Attention for Arbitrage: Transformer-Based Forecasting of ERCOT Day-Ahead vs. Real-Time Spreads under Rising Renewables

Tunahan Gumuskaya
M.S. in Operations Research - ML & AI
Columbia University
NY, USA
tg2885@columbia.edu

Abstract

ERCOT's rapid shift toward variable renewables - with wind and solar together rising from 19% to 37% of the supply mix between 2018 and 2025 - has made a small set of operating hours disproportionately important for virtual-bidding, a mechanism that allows financial participants to take offsetting positions in the day-ahead (DAM) and real-time (RT) markets, profiting from RT-DAM mispricing while helping the system converge to efficient prices. Even though headline RT-DAM convergence improved relative to 2023, the data show that spreads still "blow out" in specific system states: very high load and hours when solar or wind underperform their forecasts. ERCOT likewise reports that real-time congestion costs exceeded day-ahead congestion by about 3%, largely because renewable forecast error materializes in real-time, which strengthens the case for better day-ahead spread signals.[1] As these intermittent resources set or influence prices more often, such state-dependent deviations become harder for traditional forecasting approaches to capture.

Building on the framework introduced in "Deep Learning-Based Electricity Price Forecast for Virtual Bidding in Wholesale Electricity Market" [2], this paper examines whether modern sequence models - in particular Transformer architectures - can learn these conditional patterns more effectively than LSTMs and CNNs for ERCOT day-ahead versus real-time price-spread forecasting. The central question is whether the attention mechanism that enabled large language models such as ChatGPT and Gemini can also improve trading signals in wholesale electricity markets. To investigate this, a daily 656-feature pipeline is constructed that ingests core fundamentals (zonal load forecasts for 9 LMPs, solar and wind forecasts, fuel-price indicators, and calendar/holiday features) and classifies 24 hourly spreads into five payoff-oriented buckets aligned with economics of virtual-bidding. Models are trained in a weekly walk-forward scheme using only information available before bidding, and both all-hours and confidence-filtered bidding policies are backtested.

Empirically, the Transformer model delivers more stable week-to-week performance across different lag parameters and, when paired with selective trading, produces more consistent virtual-bidding signals than LSTMs and CNNs. In the backtest, the best-performing Transformer model captures a disproportionate share of large spreads in hour ending 6, 7 and 8 a.m., indicating that early-morning transition hours can be especially informative in the current ERCOT resource mix. When the training rolling time window is one year and trades executed either for extreme negative and positive forecast, the Transformer architecture ended up having positive PnL for 100% of all configurations in hour ending 6, 7 and 8 a.m., 92% for LSTM and 75% for CNN. This study is intended solely as methodological and market-structure research and should not be interpreted as investment advice.

I. Exploratory Data Analysis

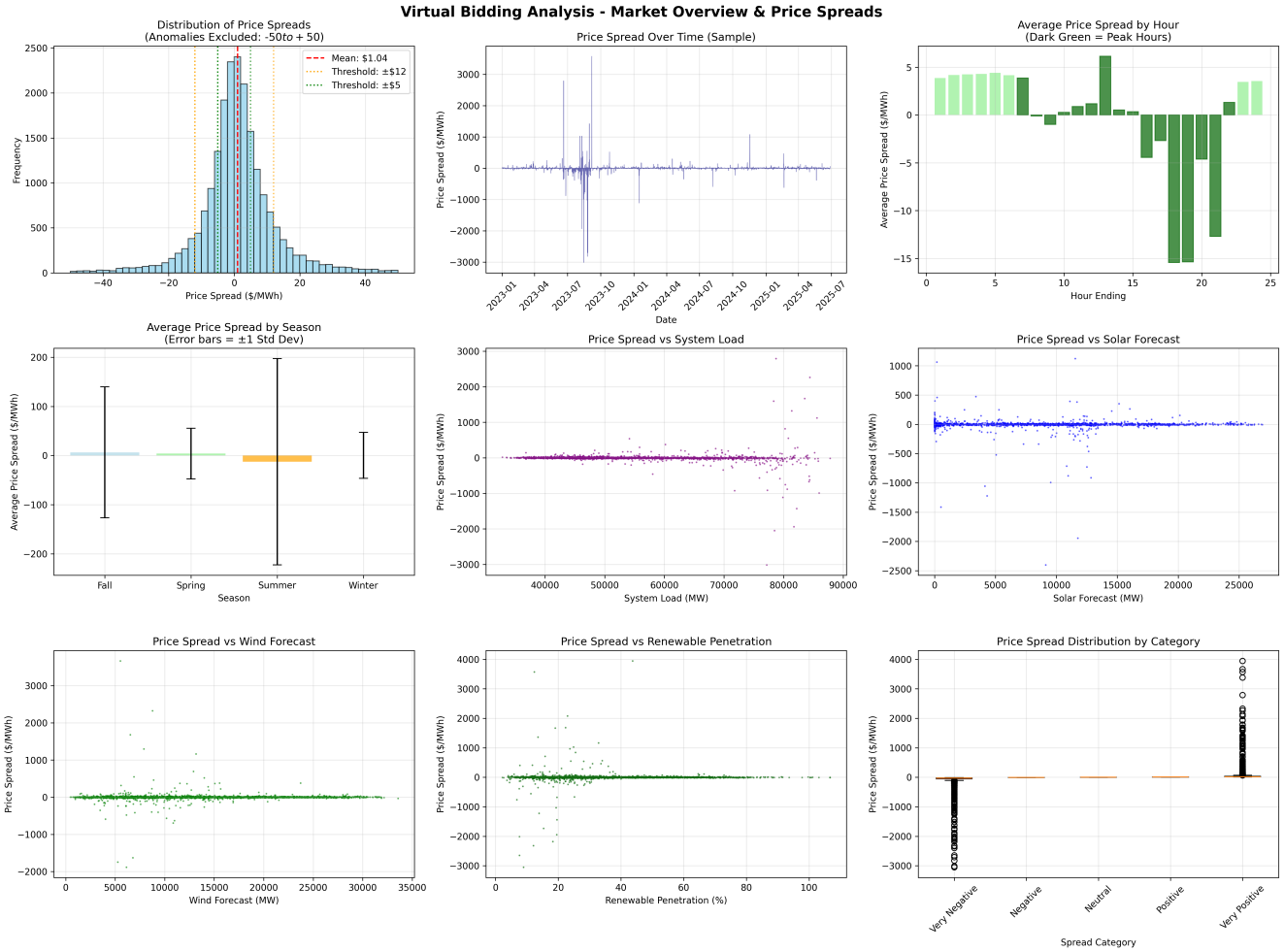


Figure1. ERCOT virtual bidding market overview: RT – DAM price spreads and fundamentals.
Panels (1)-(9) are ordered row-wise from top left to bottom right.

Note: While virtual-bid payoffs in ERCOT are typically expressed using the spread DAM − RT, this paper defines the spread as RT − DAM for modeling convenience. Under this convention, a negative RT − DAM spread simply indicates that real-time prices are lower than day-ahead prices.

Panel (1) shows the empirical distribution of system hourly price spreads (RT-DAM System Lambda), excluding extreme anomalies for visualization (-$50 and +$50/MWh); the distribution is tightly centered around about $1/MWh, with thresholds at ±$5/MWh (approximate uplift cost) and ±$12/MWh highlighting the region of interest for virtual-bidding. **Panel (2)** plots the time series of spreads, illustrating that large blow-ups are concentrated in mid-late 2023, followed by a calmer regime with episodic spikes rather than a simple trend. **Panel (3)** reports the average spread by hour ending (HE), revealing strong intraday structure: some peak hours systematically exhibit more negative or positive spreads than others. **Panel (4)** shows average spreads by season (with ±1 standard deviation error bars); summer, in particular, displays a slightly negative mean spread with large dispersion, implying frequent RT < DAM outcomes and potential value for selling DAM and buying RT. **Panel (5)** plots spreads against system load, where spreads remain close to zero at low and moderate load but fan out sharply in the 70-85 GW range, consistent with stressed system conditions. **Panel (6)** relates spreads to

the STPPF_SYSTEM_WIDE solar forecast, indicating that large spreads coincide with solar forecast errors, in line with evidence that over-forecasting solar output has required additional real-time unit commitments.

Panel (7) shows a similar but more dispersed relationship for the STWPF_SYSTEM_WIDE wind forecast; spreads widen notably when wind availability is low. **Panel (8)** plots spreads against renewable penetration, suggesting that higher realized renewables generally support RT-DAM convergence, whereas hours with low penetration exhibit much wider spreads and more extreme RT-DAM deviations. Finally, **Panel (9)** summarizes the distribution of spreads by bucket (very negative, negative, neutral, positive, very positive), illustrating that once spreads leave the normal ±$12/MWh band they can take extremely large values (down to around -$3,000/MWh and up to roughly +$4,000/MWh), underscoring the importance of focusing the modeling effort on these tail events.

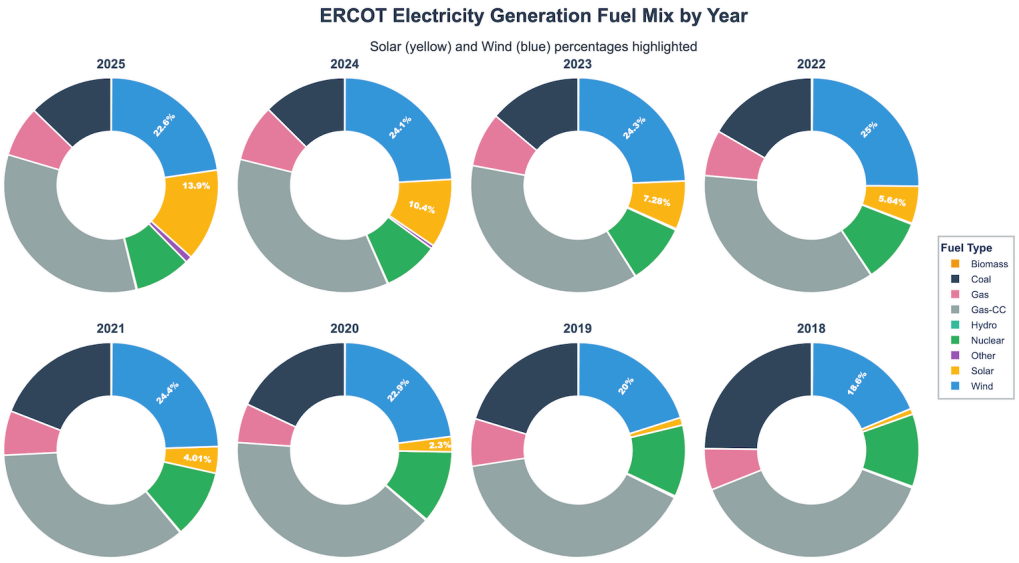


Figure 2. ERCOT electricity generation fuel mix by year.
Donut charts show the annual share of total ERCOT generation by fuel type from 2018–2025.

ERCOT's rapid shift toward variable renewables highlights how quickly the resource mix is changing. This trend is consistent with claims that solar and wind will account for most incremental electricity generation, while coal, hydro, and nuclear remain roughly flat. In such an environment, hourly normalized price "scalars" and trading rules can no longer be calibrated solely from long historical averages. They must explicitly account for the evolving contribution of solar, wind, and emerging storage resources.

II. Introduction

Virtual-bidding is a financial trading mechanism that allows market participants to profit from price differences between the day-ahead (DAM) and real-time (RT) electricity markets without physically delivering or consuming electricity. Traders submit bids that implicitly forecast whether prices will rise or fall and seek to arbitrage the spread between the two markets. This activity can improve market efficiency by enhancing price convergence and liquidity, but it also requires accurate forecasting of price movements to be profitable. In ERCOT, rising renewable penetration and emerging battery storage resources mean that historical scalars calibrated in a more thermal-dominated system no longer reliably represent cleared DAM prices.

The motivation in this study is to identify predictive signals for virtual-bidding from high-dimensional market and forecast data-specifically, a 656-dimensional feature set that includes granular solar, wind, and load day-ahead forecasts, fuel indicators, and calendar effects. For instance, a sharp decline in wind generation forecasts in West Texas during a tight summer afternoon may be associated with higher RT prices and a positive spread signal for HE 19. By capturing such relationships across space, time, and market variables, the goal is to forecast spread movements more accurately and translate them into virtual-bidding strategies. The analysis compares the performance of Convolutional Neural Networks (CNNs), Long Short-Term Memory networks (LSTMs), and Transformer architectures. An explicit question is whether Transformer, which enabled large language models such as ChatGPT and Gemini, can also provide an advantage in electricity trading.

Previous electricity price-forecasting research has established the usefulness of machine-learning models such as LSTMs. LSTM-based models have outperformed classical time-series methods like ARIMA in ERCOT day-ahead price forecasting [4] and have produced accurate forecasts in other markets (e.g., the German-Luxembourg zone) when incorporating load and renewable features [5]. Nevertheless, LSTMs remain sequential models: training is slower and less parallelizable than for CNNs or Transformers, and very long sequences can still be challenging due to vanishing or exploding gradients, even though these effects are mitigated relative to simple RNNs.

CNNs, by contrast, capture local patterns in time series via convolutional filters, effectively learning short-term dynamics and seasonal structure. They are less susceptible to vanishing gradients and can be stacked deeply to enlarge the receptive field. In practice, CNN-based architectures have also been successfully applied to price forecasting; for example, a neural network model incorporating CNN components has beaten ARIMA benchmarks for ERCOT price prediction [4]. However, pure CNN architectures have limited ability to capture long-range temporal dependencies unless the network is made very deep or uses large kernels. As a result, CNNs are often used as feature extractors and combined with LSTMs, yielding improved accuracy by pairing CNN-based local feature learning with LSTM-style sequential memory [6].

