) Correlation analysis, (b) Dependency Analysis and Multicollinearity Assessment, (c) Signal Decomposition and Noise Filtering in Sparse and Noisy Data, and (d) Probabilistic Density Estimation for Uncertainty Learning and Reliability Assessment.

Evaluation Metrics: The performance measures used in the case of CDA-EDNN in sparse and noisy data prediction consist of AUPRC, Accuracy, and F1-score to evaluate the performance of the proposed technique under uncertainty. The AUPRC is adopted to check the reliability of the noise prediction by considering the separation of relevant noise outputs in an uncertain environment. The accuracy performance measure is used to measure how well the technique performs in general terms. In predicting the correctness of the noise prediction, the metric F1-score calculates the precision of minimizing false alarms resulting from noise and the recall of the recovery of latent patterns in sparse data.

Performance Evaluation: A comparative analysis between the existing method, UANC, and the proposed method, CDA-EDNN, in terms of sparse and noisy datasets, the results from Table 1 and Figure 3 above show that the proposed model performs significantly better than the benchmarked method for all the performance metrics considered. The UANC [17] method achieves a high accuracy of noise prediction. The proposed CDA-EDNN model with better AUPRC 0.96, Accuracy 0.98, and F1-Score 0.97 measures, which indicate better precision, recall balance, overall accuracy, and robustness under uncertainties, respectively.

Table 1: Comparison Analysis of Existing and Proposed Methods

Method	AUPRC	Accuracy	F1-Score
UANC [17]	0.94	0.96	0.95
CDA-EDNN [Proposed]	0.96	0.98	0.97

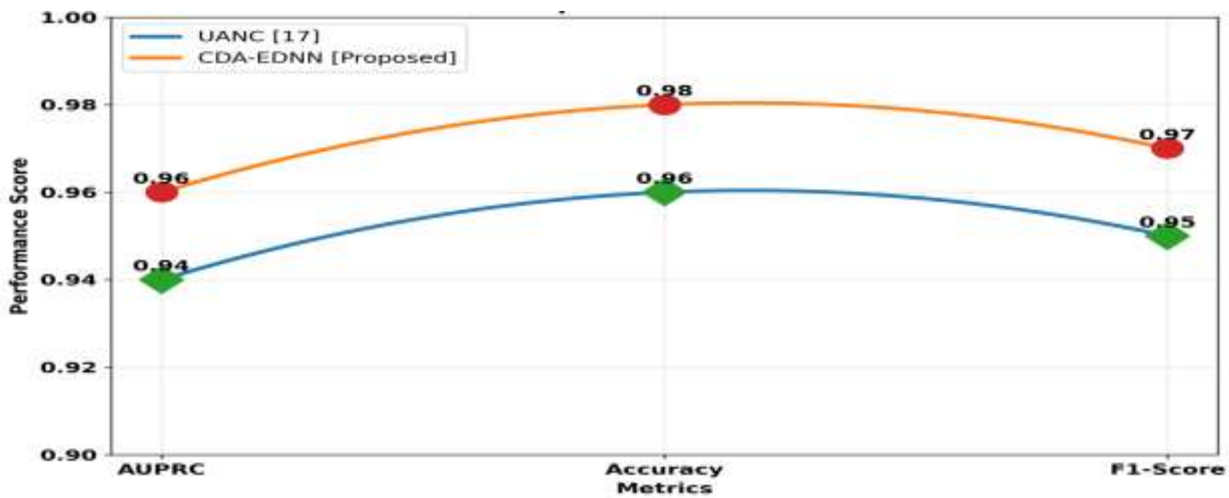

Figure 3: Comparative Analysis of the Uncertainty Learning in Sparse and Noisy Data Environments

Table 2 presents the existing UANC model and the proposed CDA-EDNN model were retrained using the Uncertainty Sparse Noisy Signal Dataset. The results suggest that both models need to be trained again, and the CDA-EDNN shows better results in all measurement metrics. The higher AUPRC, Accuracy, and F1-Score prove improved ability in dealing with sparse and noisy information. This shows the success of using the proposed model to capture the hidden patterns of signals in data and predict them successfully.

Table 2: Comparative Retrained Evaluation of Existing and Proposed Models on the Uncertainty Sparse Noisy Signal Dataset

Method	AUPRC	Accuracy	F1-Score
UANC [17]	0.95	0.97	0.96
CDA-EDNN [Proposed]	0.96	0.98	0.97

4.1 Discussion

A prediction model that works effectively in situations where there are insufficient and noisy datasets is proposed. However, the UANC [17] model is the problem of being affected by the lack of data, stochastic noise, and the incapacity of dealing with aleatory and epistemic uncertainties in such a way that the stability and accuracy of the predictions become affected. The methods based on uncertainty estimation do not utilize the information regarding the neighborhood, making the predictions more stable. To make optimal parameter tuning and architectures that enable scalability and stability when used in a large system. The proposed model overcomes all these challenges by employing both probabilistic learning and CDA algorithms. It is evident from the results that this model outperforms all the other models.

