DEPARTMENT OF INDUSTRIAL ENGINEERING & INNOVATION SCIENCES
MASTER OPERATIONS MANAGEMENT AND LOGISTICS
Information Systems IE&IS

Interpretable Deep Learning for Time Series Forecasting

A Case Study in the Semiconductor Industry

J.Y. Oostvogel – 0964621

Introduction

Effective demand forecasting is essential to maintain a seamless and efficient supply chain, especially in markets marked by high levels of uncertainty. Deep learning (DL) models could offer numerous benefits for demand forecasting applications; however, DL also brings certain challenges that need to be resolved. Often being perceived as black-box models causes a lack of trust from the user. Furthermore, these models commonly require substantial computational resources and time. This challenge leads to the formulation of the following research question that this thesis aims to answer:

"How can deep learning be used effectively in demand forecasting within the semiconductor industry to improve accuracy while ensuring interpretability and computational efficiency?"

DL Algorithms

Three DL models were selected to test forecast accuracy, computation time, and interpretability. The selected DL algorithms were the following:

- LSTCN: Long Short Term Cognitive Networks
- TFT: Temporal Fusion Transformer
- NHITS: Neural Hierarchical Interpolation for Time Series

LSTCN allows for the integration of prior knowledge through the initialization of the model. Domain information on the inner relations between the time series was gathered by surveying domain experts. After processing, it was analyzed how this prior knowledge influenced results.

Results

Table 1 shows an average of eleven accuracies, calculated from January until November 2023 of the three DL models against NXP benchmarks. Accuracy is shown for each Major Article Group (MAG) and different set of Design Win (DW) data tested. Table 2 shows the performance on a subset of the data from the M5 competition, compared to the submission of the fourth-best model (LightGBM).

MAG	DW DATA	LSTCN	LSTCN (prior)	TFT	NHITS	ADFT BP	NXP
RAN	-	52.34%	49.84%	34.40%	55.87%	13.18%	46.50%
	WON	54.19%	56.07%	32.03%	52.04%		
	+OPEN	52.93%	**58.21%**	34.00%	53.03%		
RGP	-	61.15%	59.74%	57.08%	64.74%	48.86%	60.74%
	WON	62.34%	65.36%	52.85%	62.43%		
	+OPEN	**66.44%**	63.86%	54.34%	64.72%		
RMC	-	53.11%	52.41%	49.78%	56.65%	34.09%	49.37%
	WON	55.45%	57.77%	47.41%	56.11%		
	+OPEN	55.87%	**58.09%**	47.66%	56.54%		

Table 1: Average accuracy for each model per Major Article Group (MAG) and dataset against NXP benchmarks.

Metric	LSTCN	TFT	NHITS	LightGBM
WRMSSE	1.862	4.849	8.703	**0.539**
MAE	1.231	2.683	3.555	**0.802**

Table 2: WRMSSE and MAE of DL algorithms and LightGBM benchmark.

Reviewing the accuracy over the months of January to November 2023 can reveal the consistency of the performance of the models. Figure XXXX shows these accuracies for to of the three MAGs using the dataset with OPEN opportunities.