Transformer-based models use self-attention mechanisms to model relationships between all time steps in a sequence simultaneously, enabling them to capture long-range and complex temporal dependencies very effectively. Because Transformers operate without recurrent connections, they process sequences in parallel, which improves training speed and scalability. For time-series forecasting, Transformer architectures can attend to important days or hours in the past regardless of their distance in time, a potentially valuable property for volatile markets such as ERCOT. They also handle high-dimensional, multivariate inputs well by learning inter-feature interactions through attention. However, empirical evidence in financial time-series applications is mixed: in some studies, Transformer models have not consistently outperformed LSTMs, and in at least one case they provided only a limited advantage in predicting absolute prices while underperforming LSTMs in forecasting price differences and directional movements [7].

In this context, the present study constructs multiple virtual-bidding trading strategies and compares the performance of LSTM, CNN, and Transformer models under realistic ERCOT market conditions, focusing especially on their ability to anticipate large RT-DAM spreads that matter most for virtual-bidding economics.

III. Methodology

A) Assumptions

This study adopts the same trading assumptions as Gong, Zhao, and Liao in “Deep Learning-Based Electricity Price Forecast for Virtual Bidding in Wholesale Electricity Market” [2]:

- The uplift cost of trading 1 MWh of energy is $5.00
- Bids and offers are assumed to clear by placing orders at the price cap or price floor [8]
- There is a daily budget constraint: at most 1 MWh is traded per day. If multiple hours are selected on a given day, the 1 MWh position is split evenly across those hours

- Submitted virtual-bids and offers are assumed not to affect the system lambda (i.e., the trader is price-taking)

B) Model Design

The forecasting task is approached using an encoder-only Transformer architecture designed to predict the hourly price spread between the real-time (SCED) and day-ahead (DAM) system prices. The model is constructed to capture complex temporal dependencies among multivariate electricity fundamentals, including load, wind, solar, fuel prices, and time-of-year effects.

Each calendar day is encoded as a 656-dimensional feature vector aggregating 24-hour profiles of these variables. This daily vector is first projected through a linear layer into a $d_{\text{model}} = 128$-dimensional embedding space. A learned positional encoding is then added to each day token, followed by dropout, so that the network retains information about sequence order without relying on fixed sinusoidal patterns. The core of the model consists of multiple stacked Transformer encoder layers, each comprising multi-head self-attention and position-wise feed-forward sublayers with residual connections and layer normalization. The final hidden representation corresponding to the last token in the sequence (the $T+1$ context token) is passed through a linear decoder head that outputs 24 x 5 logits, representing the hourly price-spread classes. The threshold of $12.00 is about the median of the price spreads that is outside $5.00.

Next-day ($T+1$) ERCOT system-lambda spreads, defined as RT-DAM, are formulated as a 24-hour, 5-class classification problem. For each hour, the model predicts the probability that the spread falls within one of five bins: bin 0: (-∞, -12), bin 1: [-12,-5), bin 2: [-5,5), bin 3: [5,12), bin 4: [12, ∞).

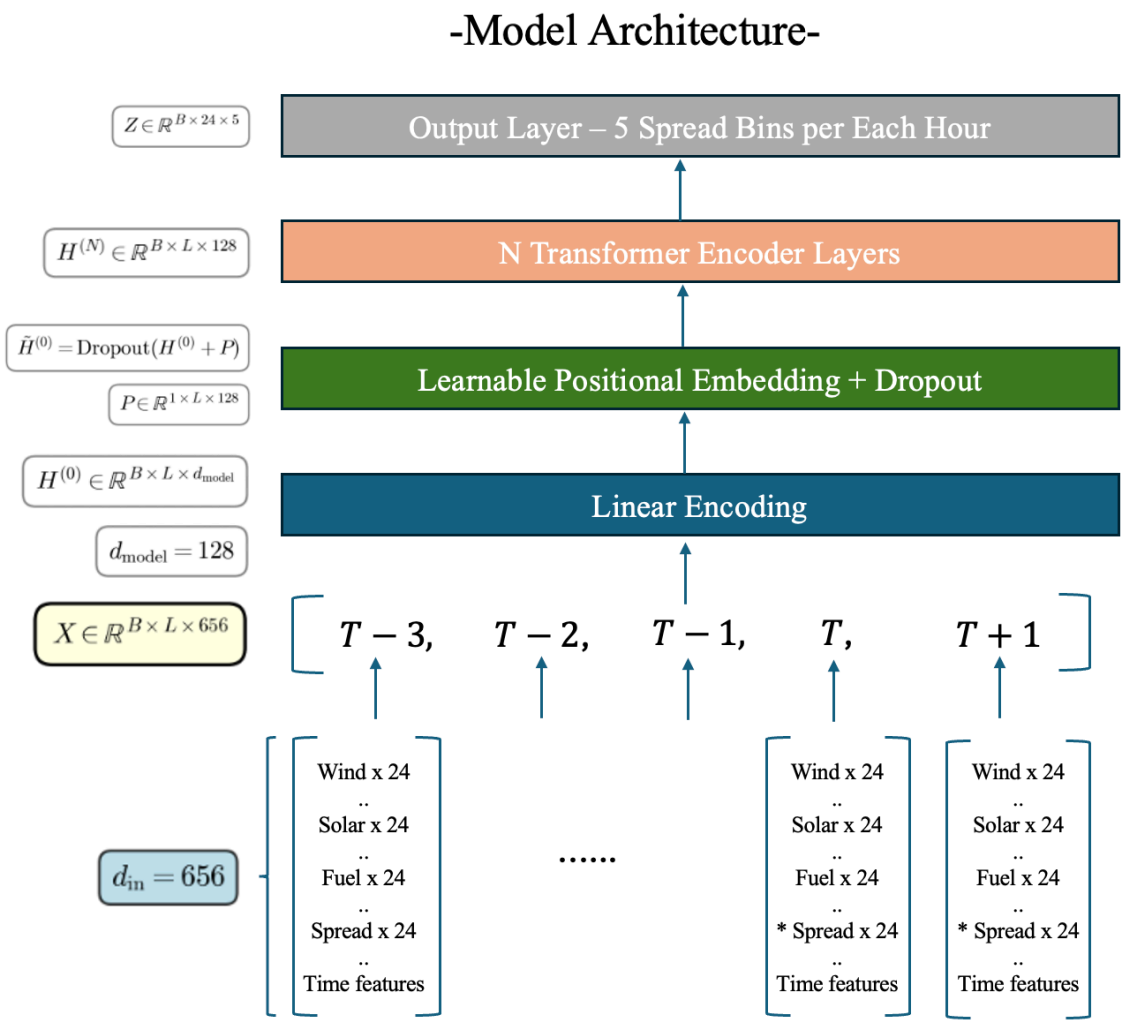


Figure 3. For each forecast $T+1$, a sequence of L = lag_days + 2 daily tokens are constructed. *The example shown corresponds to lag_days = 3, so L = 5 tokens:* $T-3, T-2, T-1, T, T+1$. *Spread = Masked as 0

The 656-dimensional daily input vector includes the following components:

- **System and zonal load forecasts:** 9 zonal load forecasts, each with 24 hourly values
- **Solar features:** 3 ERCOT day-ahead solar forecast series, each with 24 hourly values
- **Wind features:** 12 ERCOT day-ahead wind forecast series, each with 24 hourly values
- **Fuel prices:** 24-hour NYMEX natural gas price indicators
- **Time features:** sine/cosine encodings for month, day of month, and day of week; a year offset; and a holiday flag
- **Spread channel and mask:** historical system spreads are robust scaled (median/IQR) and included only for fully observed past days. For the current day T and the $T + 1$ context token, the spread channel is set to zero to prevent information leakage from future prices

With a lag parameter lag_days = L, the network consumes a sequence of $L + 2$ daily tokens: the last L fully observed days (containing spread information), plus the current day T (spread masked) and a final $T + 1$ context token (also spread masked). The Transformer encoder processes this length $L + 2$ sequence, and the model emits a single-day, 24-hour forecast for $T + 1$ from the last token's hidden state. This design preserves the benefits of self-attention over multiple past days while matching the two-settlement structure of virtual-bidding, which requires all $T + 1$ decisions to be made using only information available up to the DAM run.

C. Data Preprocessing and Dataset Split

A robust median-IQR scaling is applied to continuous features so that typical values are centered around zero and rescaled to a moderate range, while extreme outliers are dampened rather than dominating the scale. This is particularly useful in ERCOT, where fundamentals and spreads exhibit heavy tails and scarcity spikes: the transform keeps normal days numerically well-behaved for the neural network but still preserves the signal from large events without letting them destabilize training.

A walk-forward evaluation scheme is adopted to align with market reality. The global chronological splits are:

- **Training:** 2023-01-01 to 2023-12-23
- **Initial validation:** 2023-12-24 to 2023-12-30
- **Test period beginning:** first test week: 2024-01-01 to 2024-01-07

During rolling evaluation there is always a one-day operational gap between the most recent data used for training and the first test day that can be forecast, reflecting the requirement that models must be trained before bids are submitted.

D. Rolling Window and Sequence Construction

At the start of each test week, an "as-of date" is defined as Sunday. The known dataset is truncated at this as-of date, and a rolling training window of length $W \in \{30,90,180,360\}$ days is applied (for an ablation over window length) to construct the training corpus for that week. Within this window, the last seven known days are reserved as the validation set for that week; the remaining days form the weekly training set.

Sequences are then constructed by sliding a window of $L + 2$ days over each of these sets. For each slice of $L + 2$ consecutive days, the first $L + 1$ tokens serve as context and the final day provides the 24 hourly target classes, yielding one training example per day in the window once sufficient history is available.

E. Training and Hyper-Parameter Search

Training uses a weighted cross-entropy loss, with inverse-frequency class weights computed from that week's training labels to address class imbalance across the five spread bins. Optimization is performed with Adam, with learning rate and weight decay drawn from a hyper-parameter search.

For each test week, a compact Ray Tune search is conducted over the following space:

- number of encoder layers $\in \{1,2,3,4\}$
- positional-dropout rate $\in [0,0.3]$
- learning rate $\in [10^{-4}, 3 \times 10^{-3}]$ (log-uniform)
- weight decay $\in [10^{-6}, 10^{-3}]$ (log-uniform)

For every sampled configuration, the model is trained for 50 short epochs on that week's train/validation split, and the checkpoint with the lowest validation loss is selected. Using this best configuration, a final model for the week is then re-trained on the full weekly training portion for a larger number of epochs before generating forecasts.

This procedure is repeated week-by-week, allowing both model weights and hyper-parameters to adapt as new data becomes available. Within each week, the trained model is kept fixed, but forecasts are generated day by day: for each forecast day, the most recent realized information is incorporated into the input sequences, and the model produces a 24 x 5 tensor of class probabilities. For every hour, the predicted class and its associated confidence (maximum class probability) are stored for subsequent backtesting and evaluation.

For the **LSTM model**, the only change relative to the Transformer is the sequence model itself. After a linear projection into a 128-dimensional space, daily tokens are passed through a stacked LSTM network with hidden size 128 and 1-3 layers (selected via hyper-parameter search). For the **CNN model**, the difference lies in how temporal structure is modeled. After a linear projection from 656 to a lower-dimensional space (e.g., 128), the sequence of daily tokens is rearranged into a channel x time tensor and passed through a stack of 1-D convolutional layers with ReLU activations and dropout, which learn local temporal filters over the lag window. The resulting feature maps are averaged over time to produce a single sequence-level representation, which is then fed to a linear decoder that outputs the 24 x 5 logits.

IV. Results for bin 0 & configuration of time window 360 & lag 4 & confidence ≥ 95%

A) Transformer

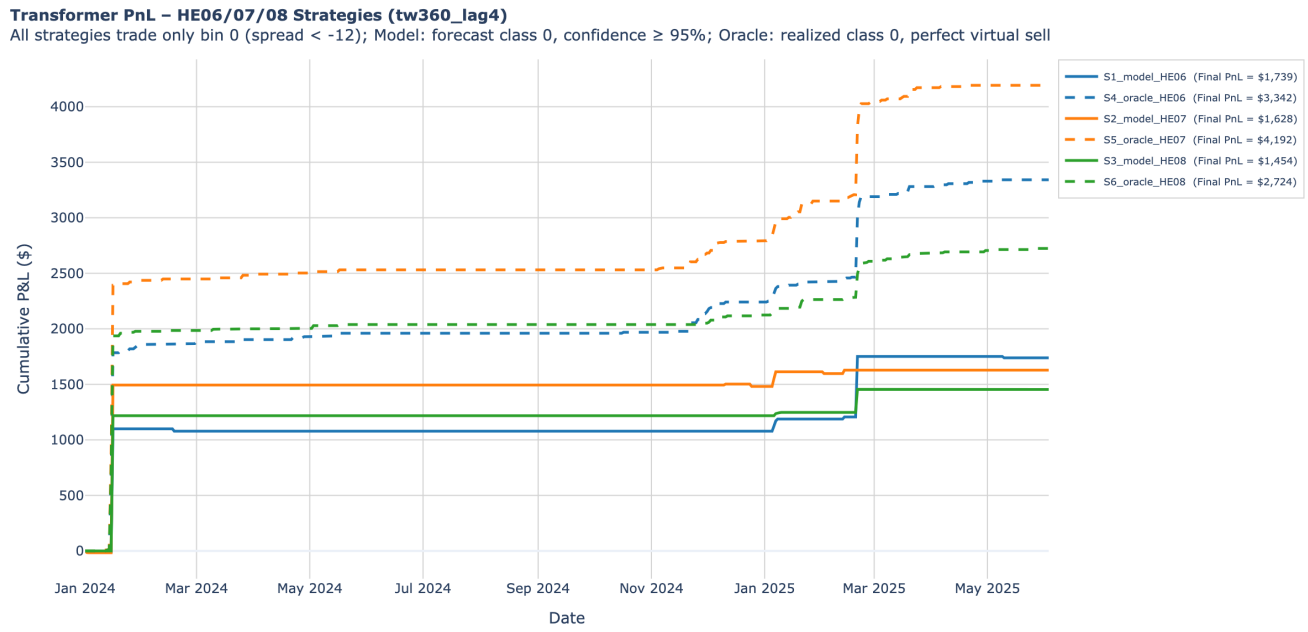


Figure 4. Transformer cumulative PnL for selected hours (HE 6, 7, and 8) under bin-0 trading strategies, for hours chosen based on economic performance and confusion-matrix diagnostics.

