Associative memory based SLFM method for estimating suspect SCADA data -NR Grid case study

S. Naresh Ram
POSOCO
New Delhi, India
naresh.ram@posoco.in

Sukumar Mishra
IIT Delhi
New Delhi, India
sukumar@ee.iitd.ac.in

S.Lakra
POSOCO
New Delhi, India
somara.lakra@posoco.in

N.Nallarasan
POSOCO
New Delhi, India
nallarasan@posoco.in

***Abstract*—We considered the problem of estimating the suspected SCADA values of POC feeders, for the drawl calculations of states under unknown network topology. Our main goal is to estimate the real-time measurements of suspected/ corrupted feeder value (due to failure of fibre optic cables) through a database of Meter data readings of the respective feeders. To achieve that, we proposed an interpretable data-driven model based on associative memory technique, Soft Lookup False data matching (SLFM) through DTW (Dynamic Time warping) . Numerical experiments conducted on State X of the Northern region with real-time data for four different cases and compared the results with(LSTM-Autoencoder) model.The experimental results demonstrate the performance of the proposed model.**

***Keywords*—Autoencoder,Dynamic Time warping, Energy Management system, False Data, India,Long Short Term Memory, SCADA,**

NOMENCLATURE

Number sets	
$\mathbb{R}$	Real Numbers
Abbreviations	
GAN	Generative Adversarial Network
$LSTM$	,Long Short Term Memory
PoC	Point of Connection
Data Representations	
ϕ: $D_S \mapsto D_M$	Mapping function
$A, B \ldots N \ \in P_g$	PoC points of Grid-Stae X
$Ax, Bx \ldots Nx \ \in P_x$	P0C points of State X-Grid
$z_{j,t}^{m} \in M^{nxd}$	Meter(True) data
$z_{j,t}^{s} \in S^{nxd}$	Measurement(SCADA) data
$z_{j,t}^{sc} \in C^{kxd}$	Corrupted Measurement data
D_M	Database of Meter data
D_S	Database of Measurement data

I. INTRODUCTION

Power is among the most critical components of infrastructure, crucial for the economic growth and welfare of nations. The existence and development of adequate infrastructure is essential for sustained growth of the Indian economy.

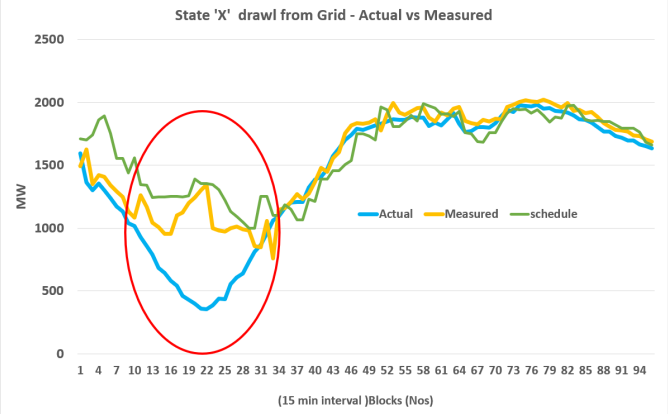


Fig. 1: State X drawl Actual vs Measure vs Schedule

India's power sector is one of the most diversified in the world, consisting more than 600 generating stations, more than 30 transmission licensees , 70 odd distribution licensees , 2 power exchanges, 40 odd trading licensees, National level Load dispatchers at the center, Regional level load dispatchers in each of the five regions and State level load dispatchers in each of the 28 States. The total installed generation capacity is 383 GW (as of April 2021), out of which 54% is from coal, about 13% Hydro, 24% Renewable, 7% Gas, and 2% Nuclear.

Recently , an event (depicted in figure 1) occurred in Northern Region of Indian Grid, due to inclement weather conditions in the state utiity network (X), its drawl deviated from schedule to a large extent. The state LDC took corrective action to control its under drawl from grid of quantum 300MW only (based on available SCADA data and system operator's experience), whereas, its actual deviation was around 800MW.

However, the state X was unaware of its actual load during that period due to suspect SCADA measurements, which led to load generation imbalance of around 800MW. To avoid such cases, there is a need for the tool which estimates the individual feeder data even under suspect measurement and topology.

