Membership Inference Attacks Against Time-Series Models

Noam Koren NOAM.KOREN1@IBM.COM
Abigail Goldsteen ABIGAILT@IL.IBM.COM
Guy Amit GUY.AMIT@IBM.COM
Ariel Farkash ARIELF@IL.IBM.COM
IBM Research, Haifa, Israel

Editors: Vu Nguyen and Hsuan-Tien Lin

Abstract

Analyzing time-series data that contains personal information, particularly in the medical field, presents serious privacy concerns. Sensitive health data from patients is often used to train machine learning models for diagnostics and ongoing care. Assessing the privacy risk of such models is crucial to making knowledgeable decisions on whether to use a model in production or share it with third parties. Membership Inference Attacks (MIA) are a key method for this kind of evaluation, however time-series prediction models have not been thoroughly studied in this context. We explore existing MIA techniques on time-series models, and introduce new features, focusing on the seasonality and trend components of the data. Seasonality is estimated using a multivariate Fourier transform, and a low-degree polynomial is used to approximate trends. We applied these techniques to various types of time-series models, using datasets from the health domain. Our results demonstrate that these new features enhance the effectiveness of MIAs in identifying membership, improving the understanding of privacy risks in medical data applications.

Keywords: Privacy, Machine Learning, Time-Series, Membership Inference

1. Introduction

There is a clear conflict between the ever-increasing interest in analyzing personal data to enhance and improve processes, and the need to preserve the privacy of data subjects. In the medical domain, sensitive data from patients is often used to train machine learning (ML) models that aid physicians in diagnostics and treatment. ML models are also utilized within medical devices and applications to predict malfunctions and improve ongoing care.

Assessing the privacy risk of such models is crucial to enable making knowledgeable decisions on whether to use a model in production, share it with third parties, or deploy it in patients' homes. Privacy risk assessment is often achieved by running membership inference attacks against the models and measuring their success rate.

Membership inference attacks (MIA) attempt to distinguish between samples that were part of a target model's training data (called members) and samples that were not (non-members), based on the model's outputs. Many such attacks are based on training a binary classifier as an attack model Shokri et al. (2016). These attacks can be applied to various model types, including classification Shokri et al. (2016), regression Truex et al. (2019), graph He et al. (2021), and generative Hayes et al. (2017) models. However, MIA against time-series prediction models, has not yet been properly researched.

This paper addresses this gap by evaluating existing membership inference approaches on time-series forecasting models and introducing new features specifically designed for these models. Our main contribution is the addition of two novel features that exploit the trend and seasonality components of time-series data. The trend is approximated by fitting a low-degree polynomial and the seasonality is estimated using the Discrete Fourier Transform (DFT).

Since time series fundamentally consist of trend and seasonality components, it is reasonable to assume that time-series models are more adept at accurately estimating these elements in series encountered during training. Additionally, various state-of-the-art forecasting models, such as Neural Fourier Transform (NFT) Koren and Radinsky (2024), TimesNet Wu et al. (2023), etc., specifically incorporate those components into their design. Consequently, when targeting a time-series prediction model, there is a significant likelihood that the model will precisely estimate the series' seasonality and trend of its training data, providing a strategic advantage in MIAs. This underscores the importance of considering these features when assessing the vulnerability of time-series models.

We empirically evaluate the impact of adding seasonality and trend as input features to MIA models by testing different combinations of existing and new features. The evaluation is performed on six time-series prediction models, using two medical datasets. The results demonstrate significant improvements across multiple prediction horizons, ranging from 3% to 26%, confirming the efficacy of the proposed attack features. This is an important first step towards proper privacy assessment methods for time-series models, which have so far been mostly overlooked.

2. Background

2.1. Time-Series Forecasting models

Time series forecasting has evolved significantly, initially relying on linear models like ARIMA Zhang (2003) and Exponential Smoothing Gardner Jr (1985). However, with deep learning advancements, neural network architectures such as LSTM Yu et al. (2019) and GRU Dey and Salem (2017) showed superior performance over traditional methods.

