A Comparative Study of Loss Functions for Short-Range Crime Forecasting: Enhancing Embedding Generation for Visualization Systems

Aline Martins Nascimento Belchior*, Raissa Rosa dos Santos Januario†,
Marvin Mendes Cabral‡ and Luis Gustavo Nonato§
Institute of Mathematics and Computer Sciences — University of São Paulo (USP)
P.O. Box 668 – 13.566-590 — São Carlos, SP – Brazil
* Email: aline.belchior@usp.br
† Email: raissa.rosa@usp.br
‡ Email: marvinmarvex@usp.br
§ Email: gnonato@icmc.usp.br

***Abstract*—Criminal incidents have complex socio-economic impacts and reduce the population's perceived security. Accurate short-term crime-rate forecasting enables more efficient allocation of policing resources and better strategic decisions to mitigate criminal activity. Transformer-based architectures are effective at capturing complex temporal dependencies in time series forecasting; however, model behaviour and the structure of learned representations are strongly influenced by the training loss. In this work, we present a comparative analysis of Transformer and recurrent architectures trained with two different losses, the Mean Squared Error (MSE) and the Soft Dynamic Time Warping (Soft-DTW), with focus on short-term forecasting of vehicle thefts in São Paulo. Our goal is to produce embeddings that are more representative of crime dynamics and thus more useful for downstream tasks such as visualization and pattern analysis. We evaluate Autoformer, ITransformer, and Informer alongside LSTM and GRU baselines, using three performance metrics: MSE, MAE and DTW. Overall, models trained with the DTW-based loss achieved performance similar to, or slightly worse than, those trained with MSE; an important exception is the Autoformer, which showed improved accuracy with Soft-DTW at the 14-day horizon. We discuss several factors that likely affected these results: (i) the short forecasting horizons studied, (ii) the formulation of the prediction task (forecasting the entire aggregated series may not be optimal), and (iii) aggregation to daily city-level counts, which discards spatial heterogeneity and may remove salient signal. These findings motivate further experiments (e.g., multi-scale and spatially resolved forecasting) to more comprehensively assess the comparative effectiveness of Soft-DTW and MSE for criminal time series prediction.**

I. Introduction

Crime is a pervasive challenge that affects societies worldwide. Elevated crime rates undermine urban functionality, harm local economies, and reduce population well-being [1]. They erode citizens' sense of security and diminish trust in institutions for maintaining public order. Consequently, accurate forecasting of crime incidence at regional scales enables strategic allocation of resources, supports preventive interventions, and helps mitigate future occurrences [2].

Recent advances in artificial intelligence have produced models capable of extracting complex patterns, identifying trends, and forecasting crimes. However, crime dynamics are intrinsically complex, featuring temporal dependencies, spatial heterogeneity, and seasonal effects, requiring robust modeling approaches for accurate prediction [2]. Transformer-based methods have emerged as a leading class of approaches: originally developed for natural language processing, the Transformer architecture has attracted widespread attention for its strong performance in sequence modeling [3]. The attention mechanism at the core of Transformers captures long-range, context-dependent relationships in sequential data, making these models well suited to time series forecasting and often improving predictive accuracy [4], [5].

Model performance depends critically on the choice of training loss, as the loss defines how prediction errors are measured and minimized. In time series forecasting, errors are typically evaluated via sequence-similarity measures. The Euclidean distance, implicitly assumed by the MSE is limited when series exhibit temporal shifts or local distortions [6]. DTW addresses this limitation by finding an optimal alignment between two sequences, allowing matches despite temporal stretching or compression, and is therefore a robust similarity metric for many forecasting tasks [6], [7]. Motivated by this property, a differentiable DTW-based loss, Soft-DTW, was proposed [6], it replaces the non-differentiable minimum operator in DTW with a soft-minimum that weights possible alignments; this makes the loss value and its gradient computable and efficient for gradient-based optimization [8], [9].

This study performs a comparative analysis of Transformer-based models trained with MSE loss and Soft-DTW loss for short-term forecasting of vehicle theft time series in São Paulo (SP), Brazil. Our objective is to learn embeddings that faithfully represent crime dynamics and can be used in visualization tools for effective analysis of criminal patterns.

