Watermarking Time Series Diffusion Models

Lucas Fatas Lynas

Supervisor(s): Dr. Lydia Y. Chen, Jeroen M. Galjaard, Chaoyi Zhu

[1]**EEMCS, Delft University of Technology, The Netherlands**

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 23, 2024

Name of the student: Lucas Fatas Lynas
Final project course: CSE3000 Research Project
Thesis committee: Rihan Hai, Dr. Lydia Y. Chen, Jeroen M. Galjaard, Chaoyi Zhu

Abstract

In many scientific fields, time series data is essential, yet maintaining the integrity and legitimacy of such data is still difficult. Traditional watermarking techniques have mainly been used for multimedia. Although approaches for watermarking non-media data have been developed recently, there is still a big gap in the development of reliable and undetectable watermarking methods for time series diffusion models. We suggest a novel modification of the tree ring watermarking method for the 2D time series model LDCast, which is intended for precipitation prediction.

Through the incorporation of watermarks into the model's process, we guarantee resilience and undetectability. Our approach preserves the LDCast model's predicted accuracy while still being able to verifying the origins of the data. We confirm the efficacy of our method through comprehensive evaluation, underscoring its potential to improve the security and integrity of time series forecasting models.

1 Introduction

Digital watermarking is the process of embedding a piece of code or a key in data in order to provide copyright information. This allows for authenticity or ownership to be confirmed. In recent years, a demand for new digital watermarking techniques has risen due to the popularity of generative models. Generative models, including text-to-image models like Midjourney and Diablo, can generate data that can be mistaken for real data. This has resulted in new watermarking techniques being developed to identify real from fake media and guaranteeing traceability.

However, media data types have been the primary target of modern digital watermarking technique. There is currently a lack of suitable solutions for non-media data, such as time series data generated by diffusion models. Time series data, which are collections of data points organized in a chronological order, are necessary for a wide range of applications. This covers financial market analysis, medical monitoring, and weather forecasts.

One of key challenges in watermarking non-media types lies in balancing the invisibility and detectability of the watermark. For media watermarking, the term "invisibility" refers to the ability of the watermark to remain invisible to the human eye. Defining invisibility for non-media data types, like time series data, is more of a challenge and dependent on the purpose of the data [1]. For a watermark to remain invisible, it should not interfere with the downstream data processing and analysis. The data's integrity must be preserved in order to accomplish this.

This research contributes by presenting a watermarking technique for 2D time series diffusion models with the aim of providing a method for verifying ownership and authenticity, while maintaining data integrity.

2 Related Work

This section discusses various works that are related to our research on watermarking diffusion time series models.

2.1 Watermarking Diffusion Models

Recent advancements in diffusion models have highlighted their capability in generating high-quality synthetic data. Watermarking techniques tailored for diffusion models was explored in a study by Duan et al. (2022) [2]. The study proposes a robust and invisibile method of embedding watermarks in images generated by diffusion models. Their method incorporates noise into the diffusion process, which allows for detectable watermarks that do that significantly impact the image quality.

2.2 Tree Ring Watermarking

The work of Zhao et al. (2021) [3] provides a unique watermarking technique called tree ring watermarking, which produces patterns resembling rings that may be included into pictures. This method has proven to be quite resilient to many image alterations, including noise addition, rotation, and compression. Because of its resilience and little effect on the data's visual quality, it is a good fit for adaptation to time series data.

2.3 LDCast

LDCast [4] is a 2D time series model designed for precipitation forecasting which utilizes both spatial and temporal information to predict future rainfall. The papers primary focus was on improving prediction accuracy and computational efficiency. It did not include any mention of watermarking or verifying ownership.

2.4 Non-Media Watermarking

Research by Chen et al. (2020) [1] delves into the invisibility aspect of watermarking non-media data. There are special difficulties when watermarking non-media data, especially time series data. The fundamental challenge is to keep the watermark invisible without interfering with the time series' natural patterns and statistical characteristics. By gently altering the data points inside the time series' statistical boundaries, they suggest an embedding strategy that makes watermarks invisible to conventional data processing methods.

3 Methodology

This section outlines the methodology employed in this research to adapt an existing watermarking technique to a Time Series Diffusion Model.

3.1 Model Selection

To adapt an existing text-to-image diffusion model watermarking technique to a time series diffusion model , we selected LDCast [4], a 2D time series model for precipitation forecasting, as a case study. LDCast is a model that uses spatial and temporal information precipitation forecasting.