Recently, Convolutional Neural Networks (CNNs) Alzubaidi et al. (2021) and Temporal Convolutional Networks (TCNs) Hewage et al. (2020) have demonstrated state-of-the-art results. The Transformer architecture Wen et al. (2022) was also adapted for forecasting, with models like AutoFormer Wu et al. (2021) and FEDformer Zhou et al. (2022) as leading architectures. However, Zeng et al. (2022) proposed DLinear, a simple linear model, challenging the efficacy of transformers. The subsequent models TimesNet Wu et al. (2023), and PatchTST Nie et al. (2022) improved upon DLinear, while the NFT model Koren and Radinsky (2024) emerged as a top-performing multivariate time-series model.

Multidimensional Fourier Transform. The Fourier Transform has been widely used in time series analysis to identify periodic patterns or cycles in the data Yi et al. (2023). By converting time-series data into the frequency domain, one can identify the main frequencies at which these cycles occur Nussbaumer (1982).

Several forecasting models use Fourier Transforms for better performance. Autoformer employs Fast Fourier Transform for autocorrelation Wu et al. (2021), FEDformer focuses on key frequencies Zhou et al. (2022), and the Fourier Neural Operator approximates partial differential equations operators with Fourier Transforms Li et al. (2020). TimesNet uses Fourier Transforms for feature decomposition to capture periodic patterns Wu et al. (2023).

The *Multidimensional Fourier Transform (MFT)* Tolimieri et al. (2012) extends the traditional Fourier Transform to handle multi-dimensional data. We drew inspiration from the Neural Fourier Transform (NFT) Koren and Radinsky (2024), and used a 2-dimensional Discrete Fourier Transform (DFT) to extract the seasonality of the time series.

2.2. Membership Inference Attacks

Membership inference attacks (MIAs) represent a significant privacy threat in machine learning. In these attacks, an adversary aims to determine whether a specific data record, x, was included in the training set of a model, D. If successful, the attack can reveal sensitive information about individuals, such as their medical history, financial status, or personal preferences. Moreover, MIAs can also be used to identify individuals who are part of a specific group or community, potentially leading to discrimination, or physical harm.

Formally, given access to a machine learning model M, the attacker seeks to ascertain the membership of a data sample x in D, i.e; to check if $x \in D$. To this end the attacker typically analyzes M's outputs, and produces numeric characteristics (features), that will enable it to distinguish members of the training data from non- members. Such features may include M's loss Shokri et al. (2016) on the sample, the log-probabilities Carlini et al. (2022b) and entropy of the outputs.

In the context of time-series forecasting models, a malicious attacker seeks to determine whether a specific time series was utilized in the model's training dataset, such as a patient's ECG test results. This type of attack poses a significant threat in industries like healthcare and finance, where sensitive time-series data is frequently leveraged to develop predictive models, and the unauthorized disclosure of such information could have severe consequences.

3. Related work

In the realm of time series, Hisamoto et al. studied membership inference on sequence-to-sequence (seq2seq) models in the context of machine translation, where the output is a chained sequence of classifications Hisamoto et al. (2020). This differs from medical sequence modeling whose input features and outputs are numerical and continuous.

Pyrgelis et al. Pyrgelis et al. (2017) presented the first study on the feasibility of MIAs on aggregate location time series, modeling the problem as a classification task to distinguish whether a target user is part of an aggregate. Their empirical evaluation on mobility datasets shows that MIAs are a privacy threat, influenced by the adversary's prior knowledge, data characteristics, number of users, and aggregation timeframe.

Similarly, Voyez et al. Voyez et al. (2022) explored the vulnerability of aggregated time-series data to MIAs, introducing a linear programming-based attack that leverages the correlation between the length of the published time series and the size of the aggregated data. Their experiments demonstrate that aggregated time-series data can be highly susceptible to privacy breaches, emphasizing the need for better privacy-preserving techniques, particularly in the medical domain.

However, to our knowledge, risk assessment in general and MIA specifically has not been thoroughly explored on ML models trained on numerical time-series data. This presented us with an opportunity for novel applications and advancements in attacking time-series models, potentially unlocking new insights and methodologies in this area.

4. Methodology

4.1. Problem Statement

This study focuses on MIA on multivariate time-series forecasting models. We assume that the attacker can access a complete sample that was either used in model training or not.

In time-series data, training samples consist of data points up to time T (lookback), denoted as $\mathbf{y} = [y_1, \ldots, y_T] \in \mathbb{R}^{M \times T}$, and the model predicts H data points onward (horizon), denoted as $\mathbf{Y} = [y_{T+1}, \ldots, y_{T+H}] \in \mathbb{R}^{M \times H}$, where $y_t \in \mathbb{R}^M$ for $t = 1, \ldots, T + H$, and M is the number of variables.

