Data Flow 2025

Mastering the Data Waves

Application of Machine Learning for Business Revenue Forecasting and Analysis: A Case Study on a 12-Year Dataset from a U.S. Fashion Company

Chi Thuong Doan[a,1], Thai Duong Quach[a,2], Tuan Long D.Nguyen[a,3], The Phong Nguyen[a,4]

[a]Team FaMI, Faculty of Mathematics and Informatics, Hanoi University of Science and Technology (HUST)

K E Y W O R D S

- Revenue
- Lasso Regression
- EMD-Hilbert.T
- Model: Hilbert.HT-Esemble-Deep-RVFL, LSTM (self-attention), Bi-LSTM, transformer, XGBoost, RF, ARIMA, SARIMA

A B S T R A C T

Revenue forecasting is a crucial aspect of business strategy, particularly for large enterprises. Accurate predictions play a decisive role in determining a company's success. This report introduces a hybrid approach using EMD-HHT for signal decomposition, Lasso Regression for feature selection, and Ensemble-Deep-RVFL for prediction. The model achieves high accuracy (**R = 0.9852, MAPE = 0.0659** for one-month forecasts), outperforming traditional methods. Even for long-term predictions (12 months), it maintains strong performance (**R = 0.9175, MAPE = 0.2359**). These results demonstrate the model's potential for improving business decision-making and financial planning.

1. Introduction

In today's dynamic business environment, forecasting plays an important role in planning and decision-making. As businesses accumulate vast amounts of data, the need for advanced forecasting models becomes much more essential. Classical time series forecasting models such as ARIMA (Box& Jenkins, 1970) and machine learning models such as LSTM (Hochreiter& Schmidhuber, 1997), Random Forest (Breiman, 2001), XGBoost (Tianqi Chen, 2016), and Transformer (Vaswani et al., 2017) have been widely studied. Beside these approachs, we develop a hybrid HHT-Esemble-Deep-RVFL model to solve the problem in this report. Specifically, we utilize Lasso Regression, which enhances generalization ability and reduce complexity ([1], Muthukrishnan R., Rohini R., 2016), for feature selection. Meanwhile, we use RVFL for its less affected by high correlation ([2], Zhang, 2016), and Ridge Regression within RVFL for mitigating mitigating correlation issues and improving robustness ([3], Toai.T.K, 2023). The effectiveness of the model has been demonstrated through the following experiments.

2. Materials and methods

a. Lasso Regression applies L1 regularization, optimizing the objective function:

$$\mathcal{L}(\beta) = \sum_{i=1}^{n}(y_i - X_i\beta)^2 + \lambda\sum_{j=1}^{p}|\beta_j| \tag{1}$$

under the optimality condition constrained by Karush-Kuhn-Tucker (KKT). A feature X_j is removed if: $|X_j^T(y - X\beta)| \leq \lambda$.

b. Hilbert-Huang Transform (HHT) separates signals into different frequency components in two steps. First, the signal is decomposed using EMD to extract Intrinsic Mode Functions (IMFs). Then, Hilbert Transform is applied to compute the instantaneous frequency for each IMF ([4], N.E. Huang , 1998).

c. Esemble-Deep-RVFL ([5], Shi&Qiushi, 2021), ([6], A.K.Malik&Gao, 2023): This model differs from others because it does not rely on gradient descent for updating weights but utilizing Ridge Regression at each hidden layer for transformation. For the initialization step, it is similar to an MLP network. However, concatenation is used to combine outputs from each hidden layer. Then, the regression weight β is computed, and an ensemble is performed on the β weight vector after the final hidden layer $\beta = (H^T H + \lambda I)^{-1} H^T Y$, $Y_{pre_i} = \sum_{i=1}^{n} X\beta_i$ (H_i is the output of the i-th hidden layer, and X is input of testset):

$$Y_{Esemble_{pre}} = \frac{1}{N}\sum_{i=1}^{n} Y_{pre_i} \qquad (2)$$

3. Results and discussion

Data analysis: The dataset consists of 901,561 training records and 74,682 test records. In this report, we focus on Revenue and COGS. To align with real-world scenarios, we handle 41 missing Revenue values in the test set and calculate the total Revenue and COGS for each month.

Table 1: Statistical values for numerical data variables.

