Convolutional Neural Networks for Robust Time Series Cleaning and Filtering

Hannia M. Hernandez[1], Alma Y. Alanis[1], Eduardo Rangel-Heras[1], Arturo Valdivia-G and Oscar D. Sanchez[1]

[1] Universidad de Guadalajara, Centro de Ciencias Exactas e Ingeniería, Guadalajara, Jalisco, México, alma.alanis@academicos.udg.mx, hannia.hernandez2934@alumnos.udg.mx, eduardo.rangel@academicos.udg.mx, arturo.valdivia@academicos.udg.mx, didier.sanchez@academicos.udg.mx

Abstract. Time series analysis involves identifying patterns, trends, and seasonal variations in data collected at specific intervals, with applications in finance, economics, and weather forecasting. A critical challenge in time series analysis is handling missing data, which can arise from non-response, errors, or equipment malfunctions. Missing data can bias results and reduce accuracy, necessitating effective techniques such as deletion, imputation (e.g., mean, regression, or multiple imputation), and algorithms capable of handling missing values natively. The choice of method depends on the nature of the missing data—Missing Completely at Random (MCAR), Missing at Random (MAR), or Missing Not at Random (MNAR). Advanced techniques like regressive, autoregressive, and vector autoregressive models have been employed for imputation in various domains, including healthcare and environmental monitoring.
This paper introduces a novel Deep Neural Network (DNN) approach for cleaning and filtering time series data, outperforming nine classical techniques such as linear interpolation, spline, and moving mean/median. The proposed DNN, a 1D convolutional autoencoder, processes noisy data directly without requiring preprocessing steps like outlier removal. Evaluated on temperature time series data from a compressor, the DNN achieved a Root Mean Square Error (RMSE) of 6°C, significantly lower than the best classical method (45°C). The DNN's ability to accurately impute missing data and follow trends, even with large gaps, demonstrates its robustness and superiority over traditional methods. This study highlights the potential of DNNs in enhancing time series data analysis and imputation.

Keywords: Time Series, Deep Neural Network, Fiters-Clean.

1 Introduction

Time series refers to data points collected or recorded at specific time intervals. The analysis of them implies identify patterns, trends, and seasonal variations in the data. It is particularly useful in fields like finance, economics, and weather forecasting. The absence of data points in a data set can occur for various reasons, such as non-response, data entry errors, or malfunction equipment. It is essential to handle missing

https://doi.org/10.61728/AE20254964

data because it can bias results and reduce the accuracy of analysis. Techniques to handle missing data include:

- **Deletion**: Removing records with missing data, which can be done listwise (removing entire records) or pairwise (removing only the missing data points).
- **Imputation**: Filling in missing data using statistical methods like mean or median imputation, regression imputation, or more advanced techniques like multiple imputation [1–6].
- **Using Algorithms**: Some machine learning algorithms can handle missing data natively.

It is crucial to select the appropriate method to handle missing data based on its nature whether it is Missing Completely at Random (MCAR), Missing at Random (MAR), or Missing Not at Random (MNAR) [4,5].

Time-series techniques are used to fill gaps in data. For example, Murad et al.[5] employed time-dependent covariates in a Cox model with Multiple Imputations by Chained Equations to address missing data related to diabetes and some cancers, showing feasibility only for small groups of contiguous missing data. Regressive, autoregressive, vector autoregressive (VAR), and autoregressive moving average (ARMA) models have also been used for imputation. Bashir and Wei[7] developed a VAR-based algorithm for multivariate time-series missing data, tested on ECG signals with 10% and 20% Missing Completely at Random (MCAR) datasets, requiring stationary time-series. Zhang[3] used a regressive model to impute clinical data in blood pressure and lactate, preserving relationships between missing values and variables. Liu & Molenaar[6] recovered missing electrodermal activity data with a VAR model, fitting it to complete data and examining directional influence between child and therapist. Dunsmuir & Robinson[8] estimated stationary time-series data with ARMA models for pollution levels with minimal missing data. Anava et al.[9] predicted time-series online with AR models, focusing on time-series prediction with missing data. Pedreschi et al.[10] evaluated methods for missing values in gel-based proteomics data, concluding BPCA showed the best results among studied methods. Choong et al.[11] proposed a new algorithm.