A. Related Works

The missing data imputation can be done through a)Naive method, which involves comparing the erroneous/suspect data with previous or average of historical data. If the difference is above the threshold, then the corrupted data be replaced with historical average values. This method failed whenever missing occurs for a longer period. b) Multivariate time series [1] [2] [16],electrical loads typical random and volatile, which makes missing imputation is difficult with classic multivariate linear systems. c) Low rank approximation [3] models captures the temporal relations but fails in time dependencies. Moreover, not every big data be represented as low rank approx.

On the other hand, estimating suspected data comes under the purview of data driven models of defending False Data Injection Attacks(FDIA) [6] , Matrix Completion [7] and Load Forecasting with missing data [8] [10].In all the areas of problem, applying Machine learning based algorithm of LSTM/Recurrent neural network with Auto encoder achieved State Of The Art results.

Here LSTM layer helps in transforming non stationary data to stationary and makes the model learns both spatial and temporal relations between them. Hence, LSTM based models better than classic models [4]. However, the problem of estimating the uncertainty in time series data remains an open question.

B. Main contributions of the paper

Due to inexplicable, the efficacy of machine learning for time series prediction has been questioned for its guarantee and robustness. We proposed a method(SLFM) based on associative memory search technique, which is inspired from mainly [11] & [12], that has high interpret ability and guarantee in convergence to unique point. Moreover, this paper highlights the use of Dynamic Time Warping (DTW) method for time series matching as similarity measure for perfect recall.

II. Soft Lookup based False data Matching (SLFM) model for estimating true data from corrupted measurement data

A. Problem formulation

Generalized modeling for predicting/estimating the true data from the suspected measured time series data of length d through the logical grouping of temporal information of feeders data ($j = 1 : n$) of state X i.e, let the K feeders of P_g and its corresponding other end in P_x has corrupted measurement data be denoted as C^{kxd} and $C \subset S \subset D_S$. For an instant T, the true measurement ($z^{*sc}_{j:k,T}$) of k suspected feeders data ($z^{sc}_{j:k,T}$) be estimated through generalized model as shown below.i.e finding a perfect matching pair to suspect data point from mapping set ϕ of databases through a loss function h.

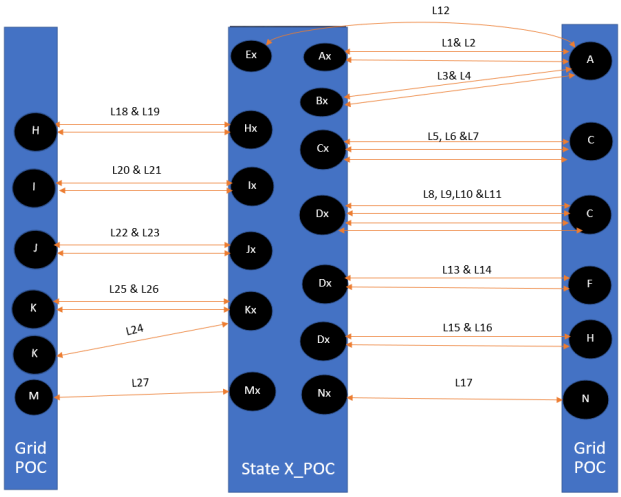


Fig. 2: State X poc feeders connectivity with Grid poc

$$z^{*sc}_{j:k,T} = \inf_{Z \in \psi} h(z^{sc}_{j:k,T}, Z) \text{ here h is loss function} \quad (1)$$

In General, under no external influences the flow pattern remains similar to very previous day pattern, If there exists any forced disturbances like fiber optical failure, weather disturbance, cyber attack etc their signature is unique to classify. With the domain knowledge, the experience operator able to bypass few of such cases and makes the grid secure and reliable.

Inspiring from human intellect, a model proposed for triggering proper recall through database of stored meter data (D_M) and Measurement data (D_S) with partial/noisy measurement (C) using associative memory technique.

