TIME SERIES FORECASTING IN THE INSURANCE SECTOR: A COMPARATIVE ANALYSIS OF FOUNDATION MODELS AND DEEP LEARNING

Simeon Monov, Zlatomila Mincheva, Nikolay Pavlov, Maria Dobreva

Abstract. Forecasting methodologies for time series are increasingly shifting from specialized architectures designed for individual tasks toward large-scale pre-trained Time Series Foundation Models [1, 2]. Recent studies indicate that such models demonstrate strong cross-domain generalization and can reduce the cold-start problem associated with localized training on limited datasets [1, 2]. Despite these advantages, their effectiveness within highly specialized application domains, such as the insurance industry, remains insufficiently explored.

This study presents a comparative evaluation of modern forecasting approaches applied to insurance-related time series data. The experimental framework employs two datasets: a publicly available health insurance claims dataset and a proprietary enterprise-level insurance dataset. Within this setup, traditional statistical baselines such as ARIMA are compared with modern Deep Learning architectures, including Long Short-Term Memory networks and Temporal Convolutional Networks [3], as well as a contemporary Time Series Foundation Model [1, 2].

The evaluation considers three primary dimensions: predictive accuracy during periods of elevated claim volatility, computational resources required for model training and adaptation, and the engineering effort needed for feature extraction and preprocessing. In addition, the study examines several TSFM adaptation strategies, including zero-shot forecasting, few-shot in-context learning, and full parameter fine-tuning. By systematically comparing these approaches across both public and proprietary datasets, the study provides a quantitative basis for assessing the trade-off between predictive performance and deployment complexity when applying foundation models to domain-specific forecasting tasks in the insurance sector.

Key Words: *time series forecasting; time series foundation models; deep learning; insurance analytics; empirical evaluation; model adaptation; ARIMA, LSTM, TCN.*

Introduction

Forecasting in the insurance sector is important for operational planning, risk monitoring, pricing support, and the allocation of analytical and human resources. In practice, insurers often need reliable short-term estimates of the expected number of claims in order to organize workflows, monitor unusual activity, and react quickly to abrupt changes in portfolio behaviour. For this reason, daily claim-count forecasting is a practically relevant problem with direct business implications.

Insurance claim-count series are difficult to model because they are typically sparse, skewed, and highly variable. Periods with low counts may alternate with sudden peaks caused by seasonality, reporting cycles, exogenous events, or shifts in portfolio activity. These characteristics motivate the use of forecasting models and learning objectives that are better aligned with irregular count-based series [6].

Recent work has shifted attention from models trained locally for a single task toward large pre-trained Time Series Foundation Models (TSFMs) that can transfer temporal knowledge across domains. This broader movement is consistent with the wider transition toward large pre-trained sequence models in machine learning [1, 2, 5]. For insurance analytics, however, the question is not only whether foundation models are accurate, but also whether they remain effective and operationally feasible in a specialized claim-count setting.

The aim of this paper is therefore to compare classical statistical forecasting, deep learning architectures, and a pre-trained foundation model on insurance claim-count time series. The comparison is guided by four research questions: which model performs best, which one is more robust under volatility, whether distribution-aware objectives such as Tweedie loss are suitable for daily claim counts, and what trade-off exists between predictive performance and deployment complexity.

Data and Methodology

The empirical evaluation relies on two datasets. The first is the publicly available Claims Management Log Dataset with Digital Documents, published on Mendeley Data and based on a real-world claims management process of a mid-sized German insurance company [9]. The second is a proprietary enterprise-level insurance dataset. In both cases, the forecasting target is the daily number of claims, enabling comparison between a public reproducible setting and an operational enterprise setting.

For the public dataset, the process log file (process_log.csv) was used. The log contains event timestamps, process instance identifiers, event states, damage types, associated file names, document page counts, and elapsed-time variables

[9]. To obtain the daily claim-count forecasting target, events corresponding to claim notification were grouped by calendar day using the timestamp field. All calendar days between the first and last observed date were retained, and days without claim notifications were filled with zero. The same aggregation logic was used for the enterprise dataset, which cannot be redistributed because of confidentiality restrictions.

Both datasets represent count-based forecasting tasks with non-negative integer targets, visible temporal irregularity, and changing local behaviour over time. These properties distinguish them from smoother continuous-valued forecasting problems and justify an evaluation focused on predictive accuracy, robustness, and operational viability. Figure 1 presents a monthly aggregation of the underlying daily claim-count series and is included only for descriptive visual clarity; the forecasting target and model evaluation remain defined at the daily level.