To simplify, we consider a lookback window of length $t \leq T$, ending at the most recent observation y_T. This window serves as the input (sample) to the model and is denoted $\mathbf{X} \in \mathbb{R}^{M \times t} = [y_{T-t+1}, \ldots, y_T]$. The forecast of $\mathbf{Y}$ is represented as $\hat{\mathbf{Y}}$.

Our task is to determine if a specific sample, $\mathbf{X}$, is part of the training data, $\mathbf{D}$, i.e; $\mathbf{X} \in \mathbf{D}$ by comparing the real future values, $\mathbf{Y}$, and the model's predicted values, $\hat{\mathbf{Y}}$.

4.2. Features for MIA

In the context of MIA, the attack features are the set of attributes or characteristics that the attack model leverages to determine whether a given data sample was a part of the training set (member) or not (non-member). As in any ML model, selecting the correct attack features is critical, as they form the basis upon which the attack model makes its predictions. An optimal set of attack features can significantly improve the success of the attack, potentially posing a much higher privacy risk.

Our goal is to find attack features that will yield good MIA results for time-series models. This involves leveraging the characteristics of time-series data by identifying features that capture its unique aspects. Hence, the introduced features isolate the seasonality and trend components of the model's prediction, contrasted with those of the true data. Figure 1 illustrates this process from the initial forecasting model to the final attack setup.

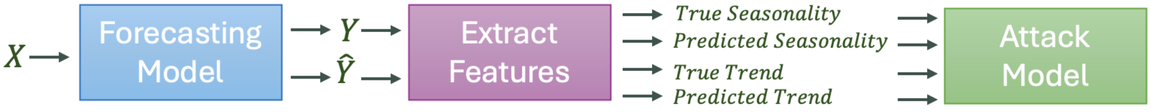

Figure 1: Flowchart illustrating the process from the initial forecasting model to the final attack model, highlighting how extracted features are used for MIAs.

Given that time-series data inherently includes components such as trend and seasonality, models trained on such data are particularly good at capturing those elements. This proficiency is leveraged by a variety of forecasting models, including Neural Fourier Transform (NFT) Koren and Radinsky (2024), TimesNet Wu et al. (2023), Fedformer Zhou et al. (2022), Autoformer Wu et al. (2021), Fourier Neural Operator (FNO) Li et al. (2020), N-BEATS Oreshkin et al. (2019), NeuralProphet Triebe et al. (2021), and ARIMA Zhang

(2003), all of which explicitly integrate these elements into their predictions. Thus, when attacking a time-series prediction model, we aim to take advantage of the model's ability to accurately predict the series' trend and seasonality, thus offering a tactical advantage in MIAs. This highlights the critical need to consider these characteristics when evaluating the susceptibility of time-series models to such privacy threats.

In this work, to effectively capture the seasonality, we employ the Multidimensional Fourier Transform, which excels in extracting periodic patterns from time-series data Musbah and El-Hawary (2019). We identify the predominant trend through a low-degree polynomial fit, allowing us to find the principal direction while filtering variations Masry (1996).

4.2.1. Seasonality

We detect seasonality in multivariate temporal data with the 2-dimensional Discrete Fourier Transform (2D-DFT) as done in Koren and Radinsky (2024). This method breaks down the dataset into its frequency components, considering both the range of variables and the timeline. This approach is essential for datasets where the interaction between different variables can create new seasonality patterns that are not seen in the univariate context.

The 2D-DFT is applied to the matrix $\mathbf{Y} \in \mathbb{R}^{M \times H}$, where M is the number of variables, and H is the number of predicted time points. This process involves two sequential 1D-DFTs. First, a column-wise 1D-DFT is applied to Y using the Fourier matrix F_M, which captures transformations across the variables. Next, a row-wise 1D-DFT is performed using the Fourier matrix F_H, which encodes the temporal structure of the data.

The 2D-DFT can be compactly represented as:

$$\mathbf{C} = \mathbf{F_M} \mathbf{Y} \mathbf{F_H}^{\top} \tag{1}$$

Here, the matrix C contains the Fourier coefficients.