B. DTW- a perfect measure for time series data

In the domain of sequence data analysis(time series), both Minkowski and Mahalanobis distances fail to reveal the true similarity between two targets. Dynamic Time Warping (DTW) [13] has been proposed as an attractive alternative. Hence DTW is widely used for similarity measure between time series data. Moreover,DTW provides more meaningful discrepancy measurements between two signals than other distance measures [14].

$$DTW(z^s, z^{sc}) = \min_{\pi} \sum_{(i,j)\in\pi}^{n} \sqrt{||z^s_i - z^{sc}_j||} \quad (2)$$

where π is time series align paths $[\pi_1 \dots \pi_p]$

The above objective function be solved through dynamic programming as depicted in [15]

$$\zeta_{i,k} = ||z^s_{j,i} - z^{sc}_{j,k}|| + \min\{\zeta_{i-1,k}, \zeta_{i,k-1}, \zeta_{i-1,k-1}\}$$
$$DTW(z^s_{i,j}, z^{sc}_{k,j}) = \sqrt{\zeta_{i,k}}$$

C. Soft Lookup False data Matching model- recall method

We now introduce new SLFM method for recalling the true measurement $z_j^m \subset M$ through a query (corrupt data) $(z_j^{sc}) \subset C$.

Proposed method mainly consists of two steps.

- Recursively multiplying(associatively) the probability distribution of similarity function ($S(D_S, z_j^{sc})$) with data base (D_S) (memory) to get nearest match of suspect measurement within database.
- Choosing the corresponding meter data from D_M for the converged signature of corrupted data through the mapping table (ϕ) between D_S & D_M.

The probability distribution of similarity points over Database (D_S) be achieved through $softmax()$ function.

Moreover, we claim that under well separated/ independent stored time series data, there exists a unique retrieval point through SLFM method for the suspected Measurement data point and the proof of such claim in Appendix 1.

$$z_{j,t}^{sc} = D_S^T.(1 - \frac{exp(S(D_S, z_{j,t-1}^{sc}))}{\sum_k exp(S(D_{Sk}, z_{j:n,t-1}^{sc}))})$$

$$= D_S^T.(1 - softmax(S(D_S, z_{j,t-1}^{sc}))) \quad (3)$$

$\sim$ converges to $z_j^s \in S \subset D_S$

Here $S(D_S, z_j^{sc}) = \sum_{i,k} DTW(z_{i,j}^s, z_{k,j}^{sc})$

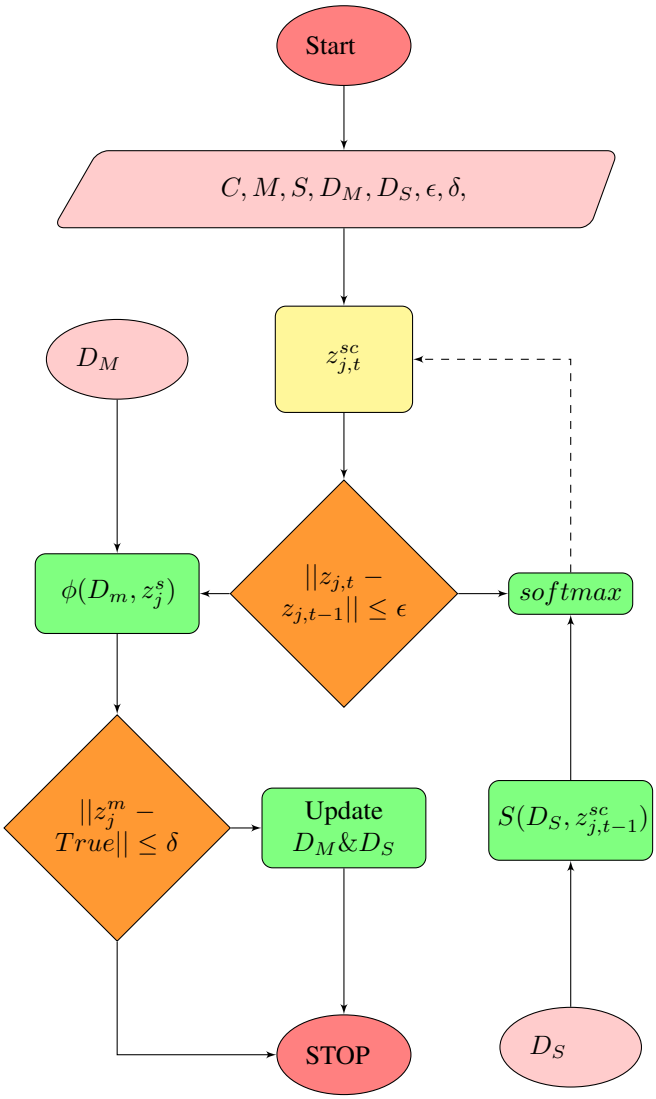
$$z_j^m = \phi(D_M, z_j^s) \text{ where } \phi\colon D_S \mapsto D_M \textbf{ one to one} \quad (4)$$