Bin 0 & time window 360 & lag 4 & confidence ≥ 95% have been shown here given the great Transformer model performance for both economical and statistical precision results. In addition, both extreme tails results with all other configurations will be shown in the next section.

- **S1, S2, S3**: trades are executed for HE 6, HE 7, and HE 8, respectively, based solely on the model's prediction when the predicted class is $(-\infty, -12)$ and the predicted probability is greater than 95%,
- **S4, S5, S6**: trades are executed for HE 6, 7 and 8, respectively, based on the ground truth when the realized class is $(-\infty, -12)$. These represent oracle benchmarks using perfect foresight; they are not implementable in practice but provide an upper bound on achievable performance.

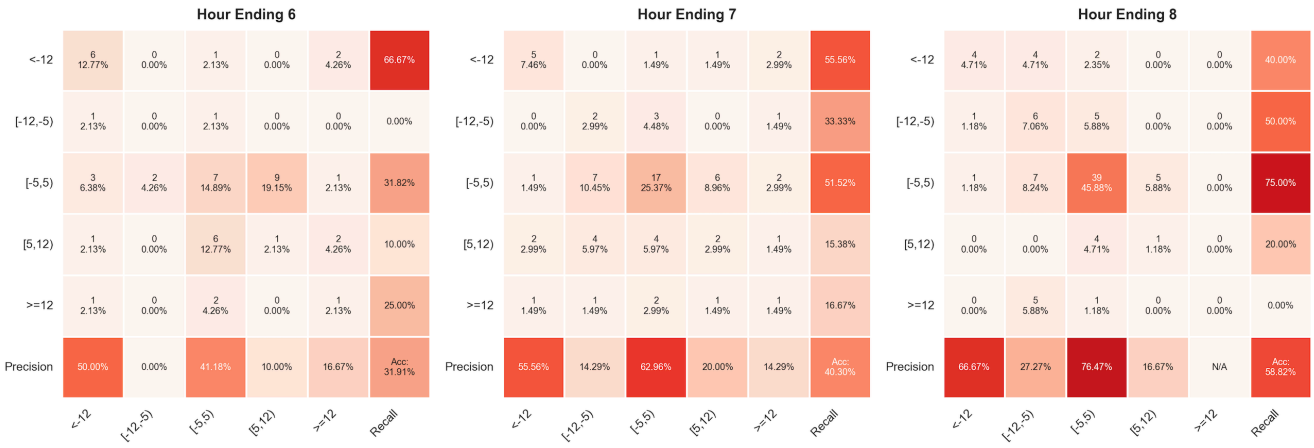


Figure 5. Transformer confusion matrices for selected hours (HE 6, 7, and 8) at 95% confidence, illustrating the model's ability to capture extreme negative spreads.

In Figure 4, the high-confidence Transformer strategies for HE 6, 7, and 8 all finish the backtest with strongly positive cumulative PnL, showing that the model repeatedly identifies the most attractive negative spread opportunities for selling DAM, buying RT and avoids extended drawdowns. In Figure 5, 95% confidence confusion matrices reinforce this picture: for the extreme negative spread bin 0, the Transformer attains precision of about 50% for HE 6, 56% for HE 7, and 67% for HE 8**,** meaning that a large fraction of its extreme-sell signals is correct. Combining the economic and classification evidence, the results indicate that the model successfully captures most of the valuable large negative spreads. When it does misclassify, the realized spreads are often still moderately negative, large enough in magnitude to cover the $5/MWh uplift but not quite beyond the -$12/MWh floor. So many of these errors remain profitable rather than severely loss-making. In the context of virtual-bidding, precision matters more than recall. A high-precision model like this one provides signals that traders can trust, whereas a low-precision model, even with higher recall, would generate frequent false positives and substantial losses.

B) LSTM

In contrast to the Transformer, the LSTM architecture does not generate strong or reliable trading signals for the selected hours. Figure 6 shows that the high-confidence LSTM strategies for HE 6 finish the backtest with only marginal positive PnL, while HE 8 ends slightly negative. In all three cases the model captures only a small fraction of the oracle benchmark PnL for bin 0 events. The confusion matrices in Figure 7 confirm this weakness: at the 95% confidence threshold the LSTM almost never issues bin 0 predictions, concentrating instead in the neutral [-5,5) bin and leaving the precision and recall for the extreme negative spread class effectively undefined because the model produced zero predictions for this class.

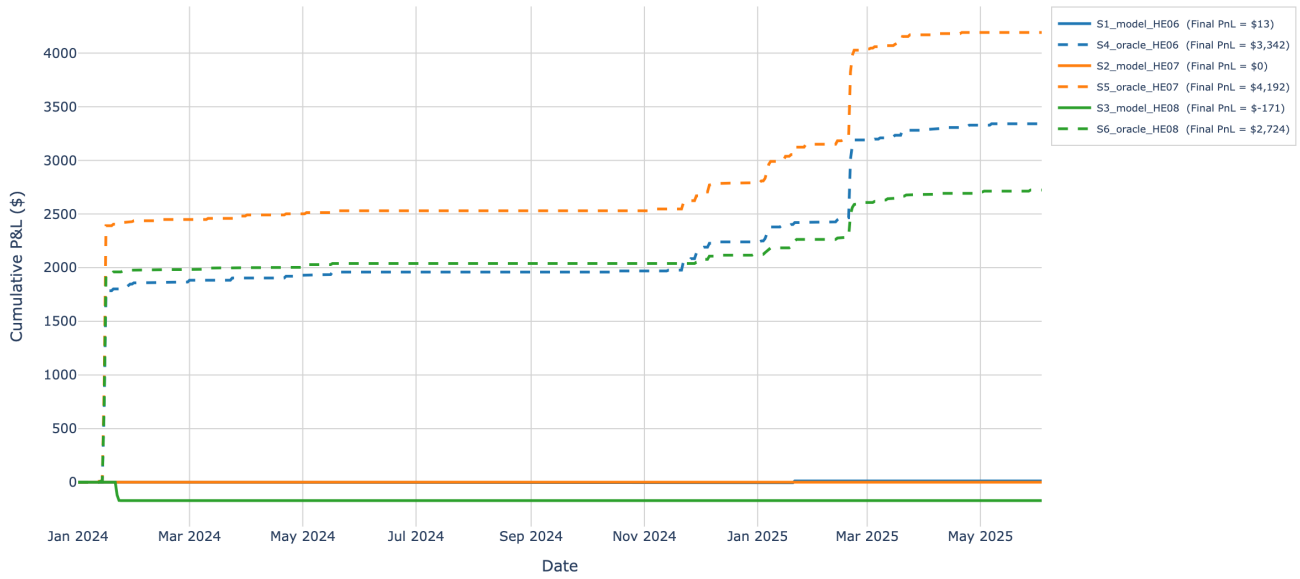


Figure 6. LSTM cumulative PnL for selected hours (HE 6, 7, and 8) under bin-0 trading strategies, showing weaker tail performance than the Transformer.

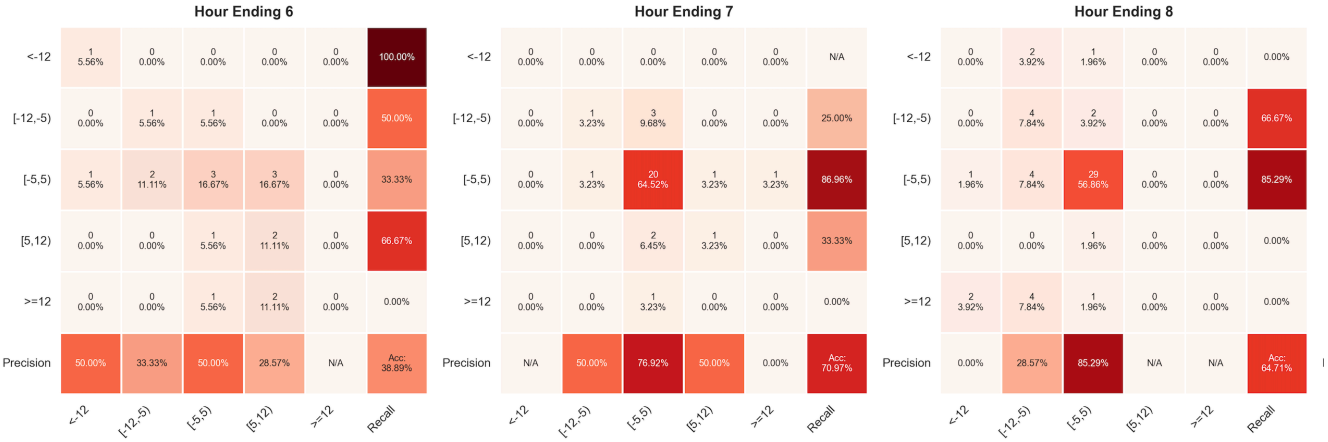


Figure 7. LSTM confusion matrices for selected hours (HE 6, 7, and 8) at 95% confidence, illustrating sparse bin-0 predictions and limited capture of extreme negative spreads.

C) CNN

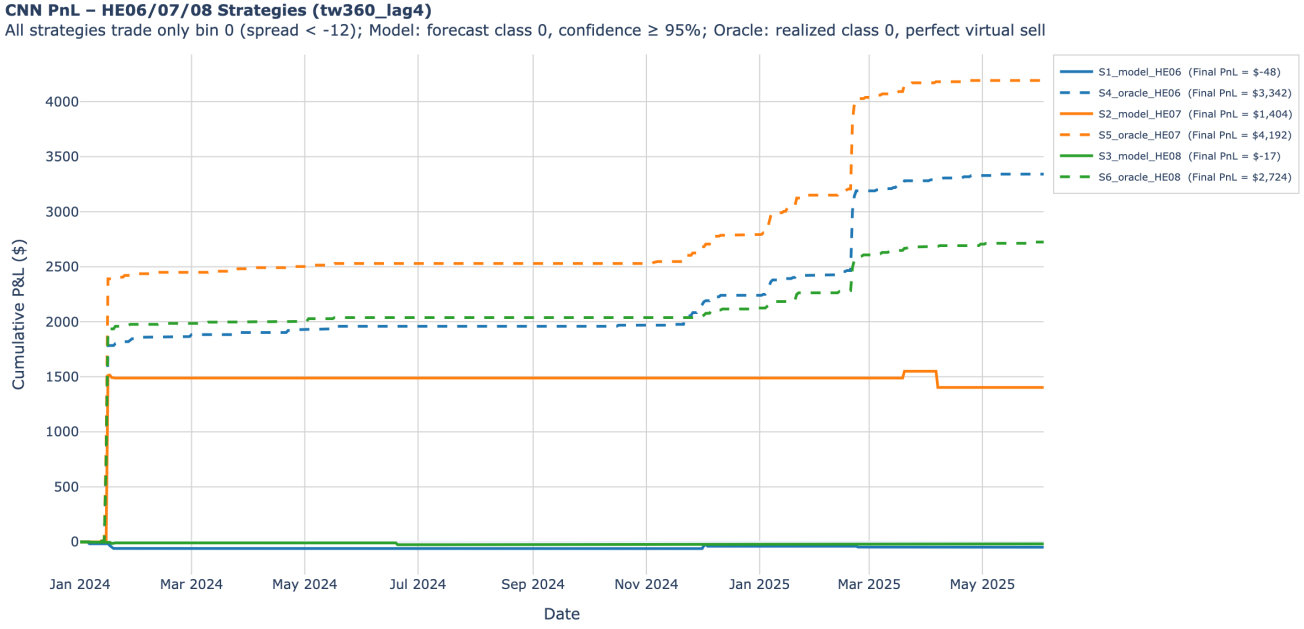


Figure 8. CNN cumulative PnL for selected hours (HE 6, 7, and 8) under bin-0 trading strategies, showing weaker tail performance than the Transformer.

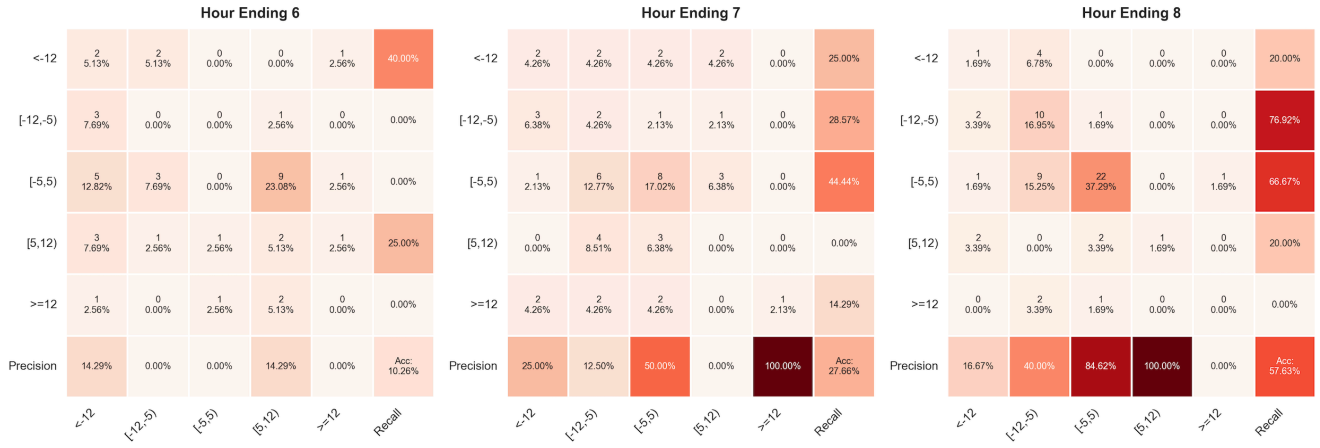


Figure 9. CNN confusion matrices for selected hours (HE 6, 7, and 8) at 95% confidence, illustrating limited precision and recall for extreme negative spreads relative to the neutral and mildly negative bins.