II. Related Work

In the domain of time series forecasting for crimes, there are several notable studies. The work [10] provides an analysis

of crime prediction in Los Angeles and Chicago. They examine predictive performance across eight Machine Learning (ML) algorithms, including Random Forest (RF), Multilayer Perceptron (MLP), and Logistic Regression; they also evaluate specialized time series models such as LSTM and ARIMA. LSTM performance was reported using Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). The number of epochs and batch size hyperparameters were configured similarly across cities: both used 40 epochs, with a batch size of 33 for Chicago (CHI) and 31 for Los Angeles (LA). The authors observed that fewer epochs sufficed for loss stabilization (13 for Chicago's loss). They as well identified trends in average crime density, notably a decline in LA toward the end of the dataset. ARIMA was fitted for projections through 2024 across major crime types and hotspots after seasonality was assessed via the Dickey–Fuller test. For CHI, ARIMA produced an RMSE of 31.8, indicating substantial variability but a stable downward trend; for LA, ARIMA achieved an RMSE of 24.65 and forecasted a pronounced decline in crime. Overall, ARIMA outperformed LSTM on the evaluated metrics.

A 6 stage methodology was proposed [11], including the development of predictive models for crimes with traditional ML algorithms and deep learning time series methods such as LSTM and Bidirectional LSTM (BLSTM). They perform forecasting at a five year horizon using statistical techniques. The dataset covers 3 U.S. cities: San Francisco, Chicago, and Philadelphia. The LSTM and BLSTM were trained for 256 epochs and used MSE as the loss function. For monthly and weekly aggregated series, BLSTM performed slightly better (e.g., for Chicago's monthly series LSTM: RMSE = 30.7510 vs. BLSTM: RMSE = 28.9522). By contrast, for daily series LSTM outperformed BLSTM (daily RMSE: 50.2144 vs. 62.3348). The authors additionally evaluated trend forecasting with statistical models, they are Seasonal ARIMA (SARIMA), Prophet, and Holt–Winters Exponential Smoothing (HWES), to project trends over the next five years. SARIMA achieved the lowest RMSE for San Francisco (RMSE = 11.2357), while Chicago produced the highest SARIMA error (RMSE = 38.73).

Complementing these studies, a methodology for analysis and forecasting of criminal time series using spatio-temporal data was proposed [12]. For prediction, the authors employed a dataset of approximately 1.2 million crime events recorded in Poland between 2013–2016. The dataset contained daily information on event counts, location, category, date, and time. For forecasting, they applied the Prophet model and evaluated using the Mean Absolute Percentage Error (MAPE) at a daily resolution. Results varied by crime type: MAPE was around 10% for traffic violations, approximately 20% for theft and robbery, and about 40% for hooliganism.

In the Brazilian context, a study [13] developed predictive models to identify trends in spatio-temporal crime data for the city of São Paulo. They evaluated RF and Support Vector Regression (SVR), as well as an ARIMA model, on data covering 2022–2023, focusing on the top 100 and top 1,000 most violent areas. ARIMA fitted using the Box–Jenkins methodology achieved the best performance for weekly-resolution forecasts. For the top 100 areas, ARIMA and SVR produced similar and satisfactory MSE values; for the top 1,000 areas, all models showed comparable MSE performance, the authors attributed this to the larger data volume. In a related Brazilian study [14] was forecasted a violent-lethality indicator using data from the state of Rio de Janeiro (2021–2023). Before modeling, they applied several diagnostic tests, including normality tests (Shapiro–Wilk, Anderson–Darling), stationarity tests (Dickey–Fuller, Phillips–Perron), and seasonality tests (Kruskal–Wallis, Friedman). Their implementations evaluated ARIMA, ETS (Exponential Smoothing State Space), and AR-NN (Autoregressive Neural Networks). According to MAE and MAPE, ARIMA performed best (MAE = 42.67; MAPE = 10.84%) and ETS had the lowest RMSE (RMSE = 55.05).

The reviewed studies demonstrate the diversity of approaches applied to crime forecasting. The range of models, parameters, and evaluation metrics encourages further experimentation, aiming to enhance predictive performance. Such advances can significantly contribute to more effective and evidence-based public policies for crime prevention.