Here $z_{j:k}^{*m} \subset z_{j:n}^m$ is the perfect recall for the corresponding corrupted data $z_{j:k}^{sc}$.

III. Proposed Model (SLFM) Flow Chart

The flow chart of the process model is shown above, here, the loop runs until the threshold ϵ reaches. However, for well separated case, not more than two loops required for attaining convergence. Here, $true$ data means the true meter data of the suspected feeder data.

In our case, the true validated meter data generally available after 7 days, hence the log for the error be maintained to update the data base. If the estimated and error below the threshold (our case, we took 100) then no need for any change/modify the databases (D_M & D_S)



IV. Results

A. Data setup and analysis

State x net drawl is being calculated from the interconnected 26 nos of feeders, the drawl details have been shown in fig 4. For estimating the corrupt data using LSTM and SLFM, all 24nos of feeder's drawl (MW) meter and as well as its respective MW values in SCADA was collected for 3 months of 15-minute interval. The collected data clustered into 40 different groups as per their DTW distances(between Measurement and Meter data).

From the heat plot of meter data in fig 3, noticed that all the feeders data are independent of each other. Whereas, dependencies/high correlation observed in between parallel feeders only i.e (feeders 7,8,9,&10). Moreover, from the histogram plot of fig 3, understood that the feeders of 2,12,13,24,25 are almost ideally loaded most of the time.

Further, with SCADA (Measurement data) plots fig 4, noticed that feeder 16 data has zero value throughout the

collected period. From fig 5, on an average 10 % of SCADA data being corrupted/ stuck at a particular value.

All the simulations and test cases are done on python, for the LSTM autoencoder we used Pytorch model run on NVIDIA GeForce GTX 1050Ti and having Ram 8GM of windows 10 PC.

TABLE I: Cases of faulty measurement data

CaseNos	Conditions	Suspect feeders
1	Parallel feeder data suspect	18, 19
2	False data injection	8,9,10,11
3	Unknown weather disturbance	22,23,25,26
4	Normal operating condition	16

We selected four different cases as shown in Table 1, We compared the results of proposed model (SLFM) with LSTM auto encoder in every case and the results discussed comprehensively.