Following are the Fourier matrices F_M and F_H that achieve the desired Fourier transformation:

$$F_M = \begin{bmatrix} \cos(2\pi \cdot 0 \cdot \frac{0}{M}) & \cdots & \cos(2\pi \cdot 0 \cdot \frac{M-1}{M}) \\ \vdots & \vdots & \vdots \\ \cos(2\pi \cdot \frac{M}{2} \cdot \frac{0}{M}) & \cdots & \cos(2\pi \cdot \frac{M}{2} \cdot \frac{M-1}{M}) \\ \sin(2\pi \cdot 0 \cdot \frac{0}{M}) & \cdots & \sin(2\pi \cdot 0 \cdot \frac{M-1}{M}) \\ \vdots & \vdots & \vdots \\ \sin(2\pi \cdot \frac{M}{2} \cdot \frac{0}{M}) & \cdots & \sin(2\pi \cdot \frac{M}{2} \cdot \frac{M-1}{M}) \end{bmatrix}$$

$$F_H = \begin{bmatrix} \cos(2\pi \cdot 0 \cdot \frac{0}{H}) & \cdots & \cos(2\pi \cdot 0 \cdot \frac{H-1}{H}) \\ \vdots & \vdots & \vdots \\ \cos(2\pi \cdot \frac{H}{2} \cdot \frac{0}{H}) & \cdots & \cos(2\pi \cdot \frac{H}{2} \cdot \frac{H-1}{H}) \\ \sin(2\pi \cdot 0 \cdot \frac{0}{H}) & \cdots & \sin(2\pi \cdot 0 \cdot \frac{H-1}{H}) \\ \vdots & \vdots & \vdots \\ \sin(2\pi \cdot \frac{H}{2} \cdot \frac{0}{H}) & \cdots & \sin(2\pi \cdot \frac{H}{2} \cdot \frac{H-1}{H}) \end{bmatrix}$$

The input features to the attack derived from this method are:

1. Coefficients of the Fourier series corresponding to the true values:

$$\mathbf{C} = \mathbf{F_1}^{\top} \times \mathbf{Y} \times \mathbf{F_2}$$

2. Coefficients of the Fourier series corresponding to the model's predicted values:

$$\hat{\mathbf{C}} = \mathbf{F_1}^\top \times \hat{\mathbf{Y}} \times \mathbf{F_2}$$

3. The L_2 norm between the coefficients of the true and predicted values:

$$||\mathbf{C} - \hat{\mathbf{C}}||_2$$

4.2.2. Trend

Consider a multivariate time series Y with H time points and M variables. Each variable's series is approximated using a polynomial of degree d. This approximation can be represented as:

$$\mathbf{Y} = \mathbf{P} \times \mathbf{A} \tag{2}$$

where A is the coefficients matrix, and P is the Vandermonde matrix, constructed from the time vector $t = \frac{[0,1,\ldots,H-1]}{H}$. P contains powers of t up to $d-1$, has dimensions $d \times H$ and is defined as:

$$P = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ t_1 & t_2 & \cdots & t_H \\ t_1^2 & t_2^2 & \cdots & t_H^2 \\ \vdots & \vdots & \ddots & \vdots \\ t_1^{d-1} & t_2^{d-1} & \cdots & t_H^{d-1} \end{bmatrix}$$

where $t_i = \frac{i-1}{H}$ for $i = 1, 2, \ldots, H$.

The coefficients matrix A is obtained by the least squares solution: $\mathbf{A} = (\mathbf{P}^T\mathbf{P})^{-1}\mathbf{P}^T\mathbf{Y}$

The input features to the attack derived from this method are:

1. Coefficients of the polynomial outlining the trend of the true values:

$$\mathbf{A} = (\mathbf{P}^T\mathbf{P})^{-1}\mathbf{P}^T\mathbf{Y}$$

2. Coefficients of the polynomial outlining the trend of the model's predicted values:

$$\hat{\mathbf{A}} = (\mathbf{P}^T\mathbf{P})^{-1}\mathbf{P}^T\hat{\mathbf{Y}}$$

3. The L_2 norm between the coefficients of the true and predicted values:

$$||\mathbf{A} - \hat{\mathbf{A}}||_2$$

4.2.3. Mean Absolute Scaled Error and Mean Squared Error

Additionally, this study investigates the use of the Mean Absolute Scaled Error (MASE), a scaled measure for assessing forecast accuracy by comparing the mean absolute error of a model against a naïve baseline forecast, as outlined by Hyndman and Koehler (2006), and the Mean Squared Error (MSE), Das et al. (2004), as features for the attack model.