For time series forecasting, prior research [15] proposed a fine-scale deep forecasting architecture that enhances Transformer-style decomposition by separately interpolating the trend-cyclical component to generate a smoothed, high-resolution representation of short-term oscillations prior to prediction. This enables the model to capture irregular cyclical patterns often overlooked. Furthermore, the authors replaced conventional MSE training with a Dynamic Time Warping–based loss, leveraging the differentiable Soft-DTW formulation to align predicted and true sequences by shape rather than point-wise correspondence. This yielded forecasts that more faithfully reproduced underlying fluctuations under high uncertainty. Results on ETT and weather benchmarks demonstrated that their interpolation-augmented, DTW-trained model consistently outperformed MSE-only variants and reduced the gap with state-of-the-art Transformer models, particularly on short-horizon tasks.

The DILATE model [16] introduces a differentiable loss function for multi-step time series forecasting that explicitly combines a Soft-DTW–based shape dissimilarity term with a relaxed Temporal Distortion Index (TDI) component. This enables deep models to generate predictions with precise temporal alignment, an advance over traditional MSE-based training, which often fails to capture abrupt non-stationary changes. By formulating the Soft-DTW term as a smooth approximation of the classical DTW cost and decoupling shape and temporal penalties through a tunable trade-off parameter, DILATE bridges the gap between training and evaluation criteria. For probabilistic forecasting the authors proposed STRIPE++, a framework that leverages determinant point processes with shape and time sensitive kernels to enforce diversity and accuracy in trajectory ensembles, marking a significant step toward structured and interpretable uncertainty quantification in deep time series models.

A systematic comparative study of loss functions for short-term residential load forecasting [17], demonstrated that replacing conventional MSE/MAE with Soft-DTW improves the model's ability to capture and predict peak loads. This effect was quantified through their newly proposed confusion-matrix framework and bespoke peak position and peak load error metrics. They implemented a residual LSTM architecture trained under Soft-DTW and hybrid losses, showing that pure Soft-DTW produced the greatest improvement in true-positive peak forecasts, despite a marginal increase in false positives, while combined losses offered limited additional gains. By introducing peak-centric evaluation measures grounded in optimal warping path alignments, this work validates the alignment-aware advantages of Soft-DTW over pointwise losses and establishes a robust benchmark for assessing peak prediction performance in smart-grid applications.

III. Data Visualization Panorama

Over the past decade, advances in internet technologies have enabled the digitization of various types of urban data, including crime incident records, socioeconomic indicators, points of interest, and urban environment characteristics, thereby making them more accessible to researchers [18]. In this context, visualization tools play a central role in large-scale data analysis. Combined with advanced machine learning models developed in recent years, these tools facilitate the exploration and identification of complex patterns in criminal activity.

Furthermore, criminology has increasingly emphasized the importance of spatio-temporal crime analysis, as this approach enables the identification of trends, the understanding of crime dynamics, and the anticipation of potential occurrences [19], [20]. Consequently, the development of visualization tools specifically designed for spatio-temporal crime analysis has attracted growing attention. For instance, CriPAV (Crime Pattern Analysis and Visualization) [19], a tool developed for visualizing spatio-temporal crime patterns at street-level granularity. The system employs street network graphs, where intersections are represented as vertices that serve as the unit of analysis. Among its components, the hotspot2vec deep learning mechanism stands out, as it generates embeddings from crime time series associated with vertices classified as hotspots. These embeddings are projected into a Cartesian space and subsequently clustered to identify groups of vertices with similar temporal patterns. This procedure enables the detection of comparable crime dynamics across geographically distant locations.

Similarly, the Space-Time Urban Explorer [20], a tool for analyzing spatio-temporal data on crime and police patrol activities. Like CriPAV, it employs street network graphs for spatial discretization, where each intersection, represented as a graph vertex, is associated with two main time series: (i) records of nearby crimes over time and (ii) records of patrol activities within the same temporal window. The tool provides a set of visualizations that support the identification of patterns and correlations across large volumes of crime and patrol data.