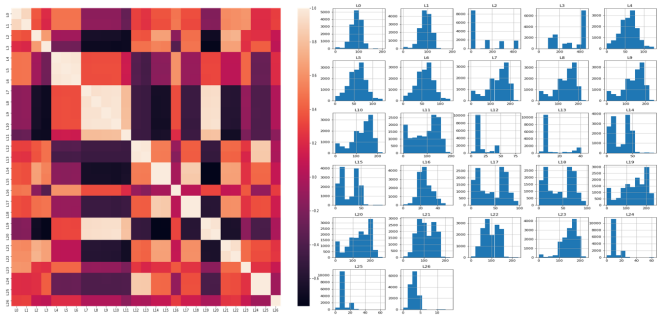


Fig. 3: Meter State X feeder's drawl Heat plot and histogram

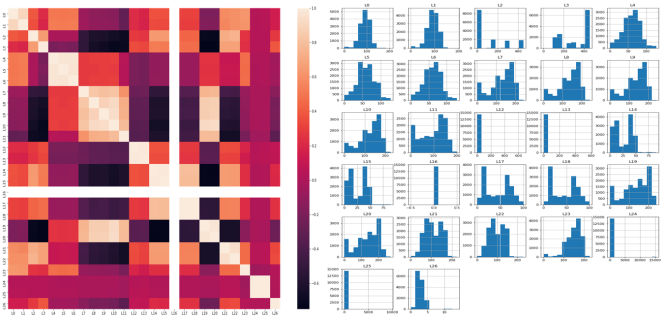


Fig. 4: SCADA State X feeder's drawl Heat plot and histogram

B. Testing results and discussions

We conducted the tests on the above depcted four cases and the results are shown, for case1 , the proposed method is robust to suspect/stuck at any particular value of any feeder data, the estimated through SLFM is the smart lookup to nearest DTW distance. The distance found to be 19, retrieved the corresponding data of nearest distance and the error wrt to true is almost zero. Whereas it's a known fact, the decisions/ estimated with ML techniques are sensitive to noise/disturbance, hence the error incurred with LSTM is more than anticipated which is shown in fig 6.

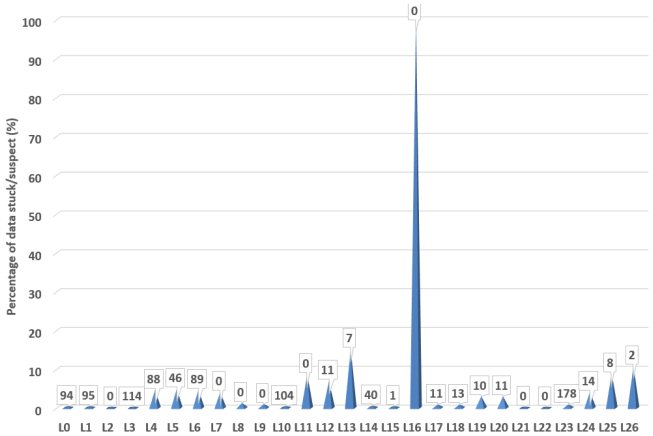


Fig. 5: Percentage of Measured feeder's data stuck at specific value of State X

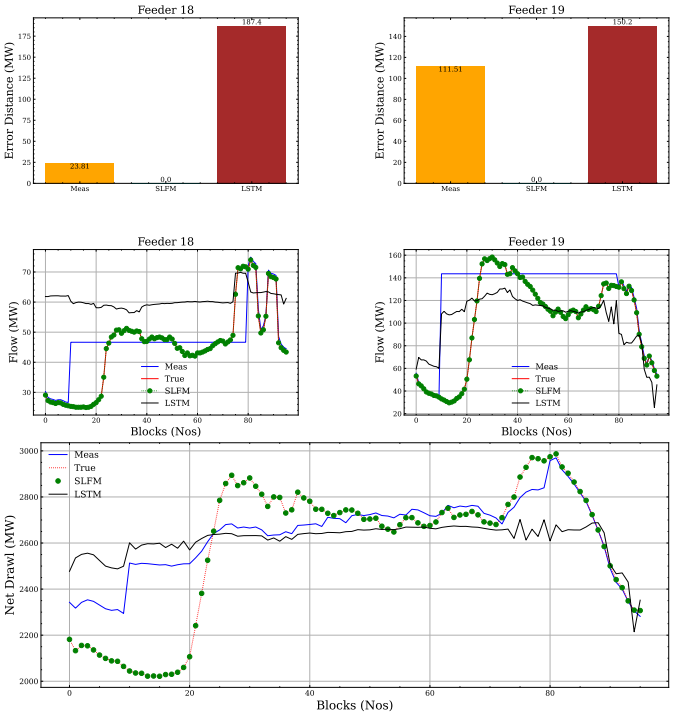


Fig. 6: case-1 comparison of SLFM with LSTM- BAR plot

For case 2, we fed a random (gaussian distribution) data injection to particular parallel feeders, again our method SLFM achieved perfect recall with the minim distance 49, fig 7.

For case 3, unknown disturbance, where that incident signature is not in the stored database, hence the perfect recall is not possible. However, the model (SLFM) achieved better and in line with the standard LSTM Auto-encoder technique , fig 8.

For case 4, Real-time SCADA data collected and fed to the model, and it perfectly captures the true flow and even identifies and estimated the feeder 16 data with minimum error, fig 9-10.