The decision to use LDCast was influenced by a series of evaluations of existing time series models. Initial focus was on finding 1D diffusion time series models. A few of the most

common models include TimeGrad [5], ScoreGrad [6] and CSDI [3] which are all autoregressive diffusion models. Autoregressive models present significant challenges due to the lack of existing watermarking techniques for these models. Current diffusion model watermarking techniques are only applied to latent diffusion models.

Since there are no existing latent 1D time series diffusion models, the focus shifted to 2D time series models, leading to the selection of LDCast. LDCast is a model designed for precipitation forecasting that uses both spatial and temporal information.

3.2 Watermarking Technique Selection

Existing watermarking techniques for text-to-image diffusion models include Stable Signature [7], Tree Ring Watermarking [3] and NaiveWM/FixedWM [2]. We focused on the tree ring watermarking method for its robustness and imperceptibility in the context of visual data.

The tree ring watermark is designed for text-to-image diffusion models. It embeds concentric ring-like patterns, also knows as the "key" into the initial noise vector that is then passed through the model to generate the output. These rings are resistant to various forms of manipulation and also do not significantly alter the final output of the model. The detect whether an image has been watermarked, the diffusion process has to be reversed to reconstruct the initial noise vector used to generate the image. This reconstructed noise vector is then compared to the key. If it is similar, this indicates that the rings are present and that the image has been watermarked. This process is depicted in figure 4.

Benefits of Tree Ring Watermark:

- **Robustness:** The watermark is resistant to a wide range of attacks, including compression, rotation, and noise addition.
- **Capacity:** It allows embedding large amounts of data without quality loss.
- **Invisibility:** It does not significantly alter the perceptual quality of the image, ensuring the watermark remains imperceptible to human senses.

To adapt this technique to LDCast, two key algorithms needed modification: watermark generation and detection.

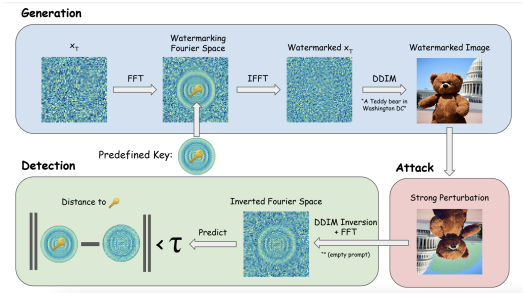

Figure 1: Tree Ring Watermark Process

3.3 Developing Watermark Generation

The primary challenge in adapting the watermark generation algorithm of Tree Ring is the dimensional differences between the data structures. The initial noise vector for the text-to-image models that tree ring watermark is intended for, are typically a 2D structure (Height x Width). The process to generate the watermark for these models involves embedding the rings, known as the "key", into the center of the initial noise vector as seen in Figure 4.

Figure 2: Original Tree Ring Watermark Diagram

The initial noise vector of LDCast on the other hand, operates with a 4D structure (Time x Window x Height x Width). The original 3D structure (Time x Height x Width) of the data is divided into smaller windows (32x32) during the encoding process, adding the fourth dimension. This requires significant modifications to the watermark generation algorithm of tree ring.

To ensure the watermark can be effectively generated and detected, the adapted watermark generation algorithm we have implemented embeds the rings into each window of each timestamp of the initial noise vector. A simplified diagram of this can been seen in Figure 4, where the rings are present in every windows of for all timestamps t_0 to t_n.

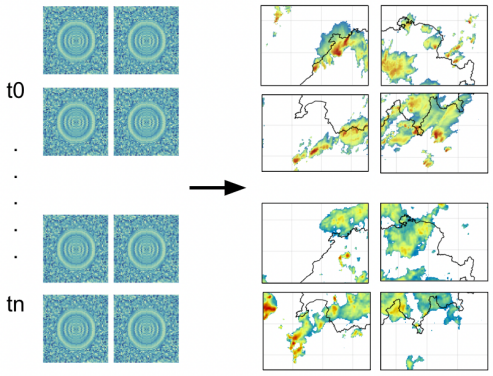

Figure 3: Adapted Tree Ring Watermark Diagram

This method of embedding multiple rings throughout the entire noise vector increases the robustness of the watermark as it will be more resilient to possible attacks. For example, if the rings were only embedded in a select few timestamps, the watermark could be removed by excluding those timestamps from the time series.