In this work, we investigate prediction models and, in particular, the use of different loss functions to generate embeddings. Based on the results obtained, we intend, in future research, to develop a visualization tool for street-level spatio-temporal crime analysis. Inspired by the functionalities of CriPAV and the Space-Time Urban Explorer, this tool will distinguish itself by enabling the visualization of higher-quality temporal embeddings derived from the forecasting methods evaluated in this study, while also allowing users to select which model and loss function to explore. The approach will employ spatial discretization through street network graphs, where each vertex will represent an intersection and be associated with the time series of crime incidents recorded in its vicinity. These series will be processed by the selected forecasting model, producing more expressive latent representations. Subsequently, these representations will be clustered to identify groups of vertices with similar temporal patterns. The tool's interface will enable users to click on a vertex within a cluster to visualize its spatial information, thereby facilitating the exploration and interpretation of patterns detected in the latent space.

IV. Methodology

A. Data preprocessing

Data on vehicle theft crimes from 2022 to 2025 were obtained from the public repository of the São Paulo Public Security Secretariat [21]. The dataset was filtered to include only incidents recorded within the city of São Paulo. Additionally, columns containing descriptive information related to the crime and its institutional context, such as 'NOME_DEPARTAMENTO', 'DESCR_TIPO_VEICULO', and 'DESCR_TIPOLOCAL', were removed, reducing the 50 attributes in the dataset to a single attribute representing, for each instance, the date of the crime occurrence, as recorded in the police report. This procedure was carried out as an initial step to organize the dataset into a time series format. Crimes registered on each day were then aggregated, with the daily total stored in the attribute 'quantity'. Thus, the final dataset used by the models contained only two attributes: 'date' and 'quantity'. Finally, the data were normalized using the StandardScaler from the Scikit-Learn library [22]. In a dataset S, for each column $c_i \subset S$, the StandardScaler standardizes the values x of column c_i according to the following expression:

$$z = \frac{x - \mu_i}{\sigma_i}$$

where z is the standardized value, and μ_i and σ_i correspond to the mean and the standard deviation of column c_i, respectively. In other words, it centers the data around the mean and scales it according to the standard deviation.

B. Models for Time Series Forecasting

Given a past series, the time series forecasting problem consists of predicting the most probable future series. In general, these models take an *I*-length series as input and

attempt to predict an O-length series as output. For our comparison, we selected two well-established models in time series forecasting tasks, the Autoformer [23] and the ITransformer [24], along with an earlier model used in their original works for comparison, the Informer [25]. In addition to these, we employed two baseline models: the Long Short-Term Memory (LSTM) network and the Gated Recurrent Unit (GRU). We conclude this subsection with an overview of the two main models and a brief description of the Informer.

The Autoformer is an encoder–decoder forecasting architecture built around two core ideas: (1) **progressive series decomposition** (trend versus seasonal) as an internal model operation, and (2) an **Auto-Correlation** module that replaces pointwise self-attention with delay- or period-based aggregation. The model ingests past series and initializes the decoder with seasonal and trend placeholders derived from recent encoder values, providing the decoder with positions to refine. This combination explicitly reduces non-stationarity (through trend removal) and focuses dependency modeling on the predictable seasonal structure.

The **SeriesDecomp** block splits a sequence X into a smoothed trend component X_t and a residual/seasonal component X_s (e.g., $X_t = \text{AvgPool}(\text{Pad}(X)),\ X_s = X - X_t$). It is applied repeatedly within layers to isolate components for targeted processing. The Auto-Correlation module identifies significant time delays τ_i by computing autocorrelations (using FFT/Wiener–Khinchin for efficiency), selects the top-k delays, and aggregates by time-rolling the values: $\text{AutoCorr}(Q, K, V) \approx \sum_{i=1}^{k} \text{Roll}(V, \tau_i) w_i$. FFT computes autocorrelations for all lags in $O(L \log L)$. The design is extended to a multi-head configuration by computing several parallel Auto-Correlation heads and projecting their concatenation.