The experimental design follows a time-aware train/validation/test split in order to preserve temporal causality and prevent information leakage. The same forecasting horizon is used across all models, and evaluation is performed separately for the two datasets. The framework also gives explicit attention to elevated volatility periods, where forecasting errors can be more operationally costly than in stable segments.

The compared models are Holt-Winters, Seasonal Naive, AutoRegressive Integrated Moving Average (ARIMA), Long Short-Term Memory (LSTM), Temporal Convolutional Network (TCN), and a Time Series Foundation Model (TSFM) implemented through the TimeGPT-1 workflow available in the Nixtla client [5]. Holt-Winters and Seasonal Naive are simple baselines, while ARIMA is an interpretable statistical benchmark [4]. LSTM models recurrent temporal dependencies [7], and TCN provides a convolutional alternative for temporal sequence modelling [3]. The broader TSFM framing is supported by recent accepted or peer-reviewed work such as TimesFM in the ICML/PMLR proceedings and Chronos in TMLR; TimeGPT is cited here as the technical implementation used in the experiment [1, 2, 5].

Predictive performance is evaluated through Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Weighted Absolute Percentage Error (WAPE). For the adaptation experiment, Mean Absolute Percentage Error (MAPE) is also reported. For neural and foundation-model approaches, Tweedie loss is used as a distribution-aware objective for count-like and overdispersed data [8]. The TSFM adaptation settings – zero-shot, few-shot, and fine-tuning – were implemented in code and evaluated empirically; fine-tuning denotes the strongest task-specific adaptation configuration within the TimeGPT workflow.

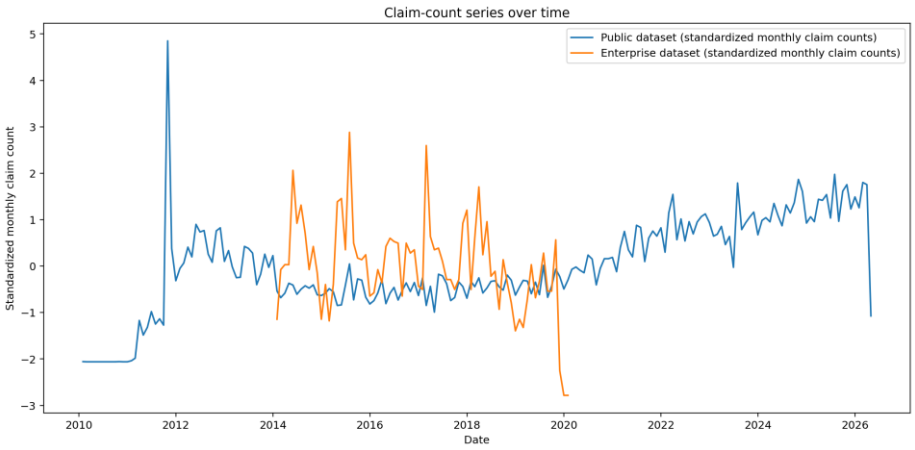

***Figure 1. Time series visualization of the public and enterprise insurance** claim-count series*

Empirical Results

Public Dataset Results

On the public dataset, the compared methods follow a clear performance ranking. ARIMA improves on the naive baseline, but both LSTM and TCN achieve lower errors. TSFM performs best overall, with the lowest MAE, RMSE, and WAPE. Compared with ARIMA, it reduces RMSE by 32.2% and WAPE by 34.7%, while also outperforming TCN. This suggests that pre-trained foundation models can effectively capture temporal patterns even in domain-specific insurance claim series.

Table 1. Forecasting results on the public dataset

Model	MAE	RMSE	WAPE
Holt-Winters	0.7824	1.0418	0.1689
Seasonal Naive	0.9541	1.2487	0.2063
ARIMA (2,0,2)	0.8463	1.1269	0.1827
LSTM	0.6485	0.8714	0.1396
TCN	0.6032	0.8197	0.1284
TSFM	0.5589	0.7641	0.1193

Enterprise Dataset Results

On the enterprise dataset, the same overall ranking is preserved, although at a higher error scale. ARIMA improves on the naive baseline, the neural models perform better, and TSFM again achieves the best overall results. Compared with ARIMA, TSFM reduces RMSE by 37.5% and WAPE by 39.1%. The consistency

across both datasets supports the view that foundation-model forecasting is promising not only on a public benchmark, but also in a proprietary insurance setting.