In contrast to the Transformer, the CNN delivers only limited economic value in the tail-focused bin 0 strategy. Figure 8 shows that, at the 95% confidence threshold, the CNN strategies for HE 6 and HE 8 finish the backtest with small losses, while HE 7 generates a modest positive PnL but still captures only a fraction of the oracle benchmark. The PnL curves are largely flat, indicating that the CNN issues very few high-confidence extreme negative (bin 0) signals and often sits out many of the most profitable trades.

The confusion matrices in Figure 9 help explain this behavior. Precision and recall for the extreme negative bin are relatively low, while the model concentrates most of its high-confidence predictions in the neutral and mildly negative bins. In other words, the CNN is more comfortable classifying normal conditions than committing to very negative spreads, which limits its ability to monetize rare but lucrative tail events.

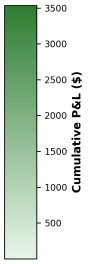
V. Results for bin 0 and 4 & all configurations

While the discussion above focuses on bin 0, the original objective of this study is to evaluate the models on both extreme spread regimes, bins 0 and 4. Accordingly, this section presents the complete set of outcomes for all 16 configurations and all three models when trading is restricted to these two extreme classes, bin 0 and bin 4, so that performance on both large negative and large positive spreads can be assessed.

A) Final Cumulative PnL When Confidence ≥ 50%

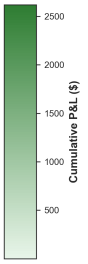
Transformer - Final Cumulative P&L by Configuration and Hour Ending (Confidence >= 50%)

	HE00	HE01	HE02	HE03	HE04	HE05	HE06	HE07	HE08	HE09	HE10	HE11	HE12	HE13	HE14	HE15	HE16	HE17	HE18	HE19	HE20	HE21	HE22	HE23
tw30_lag1	-122	-36	-184	-123	48	28	649	-628	-578	-126	-232	-13	42	144	252	69	59	-194	254	-562	1301	986	-123	-168
tw30_lag2	97	135	-2	29	97	-240	-1510	-402	-108	75	126	-91	-159	-41	0	-1	-422	-879	-497	-3348	184	863	-421	-5
tw30_lag3	-148	5	82	-98	18	-287	-74	-197	-474	-499	-338	2	-378	-157	-295	103	-247	-1541	-342	3538	930	798	-347	-115
tw30_lag4	11	17	62	-82	-89	-386	-2264	-1665	-20	-164	-119	-46	-175	-233	-417	-267	-937	-1276	-499	1432	43	1179	-391	-194
tw90_lag1	10	50	107	13	42	-277	243	-1036	-551	-373	-348	-112	-51	-353	26	-210	-755	-2033	105	-2913	-758	194	113	-28
tw90_lag2	34	49	4	81	87	-47	-1720	485	-797	-591	-447	-183	-85	-495	-26	-283	-490	-1492	-306	-5222	-3131	201	-304	14
tw90_lag3	26	34	-23	43	-106	-291	-351	-200	-532	-454	-440	-59	-255	-447	-150	-474	-752	-1922	358	-1098	-348	-304	-170	-20
tw90_lag4	65	91	-98	104	-8	-278	-1098	-1189	421	-493	-445	-181	-88	-500	-472	-582	-1054	-2183	-938	-78	872	224	-248	1
tw180_lag1	-17	-26	-113	-94	97	-16	888	-995	-304	-741	-366	-71	-472	-363	-440	640	-699	-1470	469	-4036	-773	459	-22	61
tw180_lag2	65	86	71	28	-28	-95	1895	-417	1361	-449	-120	-150	-153	-32	-3	-136	-1923	-778	92	-2455	267	-126	-83	10
tw180_lag3	25	-19	148	60	170	1	109	-42	-473	-510	-421	-139	-483	-439	-201	-360	-831	-1495	-341	1847	-1575	-363	-300	-11
tw180_lag4	77	101	91	137	-83	-52	-538	-248	-574	-523	-131	-147	-320	-12	-324	-258	-777	-812	-980	-2790	-1092	-910	-418	21
tw360_lag1	152	60	221	256	249	82	535	2952	1969	-8	-107	-106	-84	130	12	1278	1065	-1822	-168	-3927	-3280	49	-8	60
tw360_lag2	223	79	187	267	12	-205	469	1719	1201	-331	-79	-13	-173	8	72	67	-166	-1480	-77	2094	-1743	70	-62	-17
tw360_lag3	216	270	-79	-19	-105	-126	726	2678	1998	3	-69	-21	-218	-127	-38	-470	-361	-1878	242	-2177	839	431	119	75
tw360_lag4	83	96	-4	-101	109	147	1685	3005	1366	215	-156	-20	-156	254	23	-319	-282	-1335	-1723	-2639	-2883	-758	-17	-1



LSTM - Final Cumulative P&L by Configuration and Hour Ending (Confidence >= 50%)

	HE00	HE01	HE02	HE03	HE04	HE05	HE06	HE07	HE08	HE09	HE10	HE11	HE12	HE13	HE14	HE15	HE16	HE17	HE18	HE19	HE20	HE21	HE22	HE23
tw30_lag1	-173	-100	-253	-437	-368	-328	0	-1383	149	-565	-405	-53	-207	-524	-157	16	-181	-1092	-596	2620	223	17	-260	-209
tw30_lag2	28	86	-105	-64	32	-69	-451	-1320	-182	-40	-80	-99	-326	272	96	-135	-1977	-2446	-1580	-6045	-1064	779	-155	-39
tw30_lag3	-130	60	70	-34	-79	-462	-1311	55	-816	-160	-321	-89	-206	-86	-30	-160	-655	-1804	-517	-2357	1177	161	-58	-197
tw30_lag4	-255	-131	-302	-288	-262	39	-836	548	-51	-149	-191	-157	-329	-107	-419	-458	-991	-925	-595	-1655	-816	-980	-440	-116
tw90_lag1	-68	-77	-55	-119	-116	-388	-791	1029	141	-85	-395	-133	-234	74	71	-82	-617	-1410	-1599	-2665	-1668	219	-123	-150
tw90_lag2	-113	-133	-151	-137	-67	-184	1737	-1031	-260	-134	-210	26	-150	-232	-162	-509	-1055	-2146	-732	-3248	-1880	-271	-141	122
tw90_lag3	-30	34	28	26	-111	-154	-2098	-814	-997	-581	-400	-160	-224	70	-52	-253	-553	-1482	-1986	-4496	-3308	-445	-72	-124
tw90_lag4	-25	-11	-173	-78	-54	-318	-1402	-21	-82	-498	-180	-93	30	225	-75	-215	-529	-1293	-988	-1492	-1944	388	-58	156
tw180_lag1	-59	1	-14	-27	-14	-290	-1912	-592	266	-88	-350	-27	-41	-455	-385	-838	-482	-1701	-702	-2020	947	196	-39	125
tw180_lag2	47	80	8	119	108	-28	532	-431	172	-71	-146	-48	-85	35	-106	-444	-629	-1376	-1040	-422	-408	-26	-176	17
tw180_lag3	80	54	45	257	20	-141	2024	2439	-585	-79	-58	94	-28	34	-92	-509	-255	-911	-1260	-1374	1062	-3	-81	-3
tw180_lag4	111	101	-36	96	110	-105	-545	-1696	-285	-135	71	43	-68	-11	-510	-888	-560	-1269	-531	-3198	695	124	-237	-80
tw360_lag1	93	-33	-77	-50	-12	-409	1	1232	307	1	-107	-60	-161	-89	-90	-472	-235	-1546	262	1676	2085	702	119	28
tw360_lag2	-23	29	-125	-109	2	-155	1891	2215	1790	-32	-315	-77	46	598	98	36	-636	-1404	-1069	-607	804	123	70	36
tw360_lag3	57	26	152	101	88	-180	1481	1791	-709	-182	-116	-80	-126	74	-36	39	-665	-1405	-698	-1894	354	211	15	22
tw360_lag4	26	21	99	62	-23	-118	703	1428	985	-171	10	-68	-122	224	8	-480	578	-1427	292	-84	1185	11	-131	-35



CNN - Final Cumulative P&L by Configuration and Hour Ending (Confidence >= 50%)

	HE00	HE01	HE02	HE03	HE04	HE05	HE06	HE07	HE08	HE09	HE10	HE11	HE12	HE13	HE14	HE15	HE16	HE17	HE18	HE19	HE20	HE21	HE22	HE23
tw30_lag1	-385	-46	-291	-201	-375	-215	-1516	-1098	-384	-548	-317	-110	124	-1	-217	-20	-25	-1409	-440	2247	455	526	-367	-169
tw30_lag2	-206	-127	-160	-452	-416	-821	-1961	-1214	-216	-154	-125	-219	-503	-415	-51	148	-200	-1796	-501	2596	2174	12	-286	-125
tw30_lag3	-43	11	-173	-286	-110	-1076	-1977	-948	212	-8	-47	-233	-505	-88	-237	-180	-662	-1945	-671	-3904	716	-47	-85	-46
tw30_lag4	-155	-30	89	-501	-483	-564	-2888	-1793	91	-161	-169	-252	11	-78	-297	73	-550	-827	-722	1358	1071	-29	-249	-276
tw90_lag1	-203	-135	-166	117	50	-152	-35	1388	1454	-541	-429	-132	-83	-76	-59	-432	-384	-1008	-128	-2432	873	227	-3	-35
tw90_lag2	16	-30	9	-75	-82	-280	-1052	-967	-345	-654	-386	-88	-185	427	69	-452	-657	-1053	-905	-3379	-87	520	-139	-63
tw90_lag3	-72	81	-381	-334	-133	-258	-789	-690	-604	-11	-202	-218	-546	-183	-237	-250	-553	-1451	-392	-2091	-357	190	-293	-191
tw90_lag4	-50	97	-274	-263	-116	-67	440	-823	-430	-470	-28	-161	-338	360	-55	-274	-752	-980	-1833	-3041	-766	-168	-352	-315
tw180_lag1	59	-18	-23	25	-193	-480	-346	588	-619	-647	-490	-120	-338	-441	-98	-143	-356	-1818	-1684	336	861	436	-98	-64
tw180_lag2	84	45	-28	103	91	231	-1197	-1023	-109	-281	-393	-103	-55	94	-18	166	-368	-1453	-1389	-2744	-18	242	-202	33
tw180_lag3	-77	-60	-59	72	4	-90	870	297	-492	-505	-308	-51	-294	-428	-27	-98	-928	-1456	-2110	-3928	-115	-739	-122	-43
tw180_lag4	63	85	77	-4	-60	-346	1090	393	-155	-85	-177	14	-70	-5	-329	-124	-880	-442	-777	-3018	-1273	-66	-45	19
tw360_lag1	107	-93	-12	-6	-184	-126	-183	1480	-46	-728	-226	-147	-75	139	-197	-247	-1761	-1663	-1485	-2289	938	94	74	-18
tw360_lag2	178	51	103	-100	262	-101	159	2429	1904	444	-80	84	-17	28	-72	-194	-256	-1465	-341	-1176	281	194	410	175
tw360_lag3	-28	-72	31	-117	17	-63	1059	567	-415	-508	-56	-107	-94	538	-309	-377	-658	-1077	105	-1105	530	184	58	54
tw360_lag4	25	104	-115	-157	-33	-188	1727	1869	339	-633	-402	-108	124	113	-488	-117	-515	-1974	-251	833	-734	-69	-162	-37

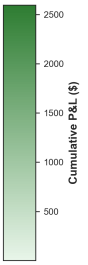


Figure 10. Final cumulative PnL by configuration and hour ending (confidence ≥ 50%) for (a) Transformer, (b) LSTM, and (c) CNN models.

For each hour and day, the model outputs a vector of logits over the five spread bins, which are converted into class probabilities using a softmax layer. The predicted class is taken as the argmax of this probability vector, and the confidence score for that hour is defined as the associated maximum softmax probability. In the trading experiments, a position is initiated only when this confidence score exceeds 50%.

As expected, none of the architectures produce systematically profitable signals across all hours and parameter settings. Instead, profitable opportunities are concentrated in specific regions of the configuration space. The objective is therefore not to trade every hour, but to identify combinations of training time window length, lag parameter, and HE for which the model provides exploitable signals.

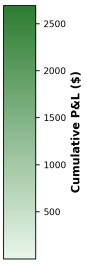
A clear pattern emerges, as the training time window length increases, a larger share of model configurations ends with positive cumulative PnL. Models trained on longer historical windows appear to learn more robust relationships between ERCOT fundamentals and RT-DAM spreads, stabilizing the mapping from loads, wind and solar forecasts, and fuel prices to spread outcomes. Longer windows also span more calendar months and seasonal regimes, exposing the model to winter peaks, shoulder periods, and summer scarcity conditions. This richer seasonal coverage improves the modeling of state-dependent behavior and increases the reliability of the trading signals.

As highlighted in Figure 10, under a 360-day training window, 100%, 92%, and 75% of lag configurations (lags 1-4) across all three architectures produce positive cumulative PnL for HE 6, 7, and 8. Considering all HE, the Transformer architecture finishes the backtest with positive cumulative PnL in 51% of configurations with a 360-day training window, compared with 53% for the LSTM and 40% for the CNN. The clustering of profitable configurations around HE 6, 7, and 8 is consistent with stronger and more systematic ramp dynamics in load and renewable output during the early-morning transition from night-time to daytime conditions.