An encoder layer repeatedly applies the Auto-Correlation module and a Feed-Forward Network (FFN) block with residual connections, each followed by SeriesDecomp, allowing the encoder to progressively extract and refine seasonal signals (encoder output = seasonal features). The decoder layer employs (a) self Auto-Correlation on decoder states, (b) encoder–decoder Auto-Correlation cross-aggregation to incorporate encoder seasonal information, and (c) FFN blocks, all interleaved with SeriesDecomp. Importantly, the decoder accumulates trend estimates across sublayers, i.e., trend accumulation via learned linear weights, while producing progressively refined seasonal outputs.

Supporting modules include standard position-wise FFNs and linear projections (for multi-head output and final readout). Residual connections and stacking (N encoder / M decoder layers) help stabilize learning. The final forecast combines the decoder's seasonal output and accumulated trend: $\hat{Y} = W_S X_{de}^M + T_{de}^M$. Computationally, Autoformer reduces the $O(L^2)$ cost of attention to approximately $O(L \log L)$ through FFT and top-k delay aggregation, while its decomposition explicitly handles non-stationarity, together enhancing long-horizon forecasting robustness and efficiency.

The ITransformer first *re-tokenizes* the input by treating each variate's entire lookback series as a single token rather than a sequence of time-step tokens. A small embedding MLP then compresses each variate-series token into a fixed-length vector. This embedding aggregates temporal information within each variate, allowing downstream layers to operate over variates (columns) instead of timestamps, thereby making the cross-variate structure explicit from the start.

Next, the core consists of a stack of identical Transformer blocks applied across the variate tokens. Each block performs multi-head self-attention across tokens to capture pairwise and higher-order relationships between variates, employs residual connections with Layer Normalization (applied per token) to stabilize gradients and preserve token-level statistics, and applies a shared feed-forward network per token to introduce nonlinear transformations and extract higher-level features from each variate's embedded history. In practice, this combination renders the attention maps interpretable as an $N \times N$ variate-interaction structure, while the FFN refines per-variate representations.

Finally, a lightweight projection MLP decodes each final token back into its forecasted horizon, allowing forecasting to be performed per variate from its refined token. Architecturally, the model is encoder-only, relies on per-token LayerNorm to preserve temporal structure, and omits explicit positional encodings because temporal order is captured within the embedding and FFN transformations. For scalability and practical training, the original work notes straightforward replacements (efficient attention kernels) and batching strategies (sampling variates) to handle large numbers of variates without altering the logical role of any component.

The Informer is designed for long-sequence time series forecasting. It introduces ProbSparse self-attention, which selects dominant queries to approximate attention with an $O(L \log L)$ computational cost, and an attention-distilling mechanism that progressively shortens intermediate representations, providing both computational efficiency and enhanced long-horizon prediction capability.

V. Results and Discussion

Table I presents, for each combination of model and loss function, the two best results, including the Sequence Prediction Length parameter, the performance metrics MSE, MAE, and DTW, and the average time per epoch in seconds. One notable observation from the table is that LSTM_MSE was the only model whose best result occurred with a prediction length of 14 days. Furthermore, the results indicate that although Autoformer_MSE with a prediction length of 28 days achieved the second-best performance in the MSE and MAE metrics, it recorded the worst DTW result for this model. Similarly, iTransformer_DTW with Seq_Pred len 14_14 obtained the highest MAE value among all Seq_Pred len configurations tested for this model.

In addition to the data presented in Table I, Figure 1 illustrates the metric results obtained by the models in combination with different loss functions, considering each Seq_Pred len

value evaluated. Figure 1a presents the MSE values, while Figure 1b displays the DTW values.

Regarding MSE as a performance metric, Figure 1a shows that Autoformer exhibits the smallest overall disparity between the results of the model trained with MSE and the one trained with DTW. For instance, with a prediction length of 14 days, the absolute difference between the MSE values of Autoformer_DTW and Autoformer_MSE is smaller than or equal to the absolute differences observed for all other models. Within the context of these experiments, this result suggests that replacing the MSE loss function with DTW in the Autoformer model does not significantly affect performance as measured by MSE.