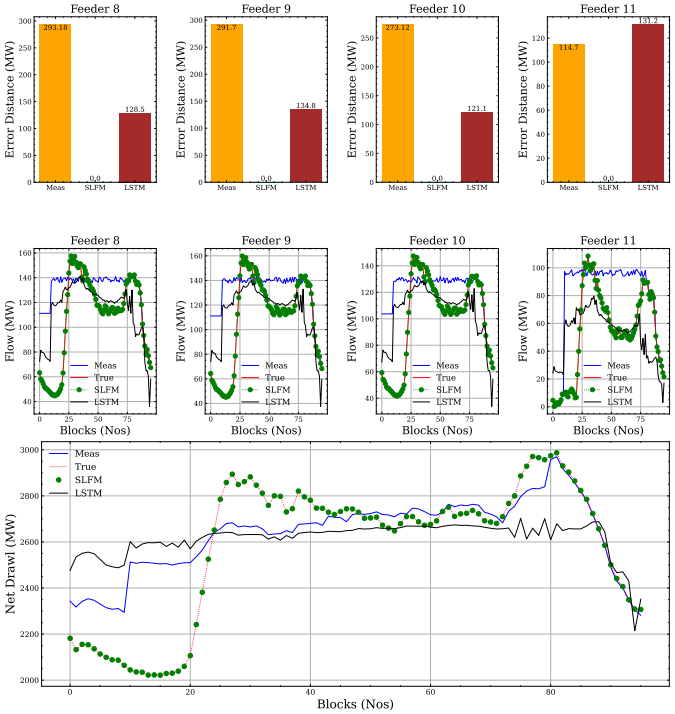


Fig. 7: case-2 comparison of SLFM with LSTM- Line plot

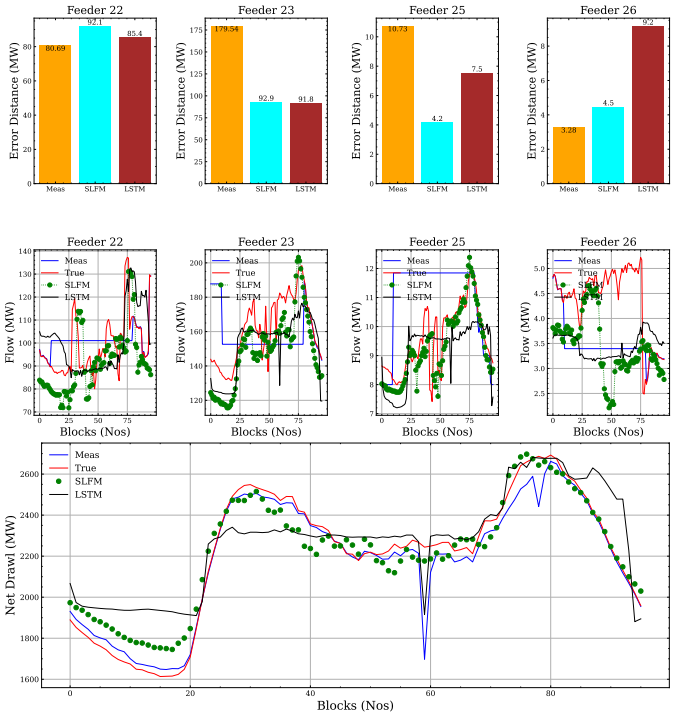


Fig. 8: case-3 comparison of SLFM with LSTM- Line plot

V. CONCLUSION

In this paper, we motivated the application of DTW in similarity measurements between time series data and proposed (SLFM) an associative memory technique model for the time series corrupt data estimation problem. We further compared the different case results with the machine learning-based algorithm (LSTM+Autoconder) model. The proposed model defeats the ML models not only in accuracy but also in interpretability, robustness and guarantee in convergence. Further, we are in the process of combining GAN model with SLFM such that the gap between database points be reduced by filling it with generative points.