5. Conclusion

To design a learning model that can incorporate uncertainties in predicting outcomes with sparse and noisy data. Experimentally generated and simulated signal datasets with varying sparsity and stochastic noise are collected and preprocessed using normalization, outlier filtering, and noise-injection techniques to ensure realistic uncertainty modeling. Feature extraction is performed through ICA, enabling effective representation learning. The proposed CDA-EDNN method incorporated probabilistic modeling with adaptive interpolation, to achieve improved performance. Evaluation using AUPRC 0.96, Accuracy 0.98, and F1-Score 0.97 demonstrated greater stability and accuracy than current techniques. Nevertheless, the model with scalability problems in very large datasets and demands a larger computing cost. Future research concentrate on designing lightweight models that are instantly deployable and expanding them to multimodal uncertainty learning scenarios.

REFERENCES

1. Wiyono, S., Adji, T.B., Nugroho, H.A. and Wibirama, S., 2026. Global mean–centered correlation: A variance-reduced and theoretically grounded similarity measure for sparse data environments. Information Sciences, p.123502. https://doi.org/10.1016/j.ins.2026.123502
2. Nguyen, H., Pham, H., Nguyen, S., Van Linh, N. and Than, K., 2022. Adaptive infinite dropout for noisy and sparse data streams. Machine Learning, 111(8), pp.3025-3060. https://doi.org/10.1007/s10994-022-06169-w
3. Jiangzhou, D., Songli, W., Jianmei, Y., Lianghao, J. and Yong, W., 2024. DGRM: Diffusion-GAN recommendation model to alleviate the mode collapse problem in sparse environments. Pattern Recognition, 155, p.110692. https://doi.org/10.1016/j.patcog.2024.110692
4. Sun, Y., Lei, S., Zhao, Y., Wei, C., Yang, X., Han, X., Li, Y., Xia, J. and Cai, Z., 2024. Spatial distribution prediction of soil heavy metals based on sparse sampling and multi-source environmental data. Journal of Hazardous Materials, 465, p.133114. https://doi.org/10.1016/j.jhazmat.2023.133114
5. Miao, M., Yu, J. and Zhao, Z., 2022. A sparse domain adaption network for remaining useful life prediction of rolling bearings under different working conditions. Reliability Engineering & System Safety, 219, p.108259. https://doi.org/10.1016/j.ress.2021.108259
6. Bassetti, D., Pospíšil, L. and Horenko, I., 2024. On entropic learning from noisy time series in the small data regime. Entropy, 26(7), p.553. https://doi.org/10.3390/e26070553
7. Ahmed, N. and Schmidt-Thieme, L., 2023. Sparse self-attention guided generative adversarial networks for time-series generation. International Journal of Data Science and Analytics, 16(4), pp.421-434. https://doi.org/10.1007/s41060-023-00416-6
8. Park, J., Pao, G.M., Sugihara, G., Stabenau, E. and Lorimer, T., 2022. Empirical mode modeling: A data-driven approach to recover and forecast nonlinear dynamics from noisy data. Nonlinear Dynamics, 108(3), pp.2147-2160. https://doi.org/10.1007/s11071-022-07311-y
9. Nezhadettehad, A., Zaslavsky, A., Rakib, A. and Loke, S.W., 2025. Uncertainty-aware parking prediction using Bayesian neural networks. Sensors, 25(11), p.3463. https://doi.org/10.3390/s25113463
10. Reddy, B., Sujan, V.S., Sastry, C.C., Krishnaiah, J. and Jitpichitchai, S., 2025. A machine learning framework for long-term forecasting of spare part demand in end-of-life product scenarios. Scientific Reports, 16(1), p.1394.https://doi.org/10.1038/s41598-025-31171-2
11. Zhang, K., 2026. Missing pattern aware elastic modeling of sparse village data with dynamic evolution simulation. Discover Applied Sciences, 8(4), p.434. https://doi.org/10.1007/s42452-026-08403-9
12. Egan, K., Li, W. and Carvalho, R., 2024. Automatically discovering ordinary differential equations from data with sparse regression. Communications Physics, 7(1), p.20. https://doi.org/10.1038/s42005-023-01516-2
13. Wang, L., Shi, L., Hu, X., Song, W., Wang, Y. and Wang, Y., 2026. Maximizing utilization of sparse and noisy data for machine learning modelling of SPAC system dynamics. Environmental Modelling & Software, p.106926. https://doi.org/10.1016/j.envsoft.2026.106926
14. Liu, Y., Li, Q. and Kim, I., 2025. Enhanced trajectory reconstruction from sparse and noisy GPS data: A progressive chunked transformer approach. Communications in Transportation Research, 5, p.100200. https://doi.org/10.1016/j.commtr.2025.100200
15. Alzraiee, A.H. and Niswonger, R.G., 2024. A probabilistic approach to training machine learning models using noisy data. Environmental Modelling & Software, 179, p.106133. https://doi.org/10.1016/j.envsoft.2024.106133
16. Poulinakis, K., Drikakis, D., Kokkinakis, I.W. and Spottswood, S.M., 2023. Machine-learning methods on noisy and sparse data. Mathematics, 11(1), p.236. https://doi.org/10.3390/math11010236
17. Zulfiqar, M.I., Khalid, A., Chen, L. and Dhelim, S., 2025. Uncertainty-Aware Neighbor Calibration for Positive and Unlabeled Learning in Large Machine Learning Models. IEEE Transactions on Emerging Topics in Computational Intelligence. https://doi.org/10.1109/TETCI.2025.3594047