Variable	Revenue		COGS	
	Train	Test	Train	Test
Mean	40448616	31684117	33253322	27386879
Std	17415717	12521258	14274530	10764054
Min	13837950	5274388	11222875	4605905
Max	87273286	48282093	72393727	40971837
CV(σ/μ)	0.4306	0.3952	0.4293	0.3930

Through statistical analysis, it is shown that both Revenue and COGS have a wide distribution range from min to max. Additionally, the coefficient of variation (CV) is very high (approximately 40% in both the train and test sets), indicating that revenue fluctuations across months are not only strong but also quite complex.

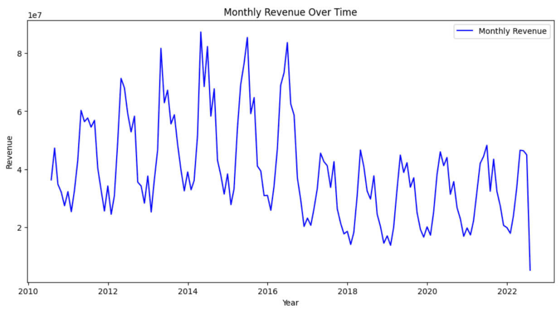

Fig. 1. The original signal revenue.

From the analysis of the original business value signal, it can be observed that the signal trend is relatively clear. Revenue often peaks in mid-year (June, July), while the lowest values are recorded at the beginning of the year. The Dickey-Fuller stationarity test yielded an ADF Statistic: -0.4401 and a p-value: 0.9032, confirming that the revenue series is non-stationary and highly nonlinear. The approach we take to the model is analyzed based on data characteristics. In ([7], Norden E. Huang, 2005), it is shown that the effective application of HHT in financial analysis is beneficial. Additionally, ([8], E. Huang & associates, 2003) indicated that HHT is highly suitable for data with clear trends, stability, and strongly nonlinear, non-stationary time series.

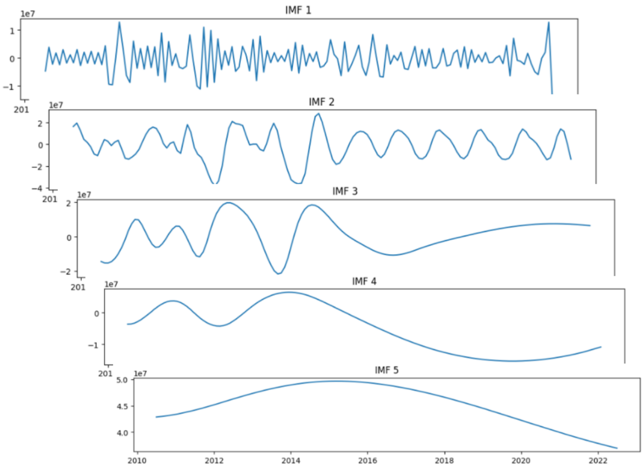

Fig. 2. The (IMFs) of the revenue signal.

Table 2: Optimized model adjustment for the forecasting.

Models	Best parameters
LSTM*	Hidden: 256, batch: 16, optim: 'Adam'
ARIMA*	(p,d,q): (5,0,5)
SARIMA*	(p,d,q): (5,0,5), Seasonal Order:(1, 0, 0, 12)
ED-RVFL+	num_nodes: 50, regular_para: 0.00001, num_layer: 15
XGBoost+	bytree: 0.5645, gamma: 7.1191, lr: 0.1334, max_depth: 1, min_child_weight: 9, n_estimators: 256
Transformer+	MH-Attention: 8, Dim-FFN: 16, batch: 32, optim: 'Adam'
Bi-LSTM+	Bidirectional: 256, Hidden: 64, Dense ouput: 64, batch: 32, optim: 'Adam'
RF+	max_depth: 3, min_samples_leaf: 2, min_samples_split: 18, n_estimators: 10, random_state: 42
LSTM (self-Attention)+	Hidden: 64, lr: 0.001, batch: 32, optim: 'Adam'

Note: *: original data, +: HHT-signal data.