Regarding DTW as a performance metric, whose interpretation, similarly to MSE, indicates that lower values represent better model performance [26], the results shown in Figure 1b indicate that LSTM exhibits the smallest overall disparity between the two tested loss functions. However, the analysis of the MSE values in Figure 1a reveals that, despite this stability in DTW, LSTM demonstrates the highest accumulated disparity between loss functions compared to all other models. In other words, by summing the absolute differences between the results obtained for each loss function at each Seq_Pred len value, LSTM achieved the largest total sum.

TABLE I
TOP TWO METRICS PER MODEL–LOSS WITH SEQUENCE PREDICTION LENGTH

Modelo	Seq_Pred len	Metrics			
		MSE	*MAE*	*DTW*	*avg_t_epoch*
Autoformer_DTW	14_7	0,5277	0,5253	2,4735	2,18
	14_14	0,6745	0,5920	4,9832	2,10
Autoformer_MSE	14_7	0,5276	0,5224	2,6625	1,81
	14_28	0,6760	0,5668	9,9331	1,88
Informer_DTW	14_7	0,4897	0,5141	2,0795	11,57
	14_14	1,3311	0,9046	4,7523	12,27
Informer_MSE	14_7	0,4250	0,4621	2,0916	7,66
	14_14	0,5715	0,5367	4,6462	8,37
iTransformer_DTW	14_7	0,5136	0,5165	1,2847	1,17
	14_14	0,8705	1,2674	1,8829	1,19
iTransformer_MSE	14_7	0,4257	0,4608	1,2725	1,04
	14_14	0,4523	0,4845	1,8318	1,05
LSTM_DTW	14_7	0,4061	0,4746	1,1798	3,03
	14_21	1,3465	0,9105	2,0131	3,39
LSTM_MSE	14_14	0,3215	0,4229	1,5779	2,62
	14_7	0,3422	0,4323	1,1600	1,92
GRU_DTW	14_7	0,5575	0,5581	1,2058	2,93
	14_14	1,2628	0,8723	1,7556	3,10
GRU_MSE	14_7	0,3544	0,4338	1,1613	2,14
	14_14	0,3177	0,4220	1,5803	2,28

Overall, it can be observed that models using Soft-DTW as the loss function achieved similar or slightly lower performance compared to those using MSE, with some exceptions in which DTW yielded superior results, such as in the case of Autoformer with a prediction length of 14 days compared to Autoformer using MSE for the same prediction length. Several factors may explain this behavior. First, the forecasts correspond to a short-range scenario, which may have favored the use of MSE, as the work [17] also reported lower performance for DTW compared to MSE in short-term prediction contexts. However, the same study found that models trained with DTW loss performed better when the forecasting task focused on predicting peaks. Therefore, another factor that may have negatively impacted the results is the type of prediction chosen. Furthermore, the granularity of the crime data may have influenced the results. Criminal dynamics are notoriously complex, affected not only by temporal factors but also by spatial ones, such as variations in the type and quantity of crimes across different regions. Thus, by aggregating data for the entire city of São Paulo solely by the day of the criminal occurrence and disregarding the spatial context, relevant information may have been lost, impacting the performance of models under both loss functions. In this sense, the results reinforce the need for further experiments to more comprehensively evaluate DTW's behavior in comparison to MSE.

VI. Final Considerations

In conclusion, we observed that the models did not respond well to the use of Soft-DTW during training in this short-range prediction scenario for crime data. Although similar studies exist regarding the use of Soft-DTW in training and comparisons of loss functions, to the best of our knowledge, the literature remains scarce for experiments conducted under the specific prediction conditions we adopted. In this study [17], we also note again, where the predictive results for the time series as a whole were similar to ours, with training using Soft-DTW performing worse than training with MSE in most cases, except when the task focused on peak prediction in the series. This suggests that the embedding generation in our experiments may have been influenced by the specific task adopted, which warrants further investigation in future work.