3.4 Developing Watermark Detection

The tree ring watermark detection method works by embedding a specific pattern, called the "key," in the initial noise vector used by the diffusion model to generate an image. The resulting image is converted back into its original noise vector in order to identify this watermark. This entails encoding the image into latent space using a the models encoder, and then retrieving the original noise vector through DDIM inversion (Denoising Diffusion Implicit Models).

The Fourier transform of the initial noise vector is then calculated. To then identify whether the image is watermarked the distance between the Fourier transform of the noise vector and the key is measure. When comparing non-watermarked images to watermarked images, the distance to the key is should be shorted. Based on this distance measurement, a distance threshold is established to determine whether or not photos are watermarked.

For our implementation, we integrated the encoder of LDCast into our pipeline and implemented the DDIM inversion process to retrieve the initial noise vector. Below is the LaTeX code demonstrating the mathematical formulation of DDIM inversion:

The inversion from the generated image x_0 to the initial noise vector x_T is given by:

$$x_t = \sqrt{\alpha_t}\hat{x}_0 + \sqrt{1-\alpha_t}\epsilon_\theta(x_t) \tag{1}$$

where α_t is a scheduling parameter and $\epsilon_\theta(x_t)$ is the noise predicted by the model at step t. The estimate of $\hat{x}_0$ is given by:

$$\hat{x}_0 = \frac{x_t - \sqrt{1-\alpha_t}\epsilon_\theta(x_t)}{\sqrt{\alpha_t}} \tag{2}$$

To find x_{t-1}, we use:

$$x_{t-1} = \sqrt{\alpha_{t-1}}\hat{x}_0 + \sqrt{1-\alpha_{t-1}}\epsilon_\theta(x_t) \tag{3}$$

The entire inversion process from x_0 to x_T is then denoted as:

$$x_T = D_\theta^\dagger(x_0) \tag{4}$$

This inversion process retrieves the initial noise vector, which is then used for watermark detection by comparing the distance in the Fourier space with the key.

Then a threshold distance is set to determine whether the time series data is watermarked or not. This threshold is based on observations and needs to ensure a balance between minimizing false positives and maximizing true positive rate. Watermarked data has a shorter distance to the key, while non-watermarked data has a larger distance. This method ensures reliable watermark detection.

4 Experimental Setup and Results

This section outlines the initial experiments conducted to evaluate the implementation of the tree ring watermark technique in the LDCast model. The focus is on generating forecasts, evaluating the invisibility of the watermark, and discussing some preliminary findings.

4.1 Simple Forecast

To begin, the LDCast model was implemented locally to produce precipitation forecasts. The model was run on a dataset of historical precipitation data, generating forecasts for future precipitation. An example of the forecast produced by LDCast is shown in Figure 4.

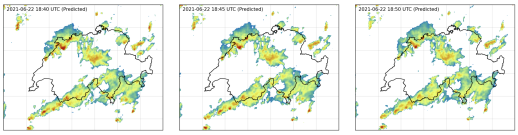

Figure 4: Example of forecast produced by LDCast.

The data used in LDCast consists of time series records of precipitation levels, structured in a 2D format that includes spatial and temporal information. Each data point represents the amount of precipitation at a specific location and time. The model processes this data to predict future precipitation patterns, providing valuable insights for weather forecasting applications.

4.2 Assessing Watermark Invisibility and Detectability

Evaluating the effectiveness of the adapted watermark involves measuring two key aspects: invisibility and detectability.

Invisibility: This refers to the watermark's ability to remain undetectable within the data, ensuring it does not interfere with the primary use of the time series data. When a watermark is invisible, it indicates that it shouldn't affect the model's ability to forecast future data. The forecast accuracy with and without the watermark will be compared in order to determine the invisibility using the Fractional Skill Score (FSS), one of the most popular spatial verification metrics in use today.

Detectability: This is the capacity to recognize with precision when a watermark appears in the data. It is crucial that the watermark can be consistently detected by the detection algorithm without leading to false positives or negatives. The measure we will use to quantify detectability is the distance of the initial noise vector of the data to the key. The distance of the watermarked data should continuously display a closer distance then the non-watermarked data so that a threshold can be set to decide whether data has been watermarked.

4.3 Evaluating Invisibility

To evaluate the invisibility of the watermark, forecasts were generated for data with a smaller watermark, larger watermark and non-watermarked data. Small watermark embedded rings with a radius of 8 (R=8) and the larger watermark embedded rings with a radius of 15 (R=15) were the two forms of watermarks that were taken into consideration. The performance of these forecasts was compared using the Fractional Skill Score (FSS), with an FSS of 1 indicating no difference between the forcasted and observed data, at different forecast intervals: 0, 5, 10, 15, and 20 minutes into the future as shows in Figure 4.