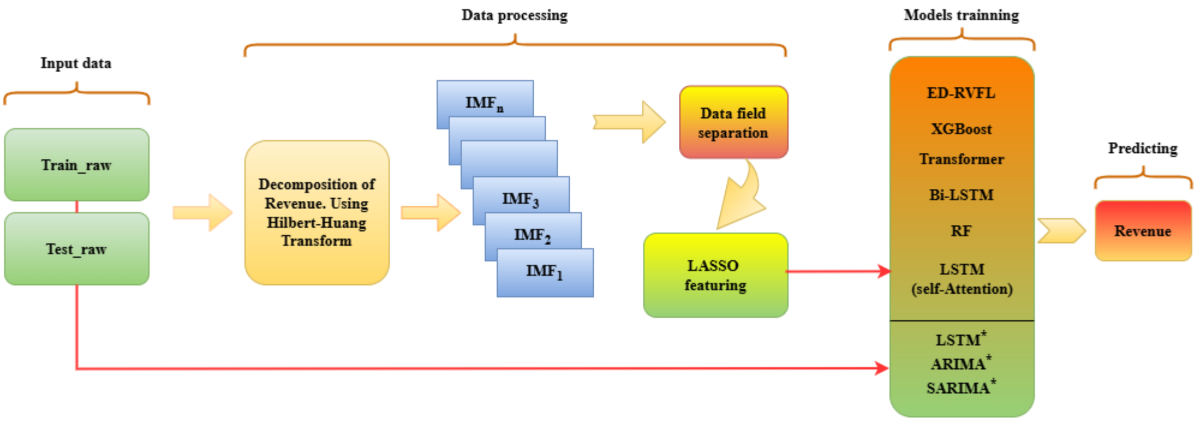

Fig. 3. The flowchart of the study.

Table 3: Results of the best one-step forecast for Revenue across models. Note: The best model is highlighted in bold.

Model	Dataset	R	MAPE	RMSE/σ	NSE	IA
LSTM*	Train	0.8362	0.2081	0.5486	0.6990	0.9032
	Test	0.4394	0.5788	0.8983	0.1931	0.7440
ARIMA*	Train	0.7774	0.3264	0.6289	0.6044	0.9118
	Test	0.6810	0.4281	0.7128	0.4637	0.7985
SARIMA*	Train	0.7774	0.3264	0.6289	0.6044	0.9118
	Test	0.7696	0.3995	0.6215	0.5923	0.8439
ED-RVFL$^+$	**Train**	**0.9891**	**0.0599**	**0.1470**	**0.9784**	**0.9944**
	Test	**0.9852**	**0.0659**	**0.1714**	**0.9706**	**0.9927**
XGBoost$^+$	Train	0.9720	0.0880	0.2342	0.9447	0.9851
	Test	0.8757	0.3091	0.4700	0.7668	0.9234
Transformer$^+$	Train	0.9208	0.1049	0.2755	0.9208	0.9786
	Test	0.6885	0.2051	0.5582	0.6885	0.9195
Bi-LSTM$^+$	Train	0.8981	0.1163	0.3124	0.8981	0.9721
	Test	0.8486	0.1626	0.3891	0.8486	0.9557
RF$^+$	Train	0.8368	0.1908	0.2372	0.7003	0.8873
	Test	0.7172	0.3994	0.6783	0.5143	0.7953
LSTM (self-Attention)$^+$	Train	0.9472	0.1170	0.3205	0.8973	0.9721
	Test	0.9269	0.2182	0.3753	0.8591	0.9576

From the obtained results, it shows that using raw data for classical time series models (ARIMA, SARIMA) or LSTM yields relatively low performance. The RMSE/σ ratio is higher than the natural variation (CV), implying the need for a model capable of deeper feature analysis. Additionally, analyzing the signal using the HHT method and applying Lasso enhances access to various multivariate models and optimizes feature extraction. The combined HHT-ED-RVFL model provides the best prediction performance with evaluation metrics R: **0.9852**, MAPE: **0.0659**, RMSE/σ: 0.1714. Moreover, the two evaluation metrics for model performance, NSE and IA record impressive results, respectively, 0.9706 and 0.9927. They indicate a very good fit between the forecast values and the actual values. These results demonstrate the robustness of the model.

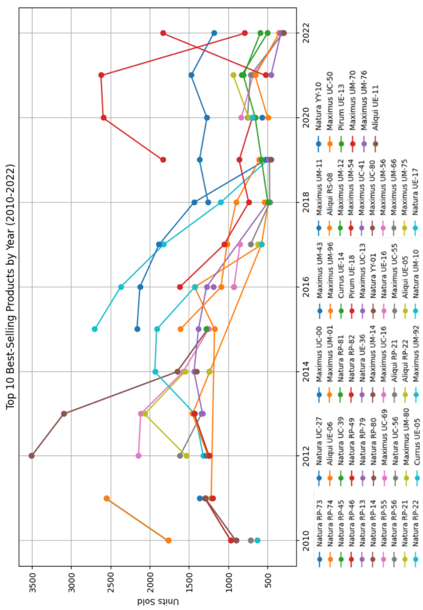

Fig. 5. Sales Trends of Top 10 Products.