$$\text{MASE} = \frac{\frac{1}{H}\sum_{i=1}^{H} |y_{T+i} - \hat{y}_{T+i}|}{\frac{1}{H-1}\sum_{i=2}^{H} |y_{T+i} - y_{T+i-1}|}, \text{ MSE} = \frac{1}{H}\sum_{i=1}^{H}(y_{T+i} - \hat{y}_{T+i})^2$$

These metrics are extended to the multivariate case by averaging the respective univariate values across all variables.

5. Experimental Setup

5.1. Threat Model and MIA Attack Setup

Following previous privacy assessment studies Shachor et al. (2023); Amit et al. (2024); Anderson et al. (2024), we adopt a gray-box threat model, assuming the attacker has access to both a subset of the model's training data and a set of non-training samples. The access to this data, allows estimating a worst case privacy risk for a model before deployment, without the need of developing shadow models Shokri et al. (2016).

To execute the attack, we leverage the privacy risk assessment framework from Shachor et al. (2023), which builds upon recent breakthroughs in MIAs Carlini et al. (2022a); Shokri et al. (2016). This framework leverages the ensemble approach, creating many specialized attack models for different subsets of the data. The framework harnesses a diverse set of input features extracted from the target model's inputs and outputs. Through an exhaustive grid search, it systematically explores various attack model architectures, hyperparameters, and preprocessing techniques to identify the optimal configuration that yields maximum attack performance.

The attack models trained by the risk assessment framework utilize various combinations of the following features: (1) Seasonality denoted as, $\boldsymbol{S}$, (2) Trend denoted as, $\boldsymbol{T}$ (with a polynomial degree of 4 like done in Koren and Radinsky (2024); Oreshkin et al. (2019)), (3) $\boldsymbol{MASE}$, (4) $\boldsymbol{MSE}$, and (5) Predicted Values denoted as, $\boldsymbol{PV}$.

In this evaluation, we applied the framework to conduct five attack instances, each with three runs. An instance refers to executing the entire attack optimization process on a different random data sample, while the runs involve different splits of the data sample to fit and infer the attack model within each instance. For each instance, a sample of 450 members and 450 non-members was chosen at random. Results were averaged across all runs and instances to ensure robustness.

5.2. Datasets

In this evaluation, two multivariate time-series medical datasets were used:

- **EEG**: 36-lead EEG database, which contains more than 1000 EEG recordings dating from 2002 to the present, sampled at a frequency of 250 Hz Obeid and Picone (2016). Our subset includes data from 32 patients, and the first 3-leads for each patient.

- **ECG**: Georgia 12-Lead ECG Challenge Database, curated by Emory University Goldberger et al. (2000). The complete database contains ECG recordings of over 10,000 individuals, sampled at a frequency of 500 Hz, and collected from various healthcare settings worldwide. Our subset features ECG time-series data from 600 individuals.

Data preprocessing included outlier removal using the Interquartile Range method, imputation of missing values via mean substitution, and data standardization.

For both datasets, The data was partitioned into three distinct subsets: 42.5% of the patients were used for training the model, 15% for validation, and the remaining 42.5% were reserved as non-member data points for the attack model. The validation set was used to tune the model's hyperparameters, thus creating strong models to attack. The non-member data points remained uninvolved in training or validating the models but were used in subsequent attack experiments. Additionally, the data was split into lookbacks and horizons using the sliding window approach. Statistics on the datasets can be found in Table 1.

Table 1: Datasets Statistics

Dataset	Num of Variables	Timesteps	Lookback	Prediction Horizons
EEG	3	9620519	100	1, 5, 10, 15, 20
ECG	12	2393563	100	1, 5, 10, 15, 20, 25, 30

5.3. Models

We performed attacks against various state-of-the-art time-series forecasting architectures:

- **DLinear** Zeng et al. (2022): Featuring a dimension of 16 and a dropout rate of 0.1.
- **Temporal Convolutional Network (TCN)** Hewage et al. (2020): Configured with channels set to [2, 2], a kernel size of 2, and a dropout rate of 0.2.
- **Long Short Term Memory (LSTM)** Yu et al. (2019): Featuring a 2-layer structure with hidden dimensions set to 50.
- **Neural Fourier Transform (NFT)** Koren and Radinsky (2024): Configured with Fourier granularity of 8 for the seasonality blocks and a polynomial degree of 4 for the trend blocks, comprising 2 blocks per stack.
- **TimesNet** Wu et al. (2023): Model dimension was set to 16 and a dropout rate of 0.1.
- **PatchTST** Nie et al. (2022): This transformer model included one encoder and decoder layer, a model dimension of 16, and a dropout rate of 0.1.