B) Final Cumulative PnL When Confidence ≥ 95%

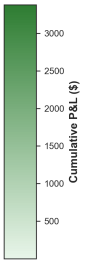
Transformer - Final Cumulative P&L by Configuration and Hour Ending (Confidence >= 95%)

	HE00	HE01	HE02	HE03	HE04	HE05	HE06	HE07	HE08	HE09	HE10	HE11	HE12	HE13	HE14	HE15	HE16	HE17	HE18	HE19	HE20	HE21	HE22	HE23
tw30_lag1	3	36	-7	12	76	-14	-34	-22	20	-20	0	4	33	24	67	95	126	-12	217	1477	-1955	437	-4	-30
tw30_lag2	14	24	17	29	28	18	-34	-56	2	0	0	0	-35	53	145	24	105	2	116	-439	-533	-105	-33	-68
tw30_lag3	-37	8	-2	16	86	30	669	287	-227	-387	-220	-5	-23	26	-18	71	-40	-49	175	-1421	-1908	167	-24	-2
tw30_lag4	-2	9	-10	-9	-23	24	34	-18	0	0	-18	0	1	-17	17	-68	-153	5	90	375	-1805	19	-14	-34
tw90_lag1	57	15	39	92	-30	-152	-1843	-73	35	-8	-29	-10	-40	71	65	1	-40	-28	44	-2166	1438	-179	26	12
tw90_lag2	40	7	21	-15	37	0	-109	9	-1	2	0	22	0	46	89	148	34	-72	221	-457	-1988	-18	-125	9
tw90_lag3	47	10	-21	-10	-0	-18	511	-28	-505	0	0	81	-3	-19	-26	119	89	-104	-252	-1768	-878	216	8	25
tw90_lag4	83	18	33	19	109	44	128	-871	-310	40	-18	-95	-16	-57	52	8	-101	31	-156	204	1138	-46	-8	20
tw180_lag1	29	29	24	21	12	22	-105	-41	-4	-444	-257	90	-310	-355	95	-57	-373	-274	-393	-994	-462	321	-85	-0
tw180_lag2	55	23	6	14	12	10	94	-193	25	0	-10	-23	-21	-27	-22	196	-146	-97	187	-127	2694	447	-10	-3
tw180_lag3	-12	19	0	-2	-25	-4	175	-140	-296	47	0	0	-78	-100	15	166	80	-25	176	-62	-507	44	46	-6
tw180_lag4	74	-7	11	42	56	37	3	-135	215	13	11	30	-35	-7	55	143	-105	59	1005	-321	982	217	-45	11
tw360_lag1	75	33	-3	5	-92	-48	1064	2437	1122	40	2	4	39	28	115	234	-110	-370	-483	-2677	-1245	251	-38	-24
tw360_lag2	61	13	-13	58	-95	-178	-91	1544	1233	0	-26	-5	-25	215	-17	241	111	-63	382	836	1905	251	45	-53
tw360_lag3	14	36	-18	-19	-24	-26	2492	2385	2092	-3	0	49	-61	111	169	220	175	123	71	-2008	1471	408	-38	-68
tw360_lag4	25	19	110	20	-70	49	1773	1585	1454	31	0	0	1	24	20	22	-105	-115	-432	-1270	603	37	70	-28



LSTM - Final Cumulative P&L by Configuration and Hour Ending (Confidence >= 95%)

	HE00	HE01	HE02	HE03	HE04	HE05	HE06	HE07	HE08	HE09	HE10	HE11	HE12	HE13	HE14	HE15	HE16	HE17	HE18	HE19	HE20	HE21	HE22	HE23
tw30_lag1	5	0	7	0	0	0	8	-1	2	0	0	0	0	-14	2	62	24	-26	68	3381	2707	193	49	0
tw30_lag2	0	0	0	0	0	0	3	0	0	0	0	85	35	57	41	-24	6	-46	-39	-2107	-1775	-8	0	0
tw30_lag3	0	0	0	0	0	-4	-64	0	-20	0	-4	-36	3	2	6	27	-21	6	-104	-2011	-134	521	0	0
tw30_lag4	-38	-20	2	2	8	-6	-88	-136	-61	-23	-31	-6	5	28	-45	-76	63	-82	-38	-679	-2364	-127	-6	-5
tw90_lag1	-5	-20	2	7	-16	-31	-1787	-213	-38	-32	-9	7	-59	27	46	344	122	164	-97	1828	-1278	28	-3	0
tw90_lag2	6	0	-19	-28	-16	0	0	15	0	0	-17	0	-26	-20	-80	-76	-71	-344	-0	522	-168	-10	-9	-6
tw90_lag3	44	37	0	-9	-31	-2	-11	62	-63	0	0	-36	16	-17	-5	-16	-4	-37	-64	-304	-248	33	0	18
tw90_lag4	2	0	0	9	21	-0	-1769	-187	1	-17	-28	-8	21	-31	-35	-15	-50	98	141	-1381	-110	69	-5	5
tw180_lag1	67	9	3	0	19	1	-12	-117	-0	0	0	0	0	-4	10	40	-61	-5	31	-1019	-380	-54	-12	-1
tw180_lag2	11	30	6	-13	-14	0	-48	-3	6	0	0	0	13	40	-0	97	7	-405	-39	-962	-615	182	17	-0
tw180_lag3	-11	-3	-10	-4	28	7	4	-24	3	0	0	2	27	49	64	75	-31	-326	180	524	-1373	-4	-1	5
tw180_lag4	13	18	2	63	-26	-13	-3	6	0	0	0	6	39	16	-20	-58	17	20	-423	552	-96	-71	-9	0
tw360_lag1	-6	4	16	14	0	24	917	1488	-169	0	0	0	-43	0	-20	-132	72	32	204	2131	3015	326	0	0
tw360_lag2	-18	-3	13	9	34	10	-46	25	-6	0	-4	0	-4	-14	-15	31	-174	-399	-194	1790	-712	310	0	6
tw360_lag3	-6	-8	8	124	202	45	7	17	0	0	0	-36	-8	-10	-19	-168	-61	-50	110	-3672	-1481	-57	-23	-14
tw360_lag4	0	51	112	-25	-11	-6	13	-1	-171	0	-5	0	-24	13	43	-104	8	59	415	-2707	-1352	1	0	-18



CNN - Final Cumulative P&L by Configuration and Hour Ending (Confidence >= 95%)

	HE00	HE01	HE02	HE03	HE04	HE05	HE06	HE07	HE08	HE09	HE10	HE11	HE12	HE13	HE14	HE15	HE16	HE17	HE18	HE19	HE20	HE21	HE22	HE23
tw30_lag1	-211	-33	-33	-5	-336	-169	-2236	-1613	-426	-452	-234	-13	-42	-40	-48	33	-117	-578	-369	139	670	322	41	-138
tw30_lag2	-32	-25	-62	-192	-276	-130	-1948	-1303	-40	12	-3	5	-329	-432	-66	140	96	-283	-45	1143	471	191	-224	-113
tw30_lag3	-21	18	-42	-7	-9	-348	-1660	-644	134	81	93	-131	7	14	-20	97	-141	-857	-100	2742	1203	171	-18	0
tw30_lag4	10	67	110	-143	-120	-302	-1930	-1429	-61	-49	-94	-54	-26	-12	-23	-300	-195	-121	-18	2001	439	-53	-167	-189
tw90_lag1	6	15	72	11	-47	-189	568	1933	2153	38	-35	-111	-27	-59	46	-132	-53	-330	330	1744	-359	367	-85	-118
tw90_lag2	-57	21	16	-15	-40	-44	-21	-934	-510	0	-3	-56	-16	-102	50	-70	-192	-319	-280	465	531	527	-88	35
tw90_lag3	-12	54	-41	-3	3	5	-1915	54	-140	-19	-46	60	-15	-106	3	84	-135	-553	-62	-4361	-315	378	-3	17
tw90_lag4	16	16	7	79	50	6	561	49	-45	-25	-13	-13	19	92	4	29	-262	-359	-772	398	-538	155	-115	5
tw180_lag1	57	22	8	-27	-187	-184	-827	-299	-300	-459	-299	-105	-93	-356	53	-121	-53	-837	-365	-1779	-45	291	-100	-27
tw180_lag2	109	10	12	-14	30	-10	540	556	-117	-107	-15	-51	7	81	284	103	317	-579	-263	-1574	292	469	-25	1
tw180_lag3	-2	-31	-24	14	56	38	524	74	60	-2	-18	38	-47	-26	43	-106	137	-379	-217	-3237	400	461	-53	-5
tw180_lag4	61	14	-2	-5	-4	-12	-123	180	-2	-26	43	57	81	121	-112	23	21	-214	-177	-3969	-936	112	-51	7
tw360_lag1	37	23	31	-7	-33	-255	481	3090	1351	-196	112	61	26	9	-6	-19	-26	-496	-167	-288	-1329	61	152	-7
tw360_lag2	-16	-12	3	5	6	23	2274	2458	1419	3	104	76	25	92	29	77	-28	-687	799	700	2065	358	44	-12
tw360_lag3	4	-7	14	5	-5	19	506	59	70	-387	-19	-79	-20	-35	-43	-116	-87	-145	237	-1473	-153	187	-10	-37
tw360_lag4	-1	21	2	-23	-28	-22	-82	1427	-26	-33	-8	-3	240	461	-109	16	-343	-304	-24	-652	-104	352	7	5

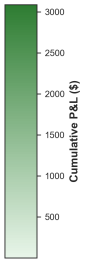


Figure 11. Final cumulative PnL by configuration and hour ending (confidence ≥ 95%) for (a) Transformer, (b) LSTM, and (c) CNN models.

In Figure 11, increasing the confidence threshold from 50% to 95% naturally reduces the number of eligible trades, because the softmax output concentrates probability mass on a single class only when the model is highly certain relative to the alternatives. In principle, higher softmax probabilities should correspond to sharper posterior beliefs and more reliable predictions. However, deep neural networks are known to produce overconfident and miscalibrated probability estimates, particularly in high-dimensional classification problems (Guo et al., 2017) [9]. Since the softmax function reflects relative logit magnitudes rather than a calibrated measure of epistemic uncertainty, a larger probability does not necessarily guarantee better economic outcomes. Nonetheless, confidence filtering remains a useful diagnostic: it isolates the subset of forecasts where the model expresses strong internal agreement, allowing an assessment of whether these self-identified high-certainty predictions translate into superior PnL performance.

Under a 360-day training window, raising the confidence threshold from 50% to 95% leaves 84% of all configurations across the three architectures with positive cumulative PnL for hours ending 6, 7, and 8. This indicates that the early-morning signal is robust even when the analysis is restricted to only the highest-confidence forecasts.

Extending the analysis to all HE shows that the impact of a higher confidence threshold is heterogeneous across models. For the Transformer, the fraction of configurations that end with positive cumulative PnL increases from 51% to 59%, suggesting that high-confidence filtering helps discard lower-quality trades. For the CNN, this share also rises, from 40% to 51%. In contrast, the LSTM's share of profitable configurations declines from 53% to 39%, indicating that LSTM confidence scores are less aligned with economic performance in this setting.

C) Cumulative PnL Graphs for HE 6, 7, and 8 when Confidence ≥ 50%

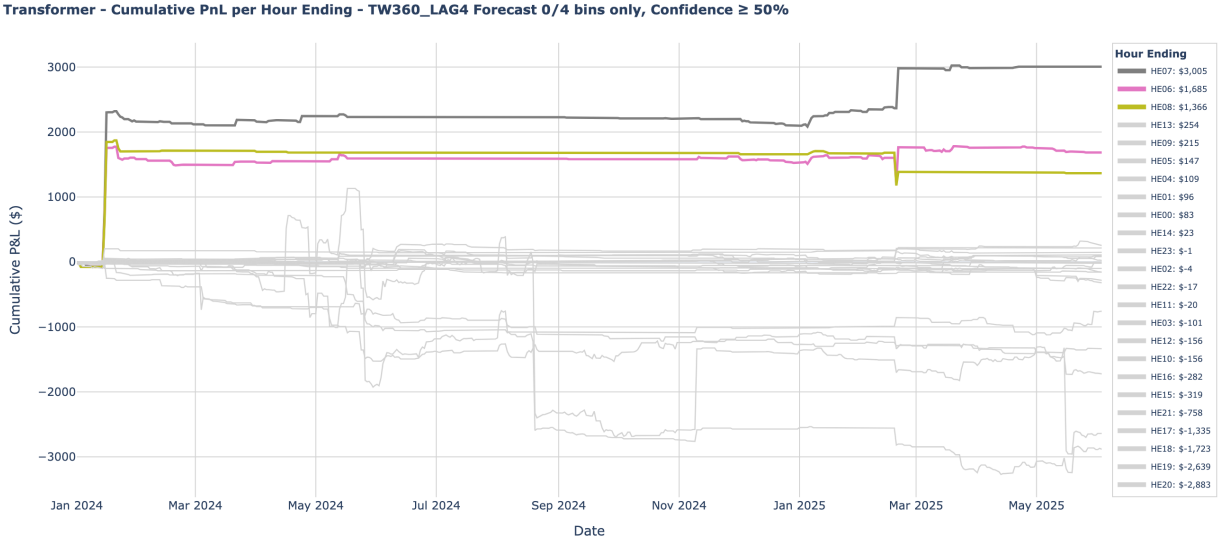


Figure 12. Transformer cumulative PnL by hour ending when confidence ≥ 50%. Colored lines correspond to HE 6, HE 7, and HE 8; all other hours are shown in gray.

Figure 12 shows that the Transformer architecture successfully captures the two major spread dislocations highlighted for HE 6, HE 7, and HE 8. Outside of these episodes, the PnL profile remains relatively flat, indicating that the model rarely forecasts extreme spreads in bin 0 or bin 4. This behavior is consistent with a model that takes directional risk primarily when it detects large, state-dependent deviations rather than continuously trading on small, noisy signals.