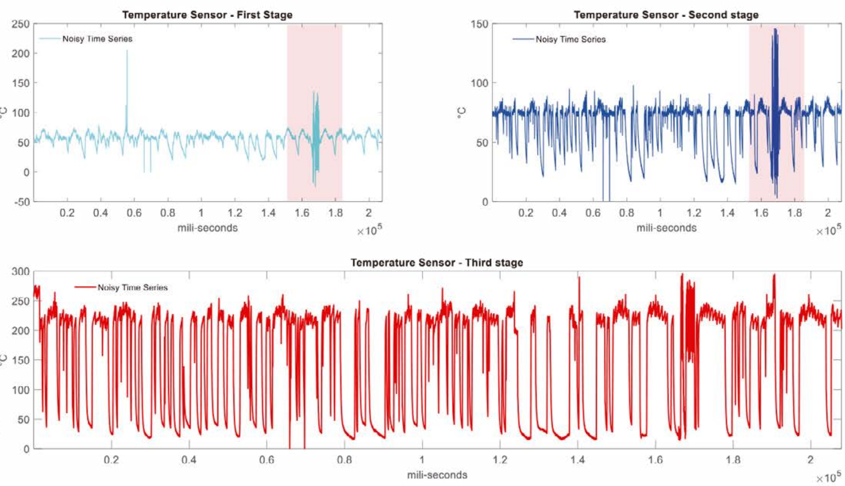
In this paper, we propose a novel Deep Neural Network (DNN) approach for cleaning and filtering time series data. To validate the effectiveness of our method, we conduct a comprehensive comparison with nine classical techniques, demonstrating the superior performance of our DNN-based solution.

2 Experiment Data Preparation

To evaluate the performance of the DNN, we utilized three temperature time series records obtained from a compressor, each comprising 207,887 data points. The time series were contaminated with noise within the range of 165,984 to 170,141, corresponding to 4,158 noisy data points. For the implementation of classical models, the noisy data segments were removed to ensure accurate analysis. In contrast, the DNN approach does not require the removal of noisy data, showcasing its robustness and ability to handle contaminated time series directly.

Fig. 1 displays the temperature across the various stages of the compressor, with the red rectangle highlighting the noisy data.

Fig. 1. Time Series – Three Stage Compressor Temperatures



3

3.1 Classic Method to Filter Data

The classic method implemented here are listed below:

1. **Previous**: Fills missing values using the last observed value.
2. **Next**: Fills missing values using the next observed value.
3. **Nearest**: Fills missing values using the closest observed value (either previous or next).
4. **Linear Interpolation**: Estimates missing values by drawing a straight line between neighboring points.
5. **Spline**: Uses smooth polynomial functions to interpolate missing values.
6. **PChip**: Piecewise Cubic Hermite Interpolation—preserves shape and monotonicity of the data.
7. **Makima**: Modified Akima interpolation—smoother than linear but less oscillatory than spline.
8. **Moving Mean**: Fills missing values using the rolling average of nearby data points.
9. **Moving Median**: Fills missing values using the rolling median of nearby data points, reducing outlier impact.

3.2 Deep Neural Network

Deep neural networks can be used to classify data and filter out data that does not meet certain criteria and imputation of missing data. For example:

- **Spam detection**: A neural network can classify emails as spam or not spam, automatically filtering out unwanted ones.
- **Content filtering**: On social networks, DNNs can identify and filter inappropriate or malicious content.
- **Imputation of missing data**: GANs can learn to fill in missing values in datasets.
- **Synthetic data generation**: Create clean and realistic data to train other models.

The presenting network is a 1D convolutional autoencoder with an encoder-decoder architecture, specifically designed to process one-dimensional signals. The input consists of a multi-channel sequence, which is passed through three convolutional layers in the encoder. These layers progressively reduce the dimensionality of the signal based on the number of down sampling levels. The decoder then reconstructs the signal using three transposed convolutional layers, gradually increasing the number of filters in each layer until the original structure is restored. A final transposed convolutional layer produces the output, matching the number of channels in the input. This architecture is well-suited for tasks such as noise removal, data compression, and synthetic signal generation. The network architecture is summarized in Table 1.