The parameters for each model were chosen based on their performance on a validation set, ensuring optimal configuration for our analysis. In Table 3, the number of parameters for each model is detailed. For all models, the MSE loss was utilized during model training. Figure 2 presents the performance (MSE) of each model on the test set (non-members) for different prediction horizons.

6. Results

To assess the robustness of the proposed features for MIAs on time-series models, we compare them to baseline attacks, also applied using the same grey-box approach, where the attacker has access to both a subset of the model's training data and a set of non-training samples. Typically, these attacks are based on the model's loss or predictions, similar to those in the black-box setting Gupta et al. (2021). In all experiments, four baselines were employed: loss-based attacks using MSE, MASE, and a combination of both, as well as attacks based on predicted values.

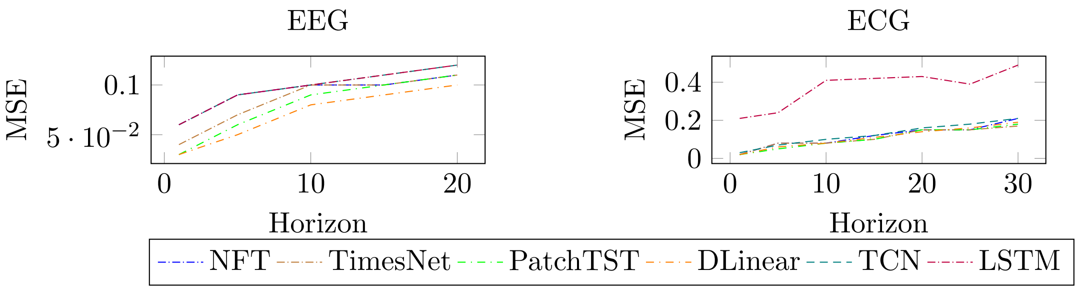

Figure 2: MSE values of models with varying horizons on ECG and EEG datasets

The results were evaluated using two key metrics: Area Under the Receiver Operating Characteristic curve (AUC-ROC) and True Positive Rate (TPR) at a fixed False Positive Rate (FPR) of 1%.

6.1. AUC-ROC Results

6.1.1. Performance Against Loss-Based Attacks

Figures 3 and 4 highlight the strong performance of the Seasonality (S) and Trend (T) features across various time-series models and datasets. The highest AUC-ROC values were achieved with feature combinations that included the Trend or Seasonality features, outperforming the baseline attacks by up to 26%. For the EEG dataset, when looking at the different models and averaging across horizons, the best-performing feature combination compared to the MSE-only attack, presented an improvement of between **8.44%** (with std **0.007**) and **26.41%** (with std **0.006**). For the ECG dataset, improvements ranged from **2.97%** (with std **0.008**) to **24.55%** (with std **0.007**). The TimesNet model showed the best improvements in both datasets.

The MASE feature achieved the lowest attack performance, however, the combination of MASE and MSE generally surpassed attacks that use the MSE feature alone.

Overall, the results indicate that the Seasonality and Trend features provide a robust solution for MIAs against time-series forecasting models. Its consistent performance across different models, datasets, and horizons highlights its effectiveness.

Analysis of Attack Performance by Prediction Horizon. We analyzed the relation between the attack AUC-ROC and the prediction horizon. To this end, we computed, for each dataset, the correlation between the prediction horizon (previously denoted by H) and the percentage of AUC-ROC improvement between the highest attack value and the MSE-only attack, presented in Table 2. We consistently observed a positive correlation for all models except for NFT and TCN on the ECG dataset. We assume that the negative correlation can be traced back to the difference in architecture. Notably, the NFT and TCN models are both based on convolutional layers, which capture the trend and seasonality in a different manner. The results presented in the table highlight the greater increase in vulnerability of time-series prediction models to attacks as the prediction horizon increases, compared to the MSE-based baseline attack. This suggests an enhanced attack vector against these models, which is not present in other model architectures.