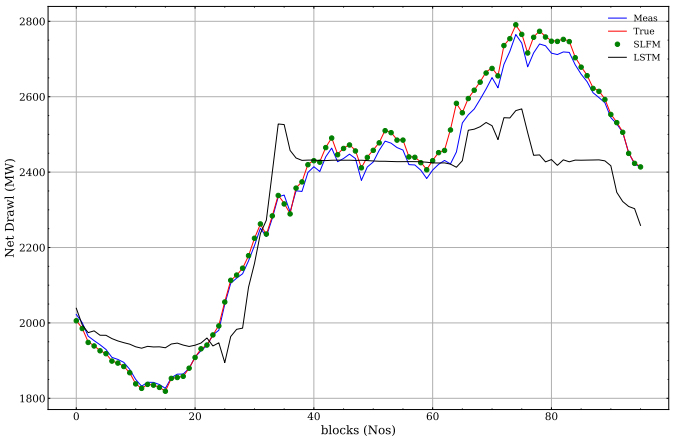


Fig. 9: case-4a comparison of Drawl SLFM with LSTM- Line plot

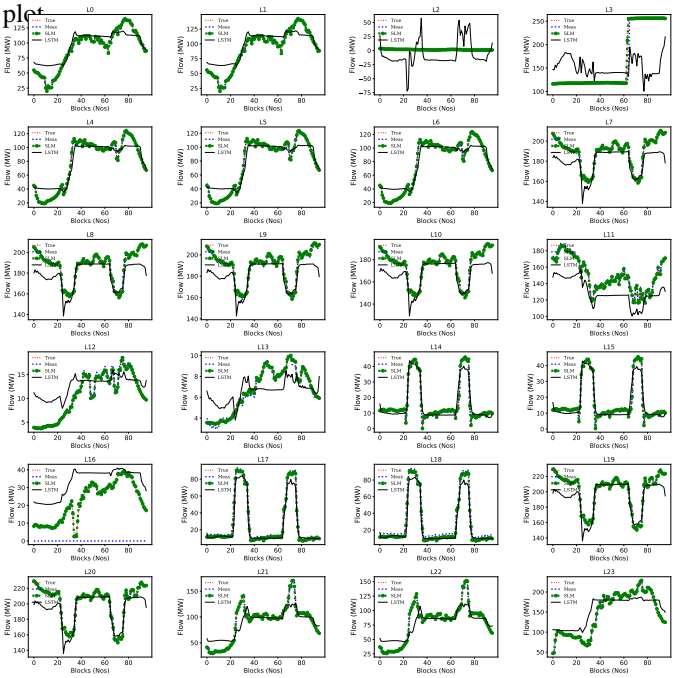


Fig. 10: case-4b comparison of flow SLFM with LSTM- Line plot

VI. ACKNOWLEDGMENT

The authors acknowledge the support and inspiration received from the POSOCO management in publication of this paper. The authors are grateful to their NRLDC colleagues for their liberal contributions for enriching the content in this paper. The views expressed in this paper are that of the authors and may or may not represent the views of the organization to which they belong.

APPENDIX-1

A. proof that associative memory retrieves data- Under well separated

Consider the space-time series X in R^{nxd} where suspected drawl feeder $x_i \in X$ and the database of stored patterns be $D_i \in D^{Lxnxd}$ and the L patterns are uncorrelated.

$$x_j^{t+1} = x_j^t + \frac{\partial x_j^t}{\partial {x^{t-1}}_j}(x_j^{t-1} - x_j^{t-2})$$

$$\frac{\partial x_j^t}{\partial x_j^{t-1}} = D^T(diag(P^{t-1}) - P^{t-1}(P^{t-1})^T)\frac{\partial S(D_k, x_j^{t-1})}{\partial x_j^{t-1}} \tag{4}$$

where $P_i^{t-1} = softmax(S(D_i, x_j^{t-1})) = \frac{e^{(S(D_i, x_j^{t-1}))}}{\sum_k e^{(S(D_k, x_j^{t-1}))}}$

$$= \frac{1}{1 + \sum_{k!=j} e^{(S(D_k, x_j^{t-1}) - S(D_i, x_j^{t-1}))}}$$

$$= \frac{1}{1 + \sum_{k!=j} * e^{-\delta_k}} \sim \frac{1}{1 + (N-1) * e^{-\delta_c}}$$

$\because$ datapoints are well separated $\delta_c >> 0,\ P_i^{t-1} \sim 1$

$$P_j^{t-1} = softmax(S(D_j, x_j^{t-1})) = \frac{e^{(S(D_j, x_j^{t-1}))}}{\sum_k e^{(S(D_k, x_j^{t-1}))}}$$