Table 1. Architecture of the Proposed Deep Neural Network.

Layer	Type	Filters/Channels	Filter Fixed
Input	Sequence input layer	Number of channels	*
Convolution 1D (Encoder) 1	Convolution 1 Layer	numFilter * (numDownSamples + 1 – i)	17
Convolution 1D (Encoder) 2	Convolution 1 Layer	numFilter * (numDownSamples)	17
Convolution 1D (Encoder) 3	Convolution 1 Layer	numFilter * (numDownSamples - 1)	17
Convolution 1D (Decoder) 1	TransposedConv1D Layer	numFilters	17
Convolution 1D (Decoder) 2	TransposedConv1D Layer	2*numFilters	17
Convolution 1D (Decoder) 3	TransposedConv1D Layer	3*numFilters	17
Output	TransposedConv1D Layer	numChannels	17

4 Results and conclusions

Table 2 presents the Root Mean Square Error (RMSE) in degrees Celsius (°C) for classic forecasting methods and Deep Neural Network (DNN) models. The best result among the classic methods is 45 °C, achieved by the moving mean technique, while the DNN achieves a significantly lower RMSE of 6 °C. These results demonstrate that classic methods are unable to compete with the performance of DNN-based approaches.

Table 2. Root Mean Square Error (RMSE) in Degrees Celsius (°C) for Classic Forecasting Methods and Deep Neural Network (DNN) Models

Previous	Next	Nearest	Linear	Spline	PChip	Makima	Moving Mean	Moving Median	DNN
95	80	85	86	404	86	87	45	46	6

Fig. 2. Imputed Data (Moving Mean) vs. Original Data

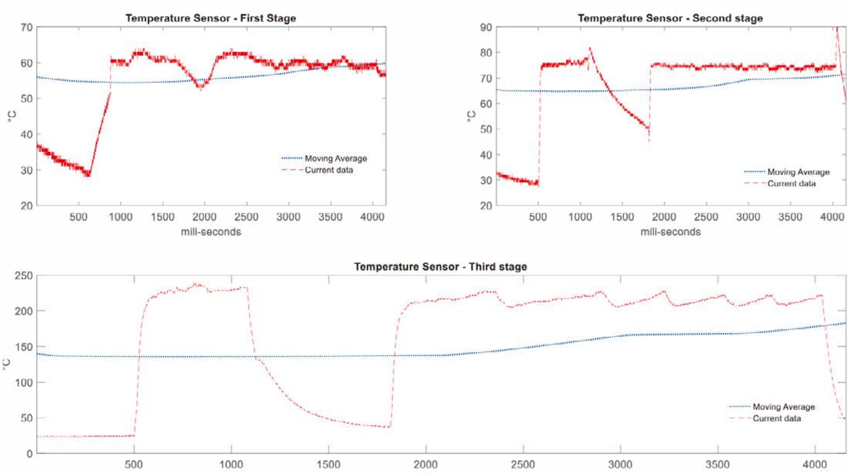
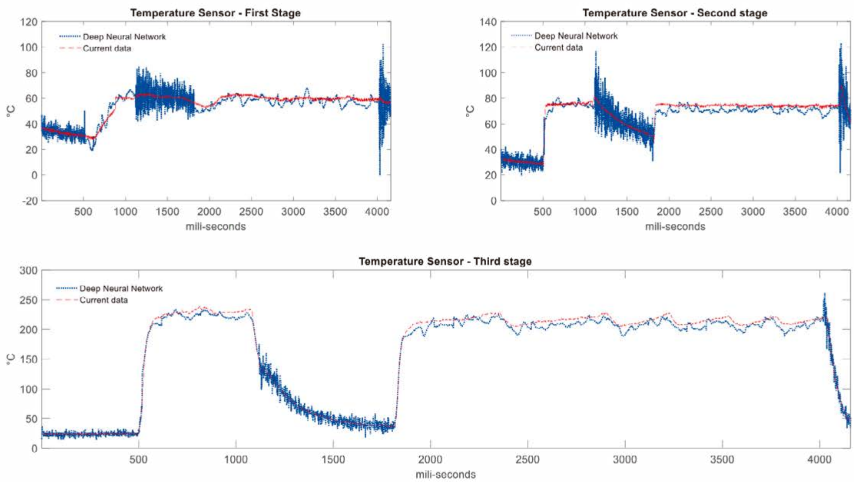


Figure 2 illustrates the results of data imputation using the classic method. To apply this method, outlier values must first be identified and replaced with imputed values. In this case, the moving mean—which yielded the best RMSE among classic methods—was used for imputation. However, due to the large volume of outlier data, the imputed data behaves more like an average, deviating from the original data's characteristics.