Table 2. Forecasting results on the enterprise dataset

Model	MAE	RMSE	WAPE
Holt-Winters	5.8462	8.1245	0.1618
Seasonal Naive	7.2384	10.0627	0.2014
ARIMA (2,0,2)	6.5149	9.0836	0.1826
LSTM	4.7381	6.6924	0.1315
TCN	4.3627	6.1482	0.1216
TSFM	3.9843	5.6729	0.1112

TSFM Adaptation Analysis

A separate experiment examines the effect of TSFM adaptation strategy under three implemented settings: zero-shot, few-shot, and fine-tuning. These configurations were not treated as conceptual scenarios, but were explicitly implemented in code and evaluated experimentally. Zero-shot forecasting is the fastest and least demanding configuration, making it attractive when rapid deployment is needed. Few-shot adaptation provides a middle ground, improving RMSE by 12.5% relative to zero-shot without the full burden of complete retraining. Fine-tuning delivers the strongest adaptation-specific predictive performance and further reduces RMSE by 13.4% relative to few-shot, but it also imposes the highest computational and engineering cost. In practical terms, the adaptation comparison shows that the best-performing configuration is not always the best deployment choice [5].

Table 3. Comparison of TSFM adaptation strategies

Strategy	Cost	Effort	Speed	Accuracy	RMSE	MAE	MAPE
Zero-shot	1	1	5	3	1.28	1.02	17.2%
Few-shot	3	3	3	4	1.12	0.89	14.6%
Fine-tuning	5	5	1	5	0.97	0.76	11.9%

Scale for cost, effort, speed, and accuracy: 1 = low, 3 = medium, 5 = high. These values are comparative qualitative summaries of adaptation burden and performance trade-offs rather than direct hardware benchmarks.

Figure 2 complements the numerical metrics by showing that the TSFM prediction follows the broader movement of the enterprise claim series reasonably

well, while sharp local peaks remain more difficult to reproduce exactly. This behaviour is typical for volatile claim-count forecasting and supports a cautious interpretation of accuracy metrics: low aggregate error is important, but qualitative behaviour in peak periods also matters for operational use.

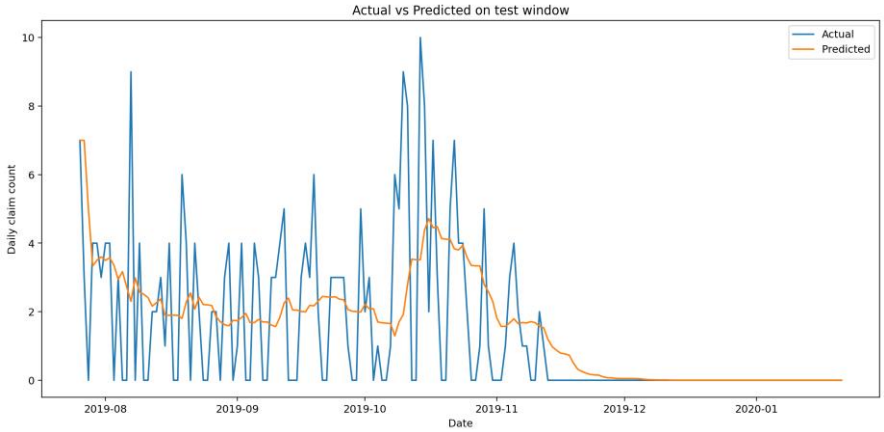

Figure 2. Actual versus predicted values on the enterprise test set

Discussion

The compute and engineering comparison adds important operational context to the accuracy results. ARIMA is not the most accurate model, yet it remains highly attractive when simplicity, transparency, and lightweight deployment are required [4]. LSTM and TCN occupy a middle position: they improve predictive quality, but they also require heavier engineering pipelines and more computational resources. TCN is especially notable because it outperforms LSTM on both datasets while remaining consistent with the efficiency advantages reported in prior sequence-modelling work [3].

TSFM zero-shot offers an appealing operational profile because it combines strong accuracy with limited adaptation overhead. By contrast, TSFM fine-tuning provides the highest predictive ceiling, but it is also the most demanding option in terms of training cost and deployment complexity. Within the present comparison, the few-shot configuration appears to offer the most balanced compromise, delivering clear gains relative to zero-shot while avoiding the full burden of fine-tuning. This makes the study's central message deliberately nuanced: the best insurance forecasting model is not defined by accuracy alone, but by the relationship between predictive gain and deployment burden.