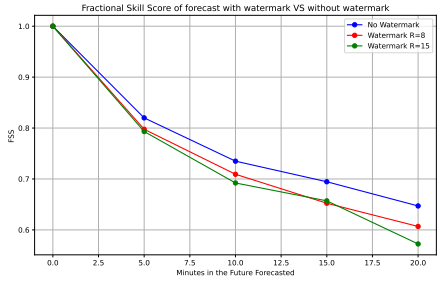

Figure 5: Fractional Skill Score comparison between watermarked and non-watermarked forecasts.

The results indicated that both the forecast with the small and large watermark performed slightly worse than the non-watermarked forecasts. The average FSS of the non-watermarked forecast was 0.779 while the small and large watermarked forecast had a FSS of 0.753 and 0.743 respectively.

4.4 Evaluating Detectability

The distribution of distances of the watermarked and non-watermarked data was analyzed in order to assess the detectability of the watermark contained in the LDCast model. Small rings with a radius of 8 (R=8) and bigger rings with a radius of 15 (R=15) were the two forms of watermarks that were taken into consideration.

The results are displayed in Figure 4, which depict the distribution of distances for the watermarked and non-watermarked data. The distance metric measures how closely the key pattern used for watermarking is to the initial noise vector.

The distribution of distances for watermarked data with R=8 is rather broad, roughly falling between 88 and 93. The distance distribution of the non-watermarked photographs ranges from 95 to 98, resulting in a discernible difference from the watermarked images.

The distances for watermarked data with R=15, on the other hand, are closely clustered between 90.3 and 91.0. The distribution of non-watermarked data is likewise narrow, spanning from 97.5 to 98.5. When employing a greater ring radius, the watermark may be detected more precisely, as evidenced by the tighter clusters for both watermarked and non-watermarked photos.

The analysis shows a stronger distinction between watermarked and non-watermarked images is achieved when employing a wider ring radius (R=15) for watermarking. The probability of overlap is greatly decreased by the close clustering of the distances for both categories, which is essential for accurate detection. The distinct separation ensures that watermarked images consistently exhibit shorter distances to the key, while non-watermarked images show longer distances.

This clear separation is evident in the distribution graphs, where the non-overlapping clusters for R=15 provide a robust basis for setting a threshold to distinguish between watermarked and non-watermarked images effectively. In contrast, the distributions for R=8 show a broader spread, making it more challenging to set an accurate detection threshold without risking false positives or false negatives.

The non-overlapping clusters for R=15 offering a strong foundation for determining a threshold that would effectively distinguish between watermarked and non-watermarked data. The distributions for R=8, on the other hand, have a wider spread, which makes it more difficult to choose a precise detection threshold without running the danger of producing false positives or false negatives.

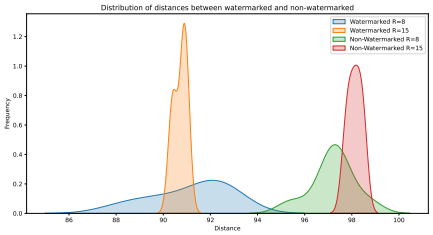

Figure 6: Distances of watermarked and non-watermarked to key in latent space

5 Responsible Research

This research was conducted with an emphasis on ethical considerations. We ensured that the methodologies and outcomes adhere to the principles of integrity and transparency.

5.1 Data Security and Privacy

All data used for this research was non-personal precipitation data. The data naturally addressed privacy and anonymity concerns because it contained no personally identifiable information. Making sure the data was secure and the watermarking method was strong enough to stop unlawful usage and distribution was the main ethical concern.

5.2 Reproducibility and Transparency

Throughout the research process we ensure the reproducibility of our findings. This is essential to maintaining our integrity as scientists. We have made all of the code and datasets used in our research publicly available to aid with replication. Because of this transparency, other researchers are able to duplicate our tests and corroborate our findings.

We go into great depth about our process, including the precise settings and parameters for both watermark detection and embedding. This extensive documentation guarantees that our studies can be precisely replicated and our findings expanded upon by other researchers.

6 Discussion

The objective of this study was to create and assess a watermarking method for 2D time series diffusion models, with a particular emphasis on modifying the tree ring watermarking method for LDCast, a precipitation forecasting model. The experiment assessed the detectability and invisibility of the watermark at different sizes (r=8 and r=15). The impact on

model performance and the watermark detection's resilience were the main evaluation measures.