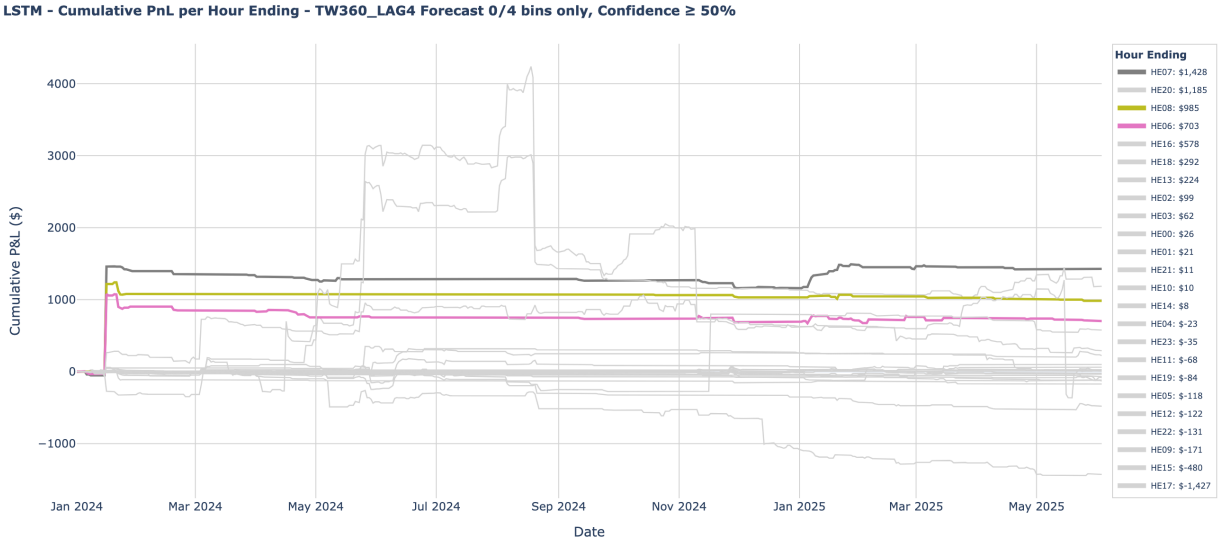


Figure 13. LSTM cumulative PnL by hour ending when confidence ≥ 50%. Colored lines correspond to HE 6, HE 7, and HE 8; all other hours are shown in gray.

In Figure 13, the LSTM architecture produces results broadly similar to the Transformer for HE 6, HE 7, and HE 8, but it fails to capture the second large spread event in February 2025.

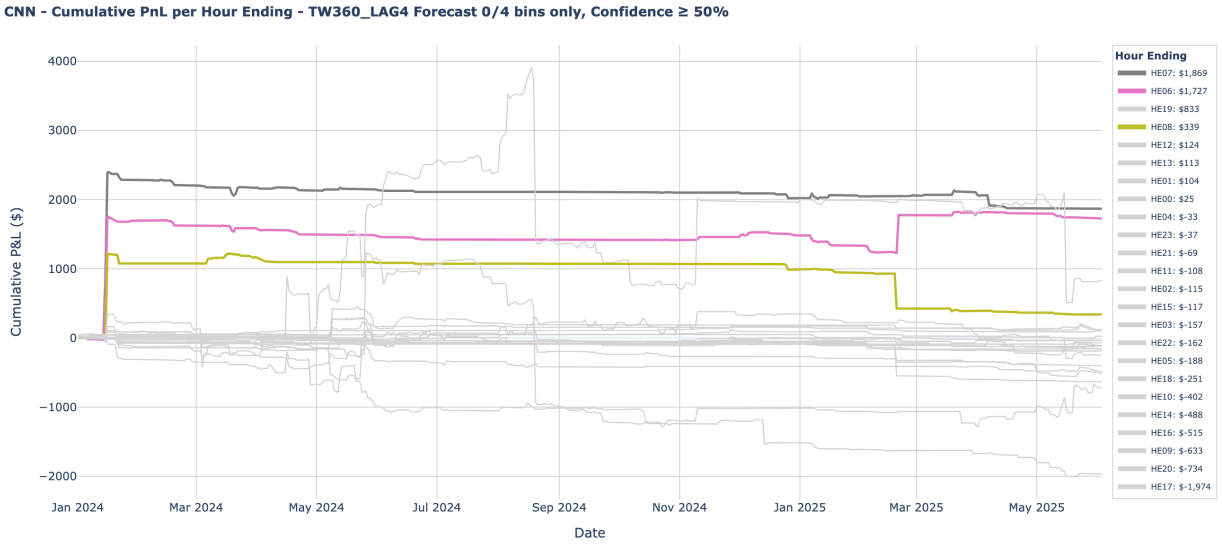


Figure 14. CNN cumulative PnL by hour ending when confidence ≥ 50%. Colored lines correspond to HE 6, HE 7, and HE 8; all other hours are shown in gray.

In Figure 14, the CNN architecture produces results broadly similar to the Transformer for HE 6, HE 7, and HE 8, but it forecasts the opposite sign of the spread for the second large spread event in February 2025 for HE 8. Similar to the LSTM in Figure 13, the CNN for HE 19 generates a relatively smooth, positive PnL trajectory up to August 2024, after which performance deteriorates. This pattern again illustrates how a single large misclassification around a major spread event can materially distort the cumulative economics of an otherwise stable strategy.

D) Cumulative PnL Graphs for HE 6, 7, and 8 when Confidence ≥ 95%

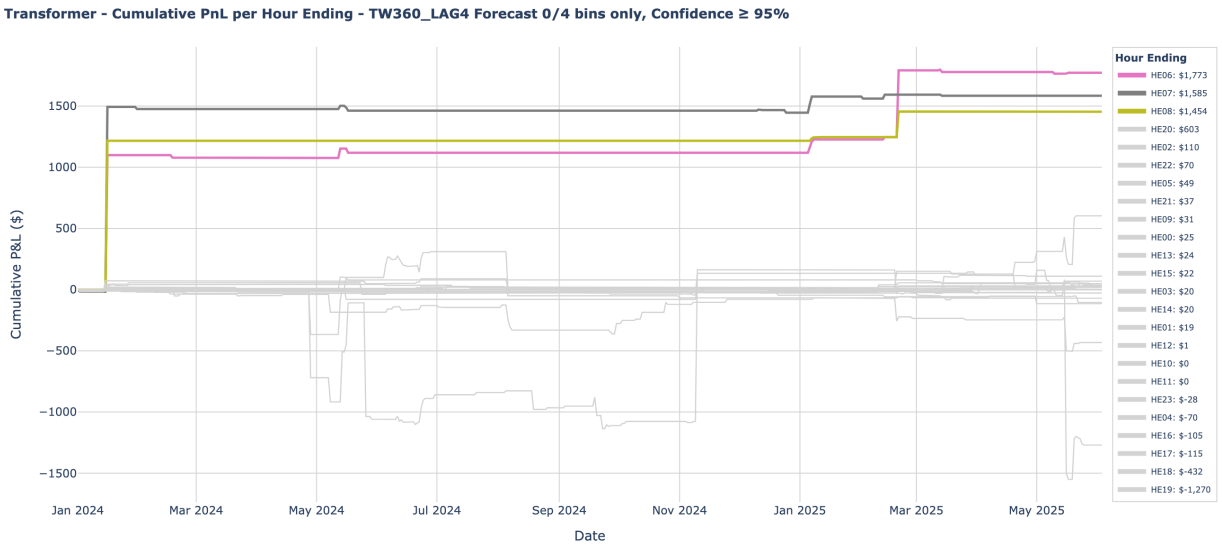


Figure 15. Transformer cumulative PnL by hour ending when confidence ≥ 95%. Colored lines correspond to HE 6, HE 7, and HE 8; all other hours are shown in gray.

In Figure 15, the 95% confidence results show that the Transformer architecture still successfully captures the two major spread dislocations for HE 6, HE 7, and HE 8.

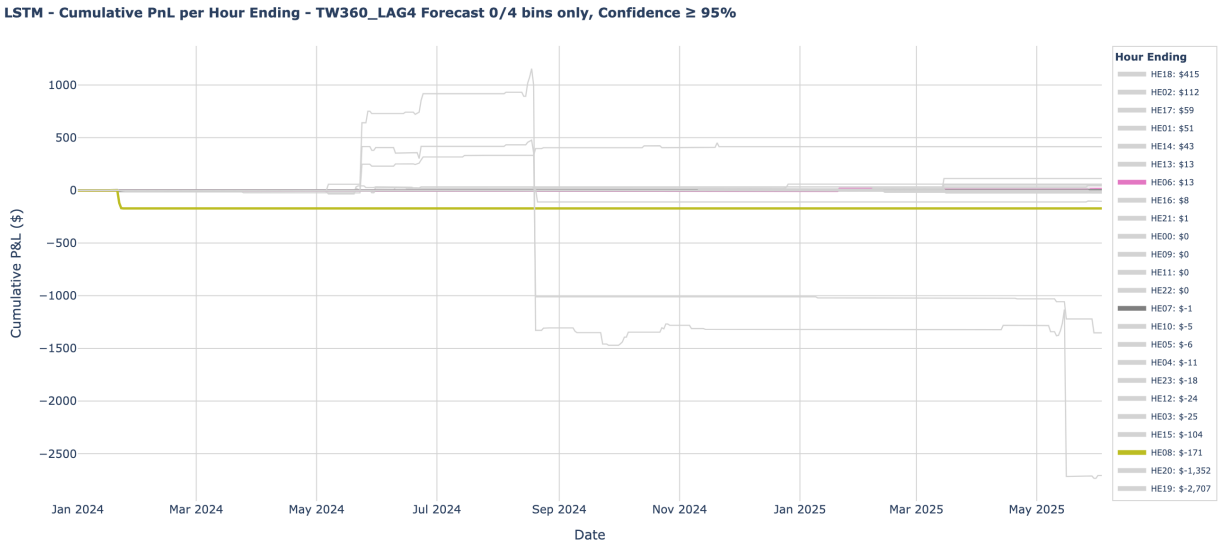


Figure 16. LSTM cumulative PnL by hour ending when confidence ≥ 95%. Colored lines correspond to HE 6, HE 7, and HE 8; all other hours are shown in gray.

Although there is no significant negative PnL for HE 6, HE 7, and HE 8 (apart from a single loss event in HE 8 in February 2024), Figure 16 shows that, under the 95% confidence threshold, the LSTM misses several opportunities it had exploited when the threshold was set at 50%. In other words, tightening the confidence filter removes some bad trades but also filters out many of the previously profitable ones, indicating that the LSTM's softmax confidence is not strongly aligned with the economic value of its signals.

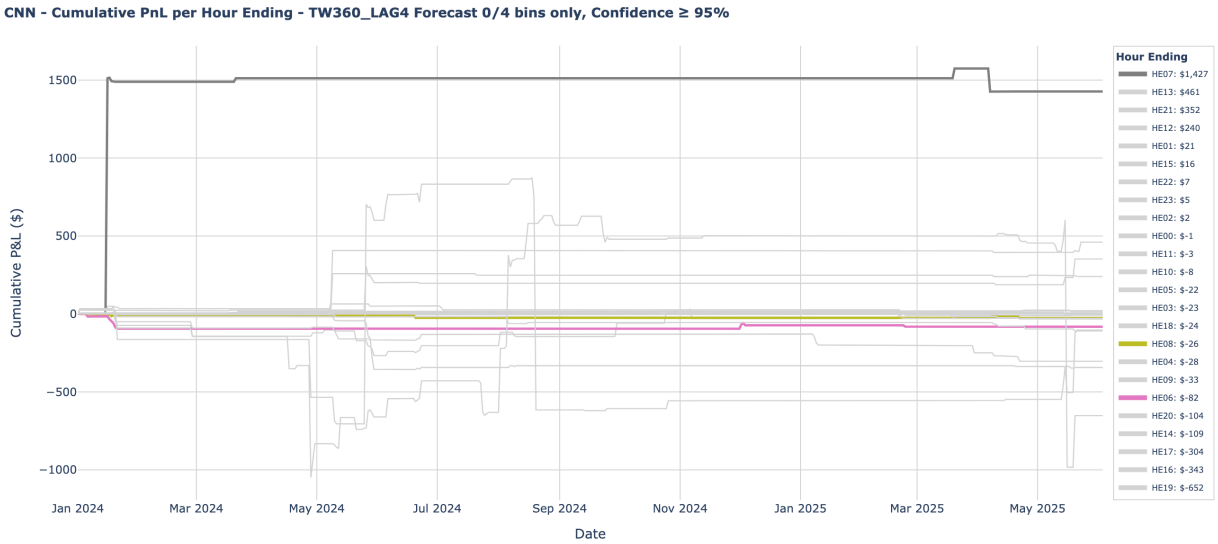


Figure 17. CNN cumulative PnL by hour ending when confidence ≥ 95%. Colored lines correspond to HE 6, HE 7, and HE 8; all other hours are shown in gray.

Although there is no significant negative PnL for HE 6 and HE 8, Figure 17 shows that, under the 95% confidence threshold, the CNN misses several opportunities it had successfully exploited when the threshold

was set at 50%. It still manages to capture one of the major spread dislocations in HE 7, but many other potentially profitable events are filtered out by the stricter confidence requirement.