Table 2: Correlation between Prediction Horizons and AUC-ROC Improvement Percentages

Model	**Correlation EEG**	**Correlation ECG**
PatchTST	0.859	0.942
TimesNet	0.796	0.732
DLinear	0.883	0.177
LSTM	0.482	0.376
TCN	0.780	-0.150
NFT	0.089	-0.931

Analysis of Attack Performance by Model. Table 3 presents the average AUC-ROC results for each model across features and horizons. For the ECG dataset, PatchTST emerged as the most vulnerable model, indicating it is highly susceptible to attacks. In contrast, DLinear was the least vulnerable and hardest to attack. Notably, PatchTST has the highest number of parameters (3,000,000), while DLinear has the fewest (800). In the EEG dataset, DLinear consistently showed superior resilience against attacks, and the vulnerability of the other models varied across the datasets.

Table 3: Number of Model Parameters vs. Average AUC-ROC Results Across Features and Horizons for EEG and ECG Datasets

Model	DLINEAR	TCN	LSTM	NFT	TimesNet	PatchTST
# Parameters	8×10^2	1×10^3	34×10^3	13×10^4	16×10^4	3×10^6
EEG Average AUC-ROC	0.59	0.82	0.83	0.82	0.6	0.72
ECG Average AUC-ROC	0.58	0.56	0.56	0.56	0.72	0.83

6.1.2. Performance Against Predicted Values Based Attacks

We examined the effects of incorporating Predicted Values (PV) as a feature in the attack model. The Predicted Values were added to every feature combination and tested independently. The results on the EEG and ECG datasets are illustrated in Figure 5, with some models omitted due to space limitations. The black curve, labeled *Benchmark Limit*, represents the upper bound of the AUC-ROC values for the feature combinations without Predicted Values, as presented in the previous section.

The analysis of the results did not reveal a consistent trend across horizons and datasets. For the EEG dataset, adding the Predicted Values consistently under-performed compared to the feature combinations without them. In contrast, the results for the ECG dataset were more varied. Although the addition of the Predicted Values feature often surpassed the Benchmark Limit for different models and horizons, the results varied widely and no feature combination consistently exceeded the Benchmark Limit across all horizons.

Overall, the results indicate that while the Predicted Values feature can in some cases improve the performance of the attack model, its impact is not straightforward and de-