The analysis revealed an important compromise between detectability and invisibility. It was discovered that smaller watermarks (r=8) were harder to detect as expected because of their smaller effect on the model's results. The inability of the detection process to distinguish between data that was watermarked and data that wasn't increased the risk of false negatives. On the other hand, larger watermarks (r=15) made patterns simpler to see and discern while having no appreciable impact on the model's functionality.

According to the results, watermarks with a bigger radius (r=15) provided a better balance between detectability and invisibility. In particular, watermarks with r=15 were very detectable and still retained their invisibility. This equilibrium was necessary to guarantee that the watermarks remained recognizable while not interfering with the precipitation forecasts' accuracy.

The results of the studies showed that a watermark with r=15 provided the optimal combination of detectability and invisibility. This size was sufficient to provide reliable identification without impairing the forecasting skills of the model. A precise detection threshold was made possible by the clear division of watermarked and non-watermarked data clusters, lowering the possibility of false positives and false negatives.

In conclusion, our study effectively created a watermarking method for 2D time series diffusion models, striking a compromise between detectability and invisibility. It was determined that r=15 was the ideal watermark size, which would maintain the watermark's detectability without degrading the model's functionality. By proving that strong, undetectable watermarks can be successfully included into time series data models, this work advances the field of non-media watermarking and opens the door to improved data authenticity and ownership verification in a variety of applications.

7 Conclusions and Future Work

The main research subject that this thesis investigated was how to create and implement a watermarking method for 2D time series diffusion models with an emphasis on balancing detectability and invisibility.

Our methodology modified the tree ring watermarking method to apply to LDCast, a 2D time series model intended for precipitation prediction. This work's principal contributions are as follows:

- **Development of a Watermarking Technique:** The research effectively employed a watermarking method that incorporates a watermark into the two-dimensional time series data, all while maintaining a minimal impact on the predictive capabilities of the model.
- **Evaluation of Invisibility and Detectability:** The watermark was assessed based on how well it balanced maintaining the integrity of the data and being able to be identified. The outcomes showed that the watermark had a limited effect on the model performance and that our detection algorithm could successfully identify it.

Even though this study has advanced the field of watermarking 2D time series models, there are still a few issues that need more research.

To make sure the watermark is resistant to different kinds of attacks, extensive robustness testing is necessary. This entails examining the effect of dilation, which enlarges the characteristics of the time series data, and assessing how resilient the watermark is to cropping, which involves removing areas of the data. Comprehensive testing with various kinds and intensities of noise will also confirm the robustness of the watermark.

It is important to apply the watermarking technique to other 2D time series models besides LDCast. This will improve the watermark's generalizability and useful applicability by enabling the assessment of its consistency across different models and data kinds.

Practical insights into the watermarking method's performance and usability in real-world settings will come from testing it in real-world scenarios, especially in operational forecasting systems. This will assist in identifying any obstacles or restrictions that might not be seen in carefully monitored experimental environments.

In summary, this thesis has shown that watermarking 2D time series models is both feasible and efficient, especially in the context of LDCast . The results provide a basis for further study and development in the area of digital watermarking, furthering its application in securing and authenticating time series data.

References

[1] Arezou Soltani Panah, Ron Van Schyndel, Timos Sellis, and Elisa Bertino. On the properties of non-media digital watermarking: A review of state of the art techniques, 2016.

[2] Yugeng Liu, Zheng Li, Michael Backes, Yun Shen, and Yang Zhang. Watermarking diffusion model, 2023.

[3] Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. Tree-ring watermarks: Fingerprints for diffusion images that are invisible and robust, 2023.

[4] Jussi Leinonen, Ulrich Hamann, Daniele Nerini, Urs Germann, and Gabriele Franch. Latent diffusion models for generative precipitation nowcasting with accurate uncertainty quantification, 2023.

[5] Kashif Rasul, Calvin Seward, Ingmar Schuster, and Roland Vollgraf. Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting, 2021.

[6] Tijin Yan, Hongwei Zhang, Tong Zhou, Yufeng Zhan, and Yuanqing Xia. Scoregrad: Multivariate probabilistic time series forecasting with continuous energy-based generative models, 2021.

[7] Pierre Fernandez, Guillaume Couairon, Hervé Jégou, Matthijs Douze, and Teddy Furon. The stable signature: Rooting watermarks in latent diffusion models, 2023.