E) Confusion Matrices - 50% Confidence

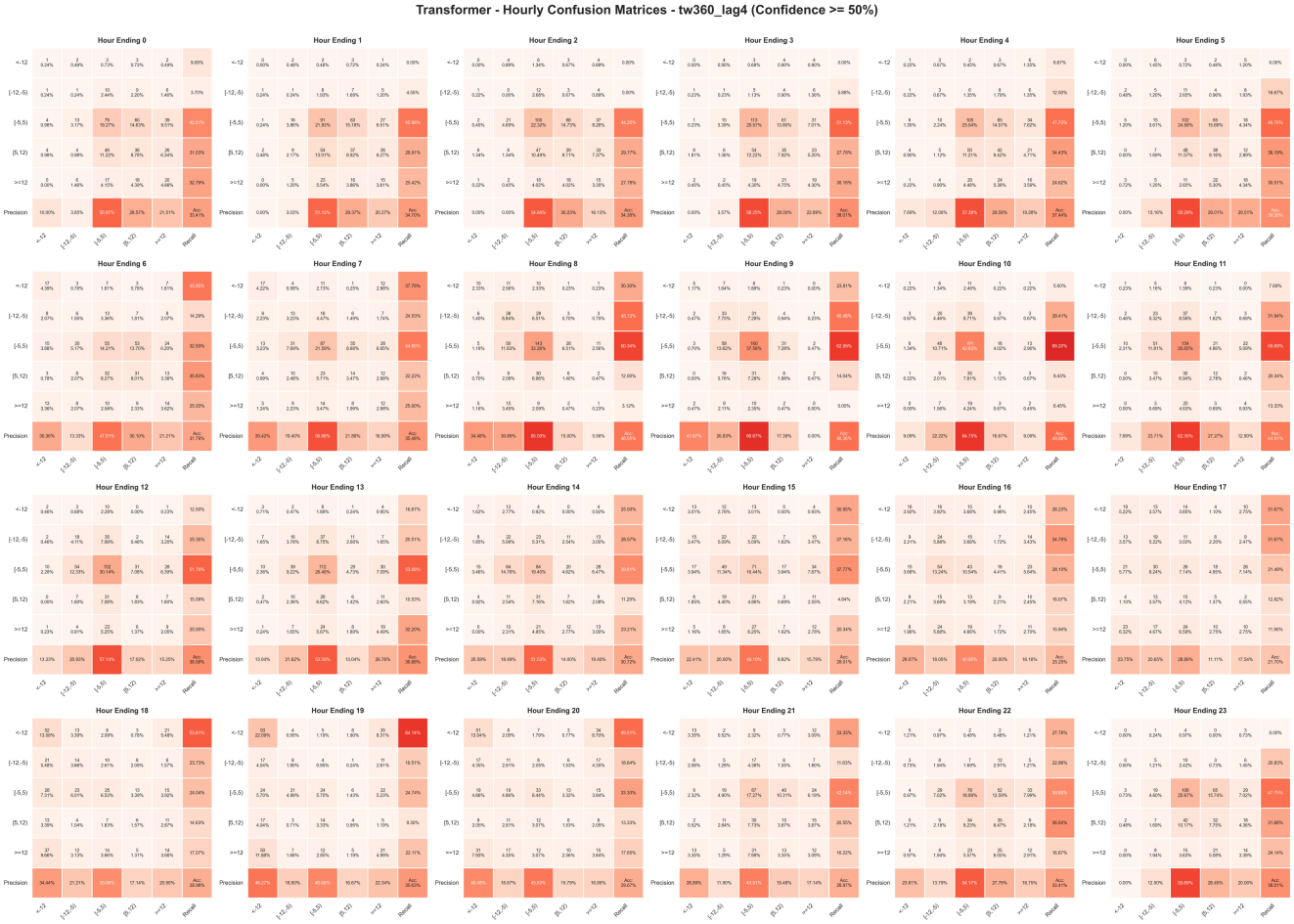


Figure 18. Hourly confusion matrices for the Transformer model (configuration tw360_lag4) across all hour endings (HE 0–23) at a 50% confidence threshold.

At the 50% confidence threshold, the cumulative PnL indicates that HE 6, HE 7, and HE 8 are economically attractive for the Transformer models, even though the corresponding precision scores are only in the 30-35% range as seen in Figure 18. This pattern suggests that the algorithm correctly identifies the largest extreme spread opportunities, while many of the misclassified trades generate only small losses on average. When the confidence threshold is raised to 95%, the precision scores for these hours exceed 50% and the strategies remain economically profitable, indicating a closer alignment between statistical precision and economic performance at higher confidence levels. In addition, HE 19 exhibits the strongest combination of recall and precision for the extreme spread classes. Across many other HE, recall for the central bin (-$5, $5) exceeds 50%, indicating that the models also perform reasonably well at identifying “normal” convergence hours with spreads near zero.

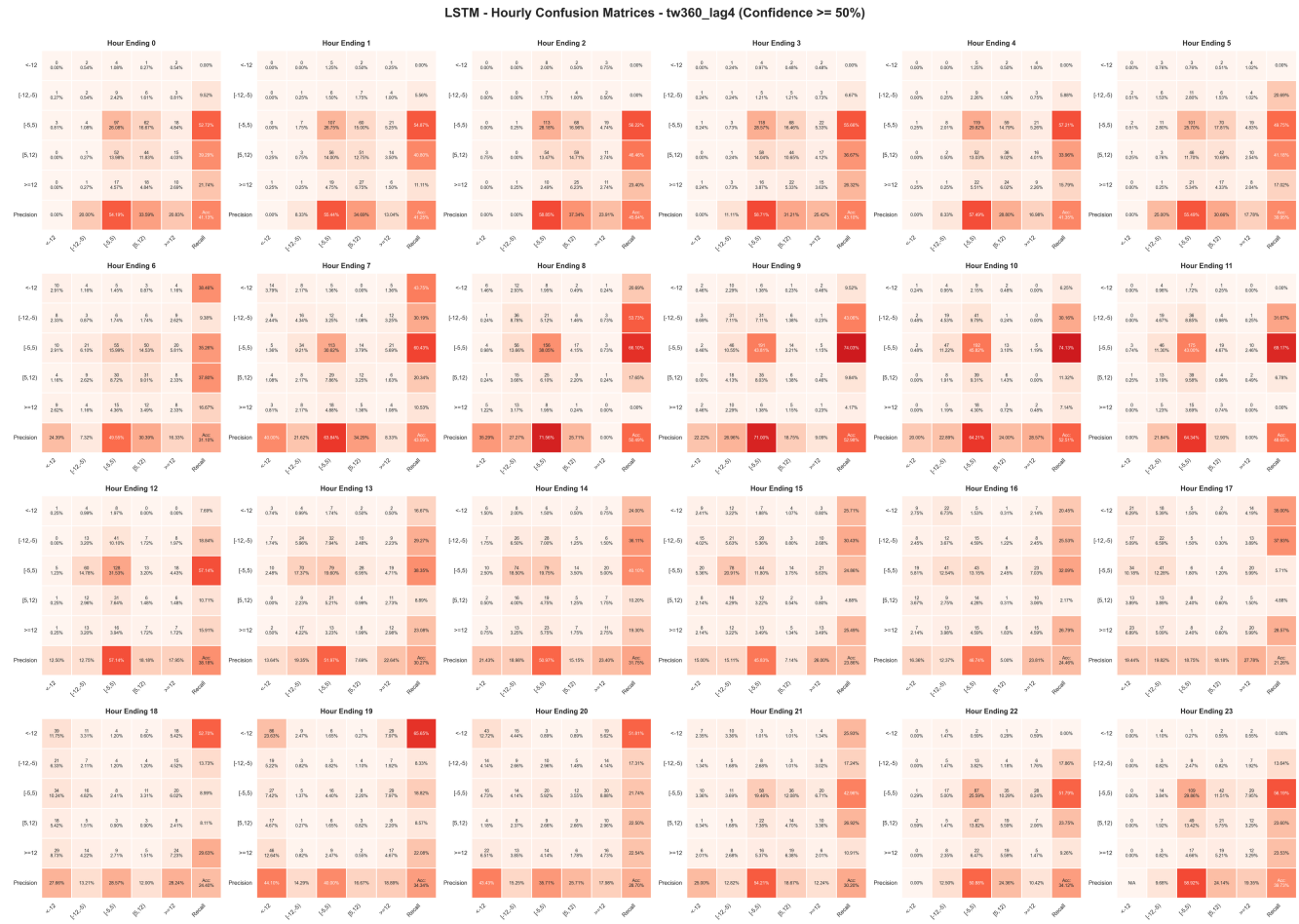


Figure 19. Hourly confusion matrices for the LSTM model (configuration tw360_lag4) across all hour endings (HE 0–23) at a 50% confidence threshold.

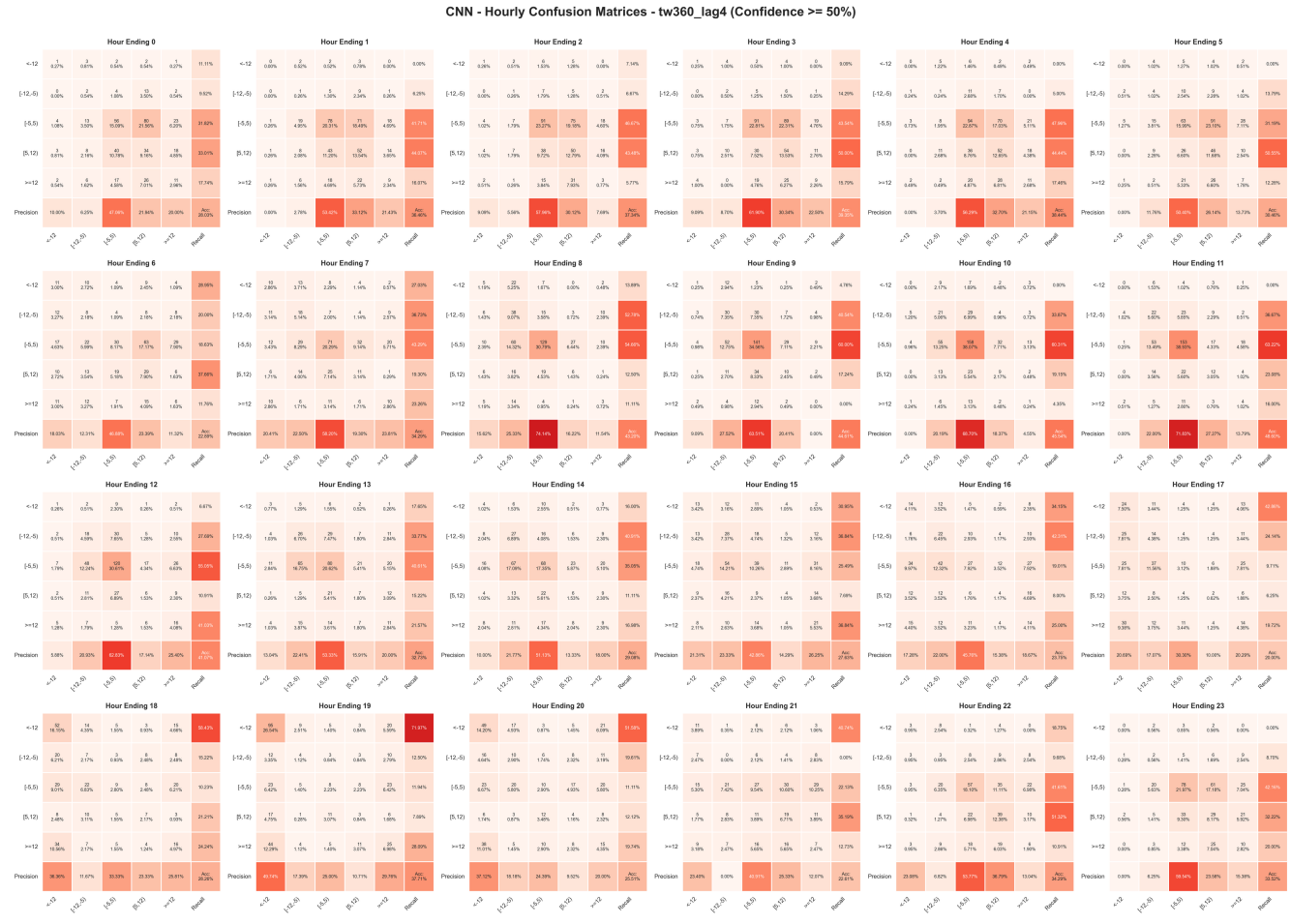


Figure 20. Hourly confusion matrices for the CNN model (configuration tw360_lag4) across all hour endings (HE 0–23) at a 50% confidence threshold.

When the confusion matrices for LSTM and CNN are examined at a 50% confidence threshold, no HE achieves precision above 50% in either model. Hour 19 stands out for CNN as the comparatively strongest case, exhibiting relatively high precision and recall. At the same confidence level and restricting attention to the extreme spread bins (0 and 4), the Transformer attains the highest precision in 34% of the hours, compared with 29% for LSTM and 27% for CNN, while in the remaining hours no single model dominates (ties).

F) Confusion Matrices - 95% Confidence

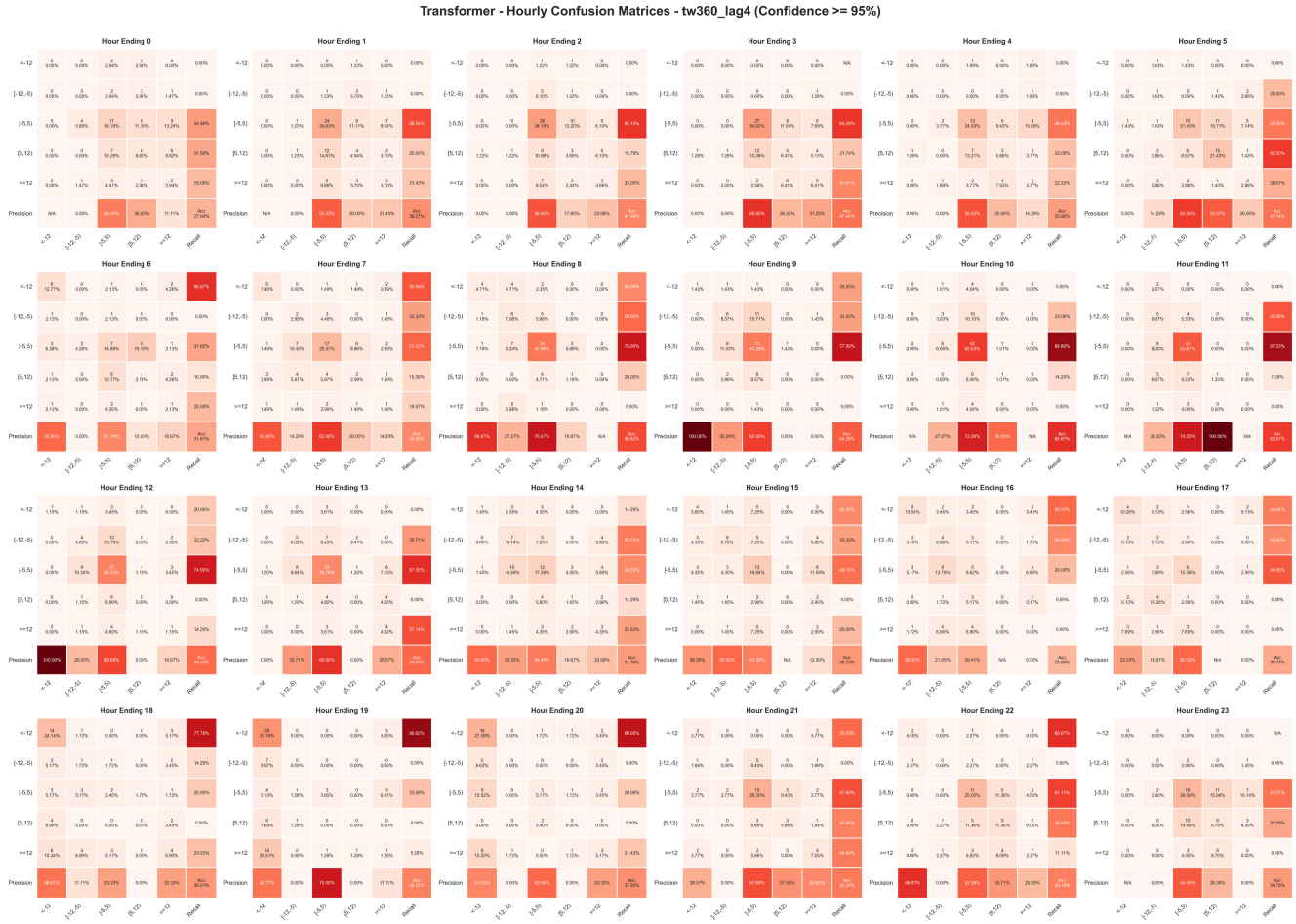


Figure 21. Hourly confusion matrices for the Transformer model (configuration tw360_lag4) across all hour endings (HE 0–23) at a 95% confidence threshold.