A. Future Work

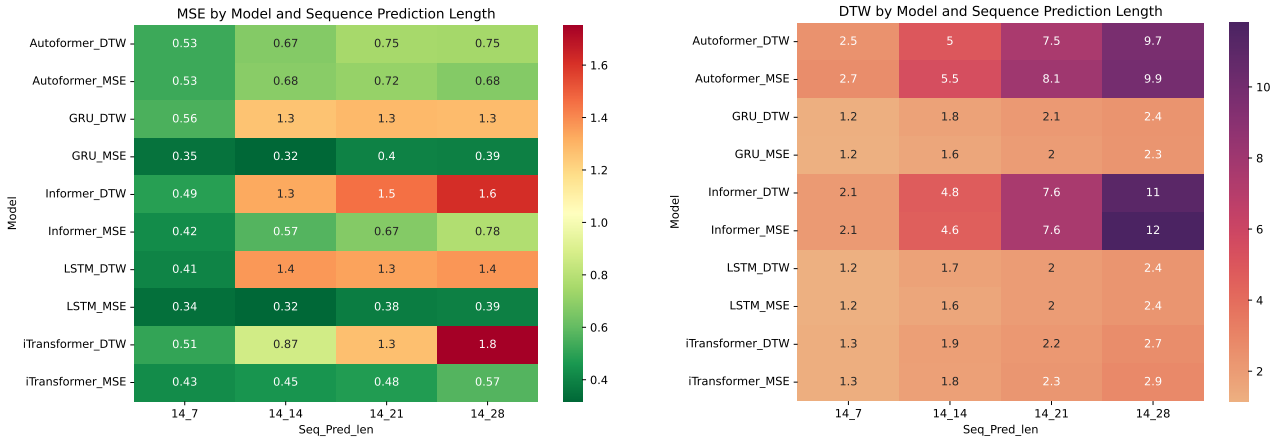
We intend to explore additional experimental configurations, such as the peak prediction approach [17], as well as alternative training and testing regimes. The main future contribution is the development of a visualization tool capable of leveraging the generated embeddings. Variations in training objectives, along with further hyperparameter tuning will be investigated. We plan to incorporate spatial information into the model, since the data used represent an average for the entire city of São Paulo and therefore do not accurately capture the criminal dynamics that the model should be able to observe. Furthermore, the experiments can be extended to models better suited for the configurations we aim to achieve with respect to embedding generation for visualization tools.

Acknowledgment

This work was supported by FAPESP (#2024/07478-8, #2024/14993-6, #2022/09091-8), CNPq (#307184/2021-8), and CAPES (#88887.103193/2025-00). The opinions, hypotheses, conclusions, and recommendations expressed in this material are the responsibility of the authors and do not necessarily reflect the views of FAPESP, CNPq, and CAPES.

References

[1] A. S. R, D. Nidhi, D. S. C, and K. Sowjanya K, "Crime forecasting : A theoretical approach," in *2022 IEEE 7th International Conference on Recent Advances and Innovations in Engineering (ICRAIE)*, vol. 7, 2022, pp. 37–41.



(a) MSE (b) DTW

Fig. 1. Results of MSE and DTW for each model according to each Seq_Pred len.

[2] R. K. Srivastava, A. Gupta, and G. Sharma, "Forecasting crime rate using artificial intelligence applications," in *2023 International Conference on Sustainable Communication Networks and Application (ICSCNA)*, 2023, pp. 1222–1226.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023. [Online]. Available: https://arxiv.org/abs/1706.03762

[4] F. Caffaro, L. Bongiovanni, and C. Rossi, *Geo-temporal Crime Forecasting Using a Deep Learning Attention-Based Model*. Cham: Springer Nature Switzerland, 2025, pp. 323–329. [Online]. Available: https://doi.org/10.1007/978-3-031-62083-6_26

[5] A. T. Nguyen Dai, T. T. Vo Thi, and T. B. Nguyen, "Applying itransformer for saigon river water level forecasting," in *2024 International Conference on Advanced Technologies for Communications (ATC)*, 2024, pp. 556–560.

[6] M. Cuturi and M. Blondel, "Soft-dtw: a differentiable loss function for time-series," in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ser. ICML'17. JMLR.org, 2017, p. 894–903.

[7] S. Salvador and P. Chan, "Toward accurate dynamic time warping in linear time and space," *Intell. Data Anal.*, vol. 11, no. 5, p. 561–580, Oct. 2007.

[8] J. Jiang, S. Lai, L. Jin, and Y. Zhu, "Dsdtw: Local representation learning with deep soft-dtw for dynamic signature verification," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 2198–2212, 2022.