$\sim 0\ \because S(D_j, x_j^{t-1}) = 0$

now $P^{t-1} = [P_j^{t-1}, P_i^{t-1}] = [1, 0]$

$$diag(P^{t-1}) - P^{t-1}(P^{t-1})^T =$$

$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} - \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

$$\frac{\partial x_j^t}{\partial {x^{t-1}}_j} = diag(P^{t-1}) - P^{t-1}(P^{t-1})^T = constant$$

$$\frac{\partial x_j^{t+1}}{\partial {x^t}_j} = 0$$

point converged to unique point, after second iteration

REFERENCES

[1] J. F. Rendon-Sanchez and L. M. de Menezes, "Structural combination of seasonal exponential smoothing forecasts applied to load forecasting", Eur. J. Oper. Res., vol. 275, no. 3, pp. 916-924, Jun. 2019.

[2] F. Bashir and H. Wei, "Neurocomputing Handling missing data in multivariate time series using a vector autoregressive model-imputation (VAR-IM) algorithm", Neurocomputing, vol. 276, pp. 23-30, 2018

[3] I. Markovsky. "Algorithms and literate programs for weighted low-rank approximation with missing data", volume 3 of Springer Proc. Mathematics, pages 255–273. Springer, 2011

[4] S. Ryu, M. Kim and H. Kim, "Denoising autoencoder-based missing value imputation for smart meters", IEEE Access, vol. 8, pp. 40656-40666, 2020.

[5] L. Liu, M. Esmalifalak, Q. Ding, V. A. Emesih, and Z. Han, "Detecting false data injection attacks on power grid by sparse optimization," IEEETransactions on Smart Grid, vol. 5, no. 2, pp. 612–621, 2014.

[6] Y. He, G. J. Mendis, and J. Wei, "Real-time detection of false data injection attacks in smart grid: A deep learning-based intelligent mechanism," IEEE Transactions on Smart Grid, vol. 8, no. 5, 2017.

[7] Jinsung Yoon, William R. Zame, Mihaela van der Schaar, "Estimating Missing Data in Temporal Data Streams Using Multi-Directional Recurrent Neural Networks," IEEE Transactions on Biomedical Engineering, 2019.

[8] K. Park et al "Missing-Insensitive Short-Term Load Forecasting Leveraging Autoencoder and LSTM", IEEE Acess 10.1109/ACCESS.2020.3036885, Nov 2020

[9] S. A. Foroutan and F. R. Salmasi, "Detection of false data injection attacks against state estimation in smart grids based on a mixture gaussian distribution learning method," IET Cyber-Physical Systems: Theory & Applications, vol. 2, no. 4, pp. 161–171, 2017.

[10] N. Laptev, J. Yosinski, E. Li, and S. Smyl, "Time-series extreme event forecasting with neural networks at uber," International Conference on Machine Learning, 2017.

[11] Michael Widrich, "Modern Hopfield Networks and Attention for Immune Repertoire Classification", 34th Conference on Neural Information Processing Systems (NeurIPS 2020)

[12] Hubert Ramsauer, "Hopfield Networks is All You Need," International Conference on Machine Learning, 2021.

[13] Donald J Berndt and James Clifford. Using dynamic time warping to find patterns in time series.In KDD workshop, volume 10, pages 359–370. Seattle, WA, 1994

[14] Xingyu Cai, "DTWNet: a Dynamic Time Warping Network",Advances in Neural Information Processing Systems, 2020

[15] Romain Tavenard ,"Tslearn, A Machine Learning Toolkit for Time Series Data",Journal of Machine Learning Research 2020,

[16] A. H. Yaacob, I. K. Tan, S. F. Chien, and H. K. Tan, "Arima basednetwork anomaly detection," in 2010 Second International Conference on Communication Software and Networks. IEEE, 2010, pp. 205–209,

[17] S.N.Ram, "nference Based Latent Space Method for False data and Topology estimation under suspect EMS data", International Conference on Power Electronics and Energy (ICPEE) 2021.

[18] A. Ashok, M. Govindarasu, and J. Wang, "Cyber-physical attackresilient wide-area monitoring, protection, and control for the power grid," Proceedings of the IEEE, vol. 105, no. 7, pp. 1389–1407, 2017.