When the confidence threshold is raised from 50% to 95%, bin 0 precision exceeds 50% for HE 6, 7, and 8, as discussed earlier. In particular, HE 9 and HE 12 achieve a bin 0 precision of 100%. This higher threshold produces a marked improvement in precision across most HE, but it also implies that trades are executed only when the model is extremely confident, causing some profitable opportunities with confidence below 95% to be skipped. In this sense, a 95% threshold represents a conservative, safety-oriented trading regime that prioritizes reliability over breadth of participation in the market.

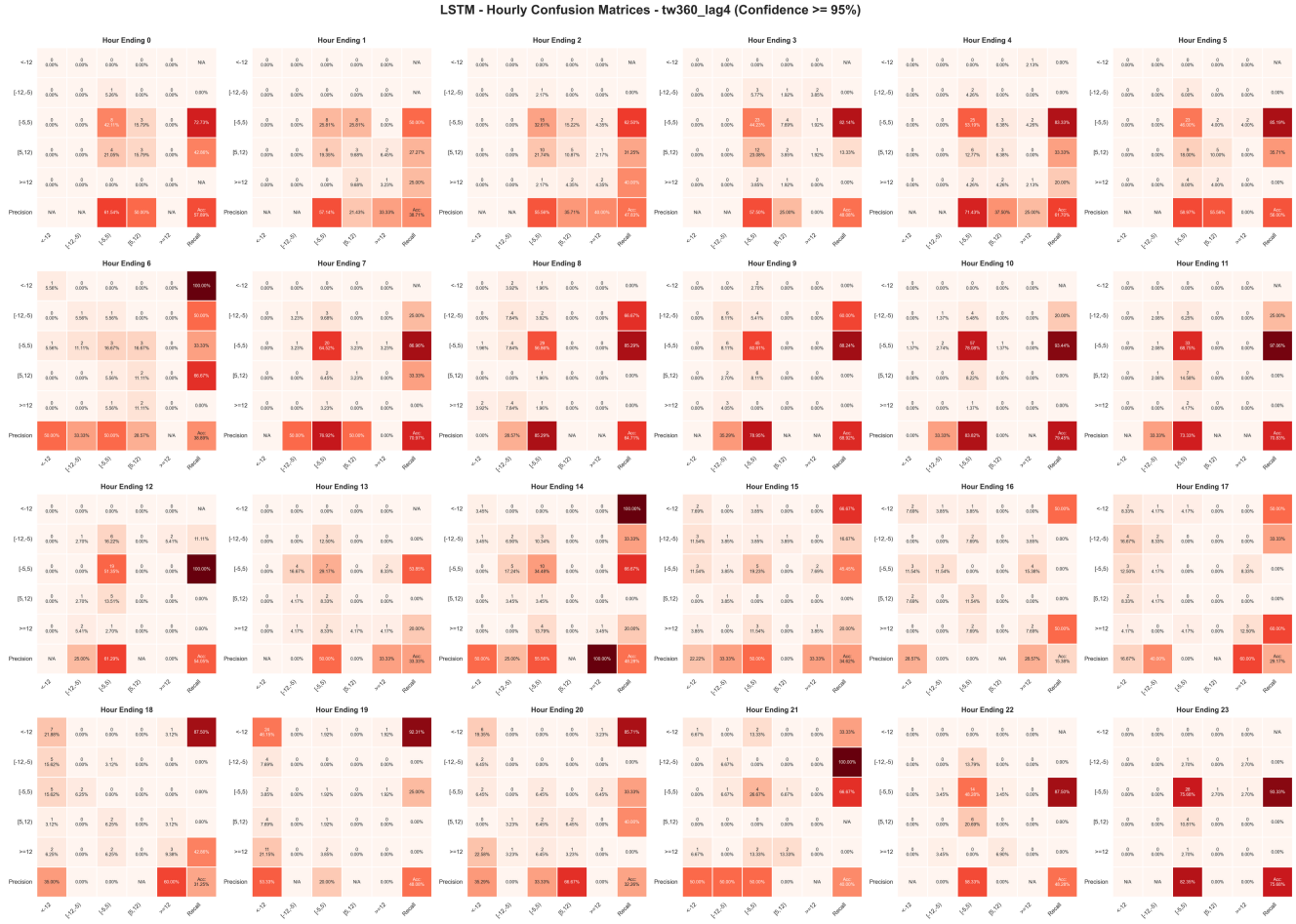


Figure 22. Hourly confusion matrices for the LSTM model (configuration tw360_lag4) across all hour endings (HE 0–23) at a 95% confidence threshold.

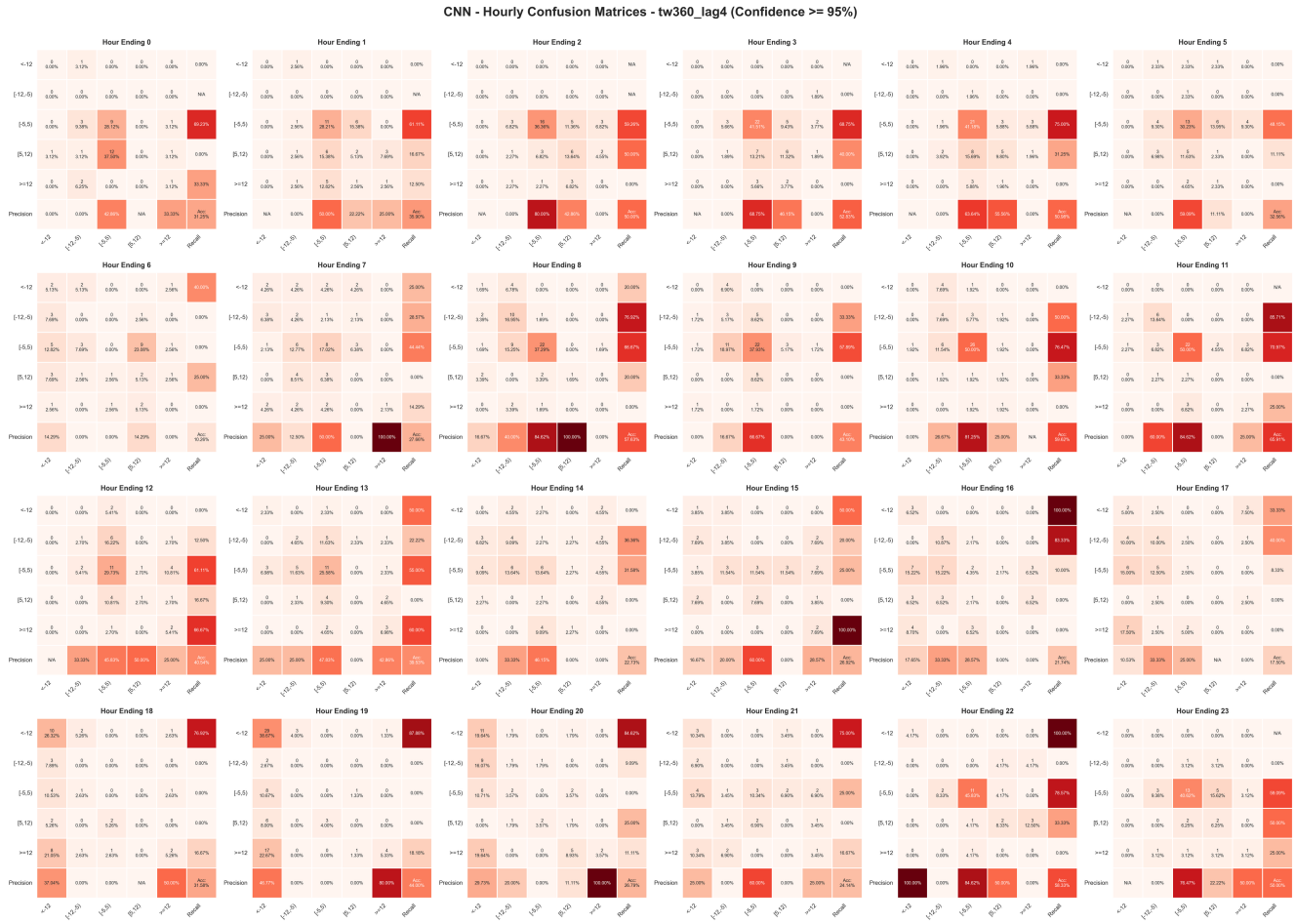


Figure 23. Hourly confusion matrices for the CNN model (configuration tw360_lag4) across all hour endings (HE 0–23) at a 95% confidence threshold.

When the confusion matrices for LSTM and CNN are examined at a 95% confidence threshold, LSTM exhibits four HE with precision above 50%, whereas CNN shows this level of precision for only two HE. HE 14 in the LSTM case attains 100% precision because it generates a single bin 4 prediction over almost two years of trading and classifies it correctly. Similarly, HE 7 and HE 20 for CNN each reach 100% precision by correctly identifying three bin 4 events during the same period.

VI. Contributions

This study develops a high-dimensional Transformer framework for forecasting ERCOT RT-DAM spreads and evaluates it under a walk-forward process that reflects real virtual-bidding workflows. The analysis focuses on the extreme spread outcomes that drive most virtual-bidding profitability and compares performance across multiple model families and configurations. The results highlight consistent, predictable patterns in early-morning hours particularly HE 6-8, where Transformer-based architecture show a clear advantage over LSTM and CNN models. The study also examines confidence filtered trading, showing that confidence thresholds can improve the economic value of Transformer forecasts and expose calibration issues in other architectures.

VII. Limitations

First, it relies on ERCOT System Lambda rather than nodal or hub-level LMPs, so congestion and basis risk central to real trading are not captured. Second, while renewable forecasts are included, the models do not account for forecast uncertainty or intraday revisions that often drive real-time movements. Third, the use of softmax confidence provides only a rough proxy for uncertainty, which can lead to unreliable economic decisions without further calibration. Finally, the trading evaluation abstracts away from practical market frictions such as liquidity, bid caps, budget constraints, and slippage, all of which would influence real-world execution.

VIII. Future Work

Expanding the framework to nodal and hub-level LMPs would enable evaluation of strategies that account for congestion and spatial price dynamics. Combining Transformer architectures with more advanced or physics informed renewable forecasting could improve prediction of extreme outcomes. Incorporating calibrated probabilistic methods may also strengthen the link between model confidence and real trading performance. Finally, applying the approach to other U.S. power markets, such as CAISO, PJM, or MISO, would help determine whether the observed patterns, including early-morning ramp predictability, hold beyond ERCOT.

Resources

[1] 2024 STATE OF THE MARKET REPORT FOR THE ERCOT ELECTRICITY MARKETS - Potomac Economics, May 2025

[2] Deep Learning-Based Electricity Price Forecast for Virtual Bidding in Wholesale Electricity Market - Jiaxuan Gong, Zhongyang Zhao, Michael H. Liao

[3] Effect of wind generation on ERCOT nodal prices Chen-Hao Tsai, Derya Eryilmaz

[4] J. Xu and R. Baldick, “Day-ahead price forecasting in ercot market using neural network approaches, in *Proceedings of the Tenth ACM International Conference on Future Energy Systems*, 2019, pp. 486-491.

[5] J. Trebbien, S. Pütz, B. Schäfer, H. S. Nygård, L. R. Gorjão, and D. Witthaut, “Probabilistic forecasting of day-ahead electricity prices and their volatility with lstms,” in *2023 IEEE PES Innovative Smart Grid Technologies Europe (ISGT EUROPE)*. IEEE, 2023, pp. 1-5.

[6] Rolling and Day-Ahead Forecasting of Electricity Market Prices: Evaluating the Performance of DL Models with Bayesian Optimization - Sajedeh Darvishi Niafenderi, Paras Mandal - Published in Annual Meeting of the IEEE - 29 October 2023

[7] Transformer versus LSTMs for electronic trading - Paul Bilokon, Yitao Qiu - Sep 2023

[8] Y. Li, “Data-driven modeling and algorithmic trading in electricity market,” Ph.D. dissertation, UC Riverside, 2024.

[9] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). *On Calibration of Modern Neural Networks*. In Proceedings of the 34th International Conference on Machine Learning (ICML 2017).