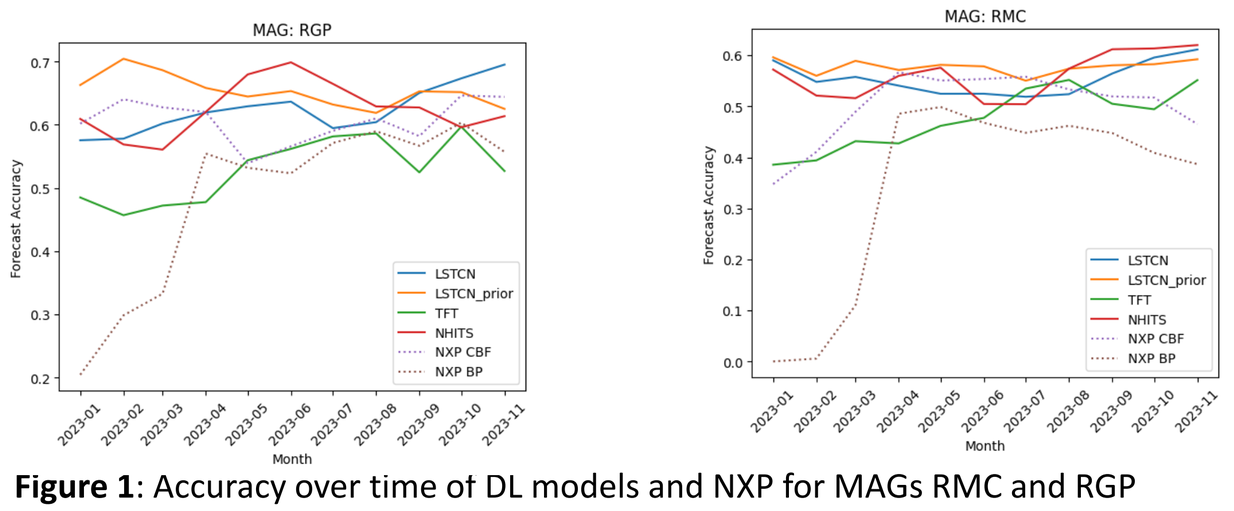

Figure 1: Accuracy over time of DL models and NXP for MAGs RMC and RGP

Concerning the computation time, shown in Table 3, every model was run locally on a PC and the Azure Kubernetes Service (AKS) with a Graphical Processing Unit for the NXP dataset. The magnitude of the M5 dataset required an impractical duration for a local run; therefore, only outcomes that use the AKS are showcased. Table 3 shows the computation times of these tests.

Data	Resource	LSTCN	TFT	NHITS
NXP	Local	00:02:16	03:13:15	00:26:07
NXP	AKS	00:01:47	01:01:33	00:03:20
M5	AKS	07:41:32	06:30:22	00:35:13

Table 3: Computation times of DL algorithms for NXP and M5 dataset.

Regarding the evaluation of interpretability, a survey was conducted by presenting the explanations of the DL algorithms from use cases to domain experts within NXP. On a Likert scale from 1 to 5, the participants had to rank their perception of the explanations on the 5 criteria shown in Table 4.

DL model	Understandability	Usefulness	Trust	Informativeness	Satisfaction
LSTCN	3.53 ± 1.11	3.67 ± 1.21	3.50 ± 1.25	3.50 ± 1.22	3.50 ± 1.20
TFT	3.53 ± 0.90	3.47 ± 1.17	3.63 ± 1.07	3.50 ± 1.07	3.37 ± 1.00
NHITS	3.33 ± 0.99	3.10 ± 1.37	3.20 ± 1.16	3.20 ± 1.19	3.13 ± 1.11

Table 4: Mean score and standard deviation on a 5-point Likert scale per evaluation criterion.

Conclusion & Recommendations

When three DL methods were tested on semiconductor industry data, it was found that deep learning could exceed company benchmarks for forecasting, enhancing the accuracy of the forecasts. In particular, the LSTCN and NHITS algorithms demonstrated increased forecasting accuracy based on historical data predictions compared to the company's forecasts. The computation time appeared to be managed at an acceptable level, even for the model with the highest computation times. Although it remains unclear which method provides the best interpretability for demand managers in the semiconductor industry, the overall interpretability was sufficient to instill trust among its users. Ultimately, interpretable deep learning models proved their ability to compete with and surpass the current standard in NXP. They maintained a low and acceptable computation time and earned user trust through their interpretability.

This research explores interpretable deep learning algorithms for time series forecasting, focusing on methods that incorporate interpretability directly into the algorithms rather than relying on post hoc agnostic techniques. Future opportunities would be to expand the evaluation to include non-interpretable DL algorithms in combination with agnostic techniques for a comprehensive understanding of interpretability metrics. Additionally, the integration of prior knowledge into LSTCN necessitates future research to grasp its impact.