Table 4: Results of Multi steps ahead Revenue forecasting.

Model	Steps	Dataset	R	MAPE	RMSE/σ	NSE	IA
Singe-Based	1	Train	0.9472	0.1170	0.3205	0.8973	0.9721
		Test	0.9269	0.2182	0.3753	0.8591	0.9576
	3	Train	0.8806	0.1765	0.4739	0.7754	0.9313
		Test	0.5966	0.5081	0.8026	0.3559	0.7161
	6	Train	0.8593	0.1894	0.5115	0.7383	0.9187
		Test	0.4324	0.6384	0.9017	0.1870	0.6744
	12	Train	0.9235	0.1170	0.3837	0.8528	0.9576
		Test	0.5097	0.9015	0.8604	0.2598	0.7768
HHT-Based	1	Train	0.9891	0.0599	0.1470	0.9784	0.9944
		Test	0.9852	0.0659	0.1714	0.9706	0.9927
	3	Train	0.9685	0.0856	0.2488	0.9381	0.9830
		Test	0.8812	0.2332	0.4728	0.7765	0.9242
	6	Train	0.9654	0.0976	0.2608	0.9320	0.9810
		Test	0.8514	0.3294	0.5245	0.7249	0.8950
	12	Train	0.9834	0.0660	0.1813	0.9671	0.9913
		Test	0.9175	0.2359	0.3978	0.8418	0.9546

Meanwhile, LSTM (self-Attention) is a single-based model with better forcasting performance than other models and is used for comparision with ED-RVFL in forcasting for multi-steps size. It is clarity that when the forecasting step size becomes more demanding, ED-RVFL performs significantly better than other single models. These results align with the analysis from **(Fig. 1.)**, where the trend follows an annual cycle. They also explain why the 12-month forecasting step (R: 0.9175, MAPE: 0.2359) performs better than the 3-month and 6-month steps. From these, it proves the suitability of the HHT-Based approach for both short-term and long-term revenue forecasting.