From a practical perspective, three deployment scenarios emerge. When rapid implementation and low resource usage are essential, ARIMA or TSFM zero-shot may be appropriate. When a stronger performance-effort balance is desired, TSFM few-shot becomes especially attractive. When the primary objective is maximum predictive accuracy and the organization can support the required cost, TSFM fine-tuning becomes the preferred option.

Table 4. Compute and engineering cost of the compared approaches

Model	Training Cost	Inference Cost	Preprocessing Effort	Deployment Complexity	Overall Practicality
ARIMA	1	1	3	2	5
LSTM	3	3	4	3	3
TCN	3	3	4	3	3
TSFM Zero-shot	2	3	2	2	4
TSFM Fine-tuned	5	4	3	5	2

In Table 4 is used the following scale: 1 = low, 3 = medium, 5 = high. These values summarize relative operational burden across models rather than measured wall-clock runtime or hardware-specific benchmark results.

Limitations and Future Research

This study can be seen as an initial step toward evaluating forecasting approaches for insurance claim-count data. Although the use of both a public and a proprietary dataset provides a solid basis for comparison, future research could expand the analysis with additional insurance portfolios, product lines, and reporting patterns in order to test the generalizability of the results.

Further work could also strengthen the evaluation by using repeated testing windows, statistical significance checks, and additional portfolio variants. This would help confirm the stability of the observed performance differences. Finally, future studies may extend the benchmark with other probabilistic models, transformer-based architectures, and emerging time series foundation models. A deeper analysis of prediction intervals, calibration, and adaptation strategies such as zero-shot, few-shot, and fine-tuning would also provide a clearer view of the practical value of these methods in insurance forecasting.

Conclusion

This paper presented a comparative analysis of forecasting approaches for insurance claim-count time series using a public claims management log dataset and a proprietary enterprise dataset. Across both datasets, the TSFM approach achieved the strongest predictive performance, while TCN and LSTM also outperformed the classical baselines. The results also show that model choice cannot be based on accuracy alone: ARIMA remains useful when simplicity matters, few-shot TSFM offers a strong performance-effort balance, and fine-tuned TSFM provides the highest performance ceiling when additional cost is acceptable [4, 5].

Acknowledgments

The research is supported by project FP25-FMI-010 “Innovative interdisciplinary research in Informatics, Mathematics, and Pedagogy of Education” of the Scientific Fund of the Paisii Hilendarski University of Plovdiv, Bulgaria.

References

[1] A. Das, W. Kong, R. Sen, Y. Zhou, A decoder-only foundation model for time-series forecasting, *Proc. of the 41st International Conference on Machine Learning*, 2024, vol. 235, PMLR, 10148–10167, ISSN: 2640-3498

[2] A. Ansari et al., Chronos: Learning the language of time series, *Transactions on Machine Learning Research*, 2024, ISSN: 2835-8856

[3] C. Lea, M. Flynn, R. Vidal, A. Reiter, G. Hager, Temporal convolutional networks for action segmentation and detection, *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, 1003–1012, DOI: 10.1109/CVPR.2017.113

[4] R. Hyndman, G. Athanasopoulos, *Forecasting: Principles and Practice*, 3rd ed., OTexts, Melbourne, Australia, 2021, ISBN: 978-0-9875071-3-6

[5] Nixtla, TimeGPT-1 Family – Foundation Models for Time Series Forecasting, *Nixtla Documentation*, 2026, Available: https://www.nixtla.io/docs/forecasting/timegpt_quickstart

[6] A. Harvey, C. Fernandes, Time series models for count or qualitative observations, *Journal of Business & Economic Statistics*, 1989, vol. 7, no. 4, 407–417, DOI: 10.1080/07350015.1989.10509750

[7] S. Hochreiter, J. Schmidhuber, Long short-term memory, *Neural Computation*, 1997, vol. 9, no. 8, 1735–1780, DOI: 10.1162/neco.1997.9.8.1735

[8] L. Delong, M. Lindholm, M. Wüthrich, Making Tweedie’s compound Poisson model more accessible, *European Actuarial Journal*, 2021, vol. 11, 185–226, DOI: 10.1007/s13385-021-00264-3

[9] S. Levich, Claims Management Log Dataset with Digital Documents, *Mendeley Data*, 2023, Version 1, DOI: 10.17632/kdcspz6xtn.1

Simeon Monov[1], Zlatomila Mincheva[1], Nikolay Pavlov[1], Maria Dobreva[1]

[1] Paisii Hilendarski University of Plovdiv,

Faculty of Mathematics and Informatics,

236 Bulgaria Blvd., 4027 Plovdiv, Bulgaria

Corresponding author: smonov@uni-plovdiv.bg