[9] K.-H. Ho, P.-S. Huang, I.-C. Wu, and F.-J. Wang, "Prediction of time series data based on transformer with soft dynamic time wrapping," in *2020 IEEE International Conference on Consumer Electronics - Taiwan (ICCE-Taiwan)*, 2020, pp. 1–2.

[10] W. Safat, S. Asghar, and S. A. Gillani, "Empirical analysis for crime prediction and forecasting using machine learning and deep learning techniques," *IEEE access*, vol. 9, pp. 70 080–70 094, 2021.

[11] E. G. İlgün and M. Dener, "Exploratory data analysis, time series analysis, crime type prediction, and trend forecasting in crime data using machine learning, deep learning, and statistical methods," *Neural Computing and Applications*, vol. 37, no. 18, pp. 11 773–11 798, 2025.

[12] G. Borowik, Z. M. Wawrzyniak, and P. Cichosz, "Time series analysis for crime forecasting," in *2018 26th international conference on systems engineering (ICSEng)*. IEEE, 2018, pp. 1–10.

[13] H. Silva, S. Rocha, and G. Gonçalves, "Técnicas para predição de crimes utilizando dados oficiais considerando tempo e espaço," in *Anais da XII Escola Regional de Computação do Ceará, Maranhão e Piauí*. Porto Alegre, RS, Brasil: SBC, 2024, pp. 229–238. [Online]. Available: https://sol.sbc.org.br/index.php/ercemapi/article/view/30190

[14] J. L. de Jesus Goulart, M. M. Provenza, V. L. Xavier, I. C. de Almeida Lima, P. H. C. Simões, and J. C. Siqueira, "Previsões de séries temporais para os crimes de letalidade violenta no rio de janeiro através dos modelos de estado e suavização exponencial, arima e redes neurais autorregressivas," *Caderno Pedagógico*, vol. 21, no. 10, pp. e8626–e8626, 2024.

[15] Y. Chen, W. Jia, and Q. Wu, "Fine-scale deep learning model for time series forecasting," *Applied Intelligence*, 2024.

[16] V. Le Guen and N. Thome, "Deep time series forecasting with shape and temporal criteria," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 342–355, 2022.

[17] Y. Chen, C. Obrecht, and F. Kuznik, "Enhancing peak prediction in residential load forecasting with soft dynamic time wrapping loss functions," *Integrated Computer-Aided Engineering*, vol. 31, no. 3, pp. 327–340, 2024.

[18] D. B. L. Silva, T. Vieira, E. de Barros Costa, A. Paiva, and L. G. Nonato, "A street corner-level methodology to analyze the influence of points of interest on urban crime," *Socio-Economic Planning Sciences*, p. 102297, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0038012125001466

[19] G. Garcia-Zanabria and L. Nonato, "Visual crime pattern analysis," in *Anais Estendidos da XXXV Conference on Graphics, Patterns and Images*. Porto Alegre, RS, Brasil: SBC, 2022, pp. 55–61. [Online]. Available: https://sol.sbc.org.br/index.php/sibgrapi_estendido/article/view/23261

[20] T. P. Santos, J. M. S. Souza, T. Vieira, and L. G. Nonato, "Space-time urban explorer: A visual tool for exploring spatiotemporal crime and patrolling data," in *2024 37th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, 2024, pp. 1–6.

[21] SSP, "Portal da transparência da ssp-sp," 2025. [Online]. Available: https://www.ssp.sp.gov.br/estatistica/consultas

[22] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[23] H. Wu, J. Xu, J. Wang, and M. Long, "Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting," *Advances in neural information processing systems*, vol. 34, pp. 22 419–22 430, 2021.

[24] Y. Liu, T. Hu, H. Zhang, H. Wu, S. Wang, L. Ma, and M. Long, "itransformer: Inverted transformers are effective for time series forecasting," *arXiv preprint arXiv:2310.06625*, 2023.

[25] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, "Informer: Beyond efficient transformer for long sequence time-series forecasting," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 12, 2021, pp. 11 106–11 115.

[26] M. Müller, *Information retrieval for music and motion*. Springer, 2007.