In contrast, Figure 3 demonstrates the DNN's ability to accurately follow the data trend, even in the presence of significant data gaps. A key advantage of the DNN over classic methods is its ability to impute data without the need for prior outlier detection. Unlike classic approaches, the DNN does not require preprocessing steps such as z-score normalization or other outlier-cleaning techniques, streamlining the imputation process.

Fig. 3. Imputed Data (DNN) vs. Original Data



References

1. Dan, E.L.; Dinsoreanu, M.; Muresan, R.C. Accuracy of Six Interpolation Methods Applied on Pupil Diameter Data. In Proceedings of the 2020 22nd IEEE International Conference on Automation, Quality and Testing, Robotics - THETA, AQTR 2020 - Proceedings; Institute of Electrical and Electronics Engineers Inc., May 1 2020.
2. Noor, N.M.; Al Bakri Abdullah, M.M.; Yahaya, A.S.; Ramli, N.A. Comparison of Linear Interpolation Method and Mean Method to Replace the Missing Values in Environmental Data Set. In Proceedings of the Materials Science Forum; Trans Tech Publications Ltd, 2015; Vol. 803, pp. 278–281.
3. Zhang, Z. Missing Data Imputation: Focusing on Single Imputation. *Ann Transl Med* **2016**, *4*, doi:10.3978/j.issn.2305-5839.2015.12.38.
4. Junger, W.L.; Ponce de Leon, A. Imputation of Missing Data in Time Series for Air Pollutants. *Atmos Environ* **2015**, *102*, 96–104, doi:10.1016/j.atmosenv.2014.11.049.
5. Murad, H.; Dankner, R.; Berlin, A.; Olmer, L.; Freedman, L.S. Imputing Missing Time-Dependent Covariate Values for the Discrete Time Cox Model. *Stat Methods Med Res* **2020**, *29*, 2074–2086, doi:10.1177/0962280219881168.
6. Liu, S.; Molenaar, P.C.M. IVAR: A Program for Imputing Missing Data in Multivariate Time Series Using Vector Autoregressive Models. *Behav Res Methods* **2014**, *46*, 1138–1148, doi:10.3758/s13428-014-0444-4.

7. Bashir, F.; Wei, H.L. Handling Missing Data in Multivariate Time Series Using a Vector Autoregressive Model-Imputation (VAR-IM) Algorithm. *Neurocomputing* **2018**, *276*, 23–30, doi:10.1016/j.neucom.2017.03.097.
8. Dunsmuir, W.; Robinson, P.M. Estimation of Time Series Models in the Presence of Missing Data. *J Am Stat Assoc* **1981**, *76*, 560–568, doi:10.1080/01621459.1981.10477687.
9. Anava, O.; Hazan, E.; Zeevi, A. Online Time Series Prediction with Missing Data. In Proceedings of the International Conference on MachineLearning, Lille, France; Lille, France, 2015; pp. 1–9.
10. Pedreschi, R.; Hertog, M.L.A.T.M.; Carpentier, S.C.; Lammertyn, J.; Robben, J.; Noben, J.P.; Panis, B.; Swennen, R.; Nicolai, B.M. Treatment of Missing Values for Multivariate Statistical Analysis of Gel-Based Proteomics Data. *Proteomics* **2008**, *8*, 1371–1383, doi:10.1002/pmic.200700975.
11. Choong, M.K.; Charbit, M.; Yan, H. Autoregressive-Model-Based Missing Value Estimation for DNA Microarray Time Series Data. *IEEE Transactions on Information Technology in Biomedicine* **2009**, *13*, 131–137, doi:10.1109/TITB.2008.2007421.