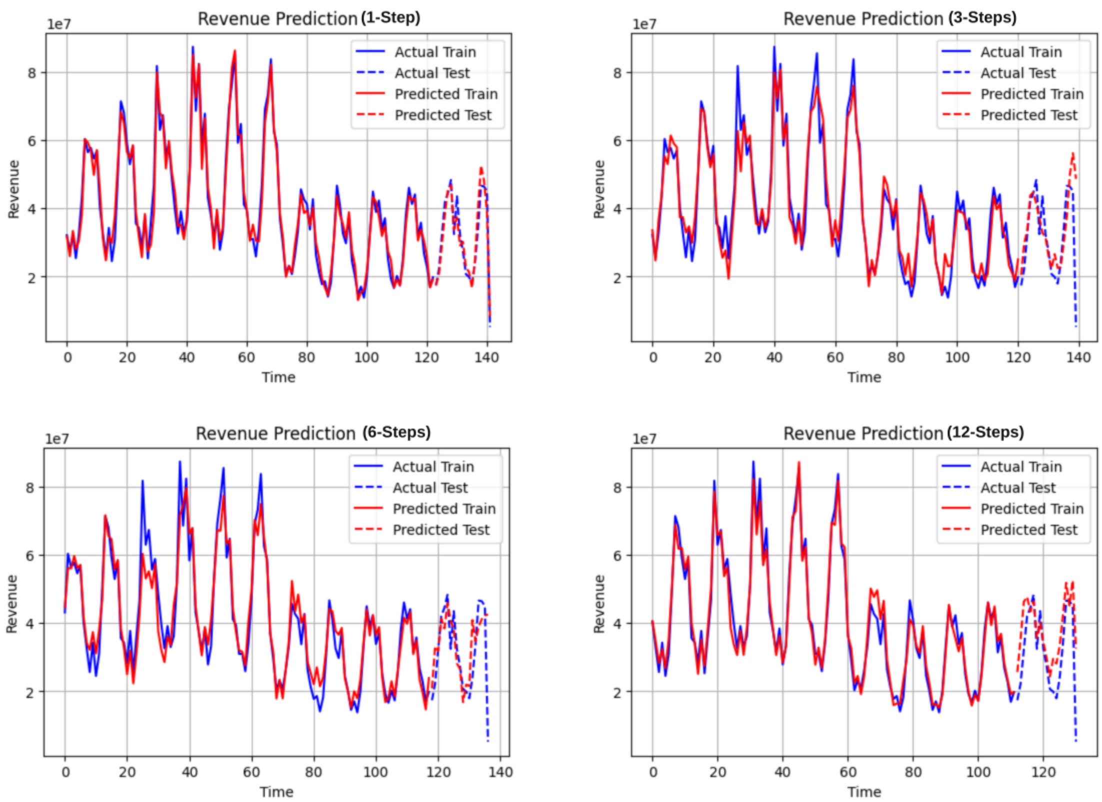

Fig. 4. Observed time series for forecast revenue at multiple steps (ED-RVFL)

.

4. Conclusion

In this report, our team propose a hybrid modeling approach. Signal frequency analysis is performed using HHT to decompose the data into IMFs, followed by feature selection with Lasso ($\alpha = 15000$), and trained with Ensemble Deep-RVFL, which yields excellent results **(Table.3, 4.)**. Specifically, HHT-ED-RVFL achieves the highest accuracy in forecasting revenue for 1, 3, 6, and 12-month steps. The HHT-ED-RVFL model in the testing period records goodness-of-fit metrics (R: 0.9852, MAPE: 0.0659, RMSE/σ: 0.1714, NSE: 0.9706, IA: 0.9927). For a trustworthy model combined with an analysis of the top-selling product trends **(Fig. 5.)**, we believe our research can be extended to support business decision-making.

Nomenclature	
HHT	Hilbert-Huang Transform
IMFs	Intrinsic Mode Functions
RVFL	Random Vectors Functional Link
EMD	Empirical Mode Decomposition
LSTM	Long-Short term memory
ARIMA	Autoregressive integrated moving average
SARIMA	Seasonal Autoregressive Integrated Moving Average
MAPE	Mean Absolute Percentage Error
RMSE	Root Mean Square Error
NSE	Nash-Sutcliffe Efficiency
IA	Index of Agreement
RF	Random Forest
ED	Esemble Deep

Data availability

The dataset is provided from the Dataflow2025 competition, HAMIC, HUS.

References

[1] Muthukrishnan, R., & Rohini, R. (2016, October). LASSO: A feature selection technique in predictive modeling for machine learning. In 2016 IEEE international conference on advances in computer applications (ICACA) (pp. 18-20). Ieee. ieeexplore.ieee.org.7887916.

[2] Zhang, Y., & Suganthan, P. N. (2016). A survey of randomization-based deep and ensemble learning algorithms. Neurocomputing, 275, 278–287.

[3] Toại, T. K., Hạnh, V. T. X., & Huân, V. M. (2023). Áp dụng hồi quy Ridge và mạng nơron nhân tạo để dự báo giá ICO sau sáu tháng. TẠP CHÍ KHOA HỌC ĐẠI HỌC MỞ THÀNH PHỐ HỒ CHÍ MINH-KINH TẾ VÀ QUẢN TRỊ KINH DOANH, 18(4), 131-144.

[4] Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Shih, H. H., Zheng, Q., ... & Liu, H. H. (1998). The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proceedings of the Royal Society of London. Series A: mathematical, physical and engineering sciences, 454(1971), 903-995. semanticscholar.org/paper.

[5] Shi, Q., Katuwal, R., Suganthan, P. N., & Tanveer, M. (2021). Random vector functional link neural network based ensemble deep learning. Pattern Recognition, 117, 107978. sciencedirect.article.S0031320321001655.

[6] Malik, A. K., Gao, R., Ganaie, M. A., Tanveer, M., & Suganthan, P. N. (2023). Random vector functional link network: recent developments, applications, and future directions. Applied Soft Computing, 143, 110377. arxiV.2203.11316.

[7] Huang, N. (2005). Application of the Hilbert-Huang Transform to Financial Data (No. GSC-14807-1). ntrs.nasa.gov.20110014861.

[8] Huang, N. E., Wu, M. L., Qu, W., Long, S. R., & Shen, S. S. (2003). Applications of Hilbert–Huang transform to non-stationary financial time series analysis. Applied stochastic models in business and industry, 19(3), 245-268. shen.sdsu.edu.huan/asmbi/2003.