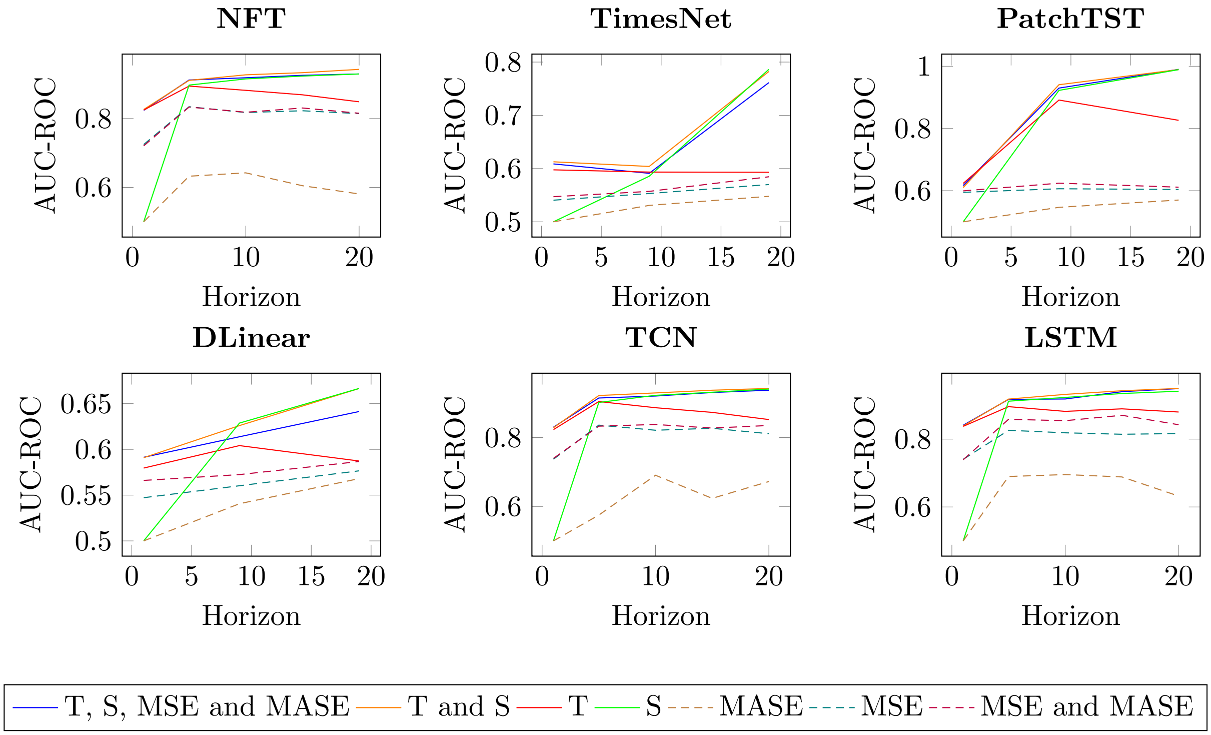

Figure 3: AUC-ROC Performance Across Different Models on EEG Data

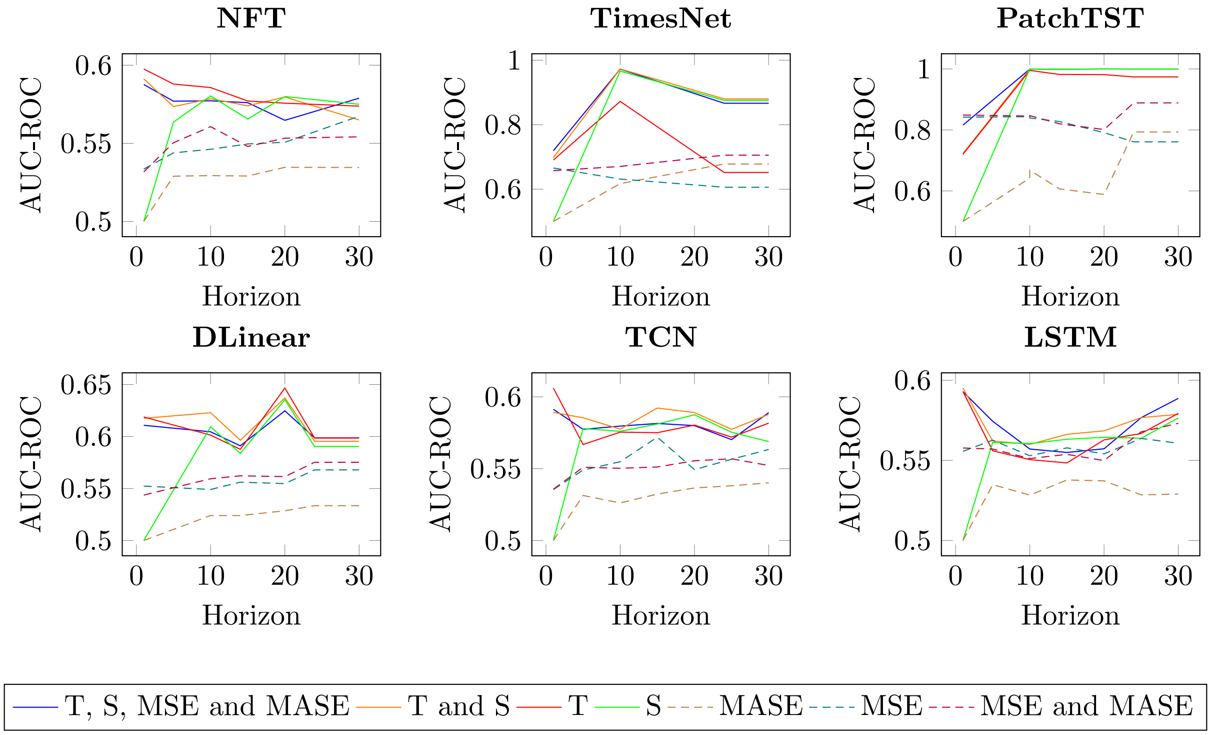

Figure 4: AUC-ROC Performance Across Different Models on ECG Data

pends on the specific model, dataset and horizon. The figures highlight the complexity and variability of incorporating Predicted Values in this context.

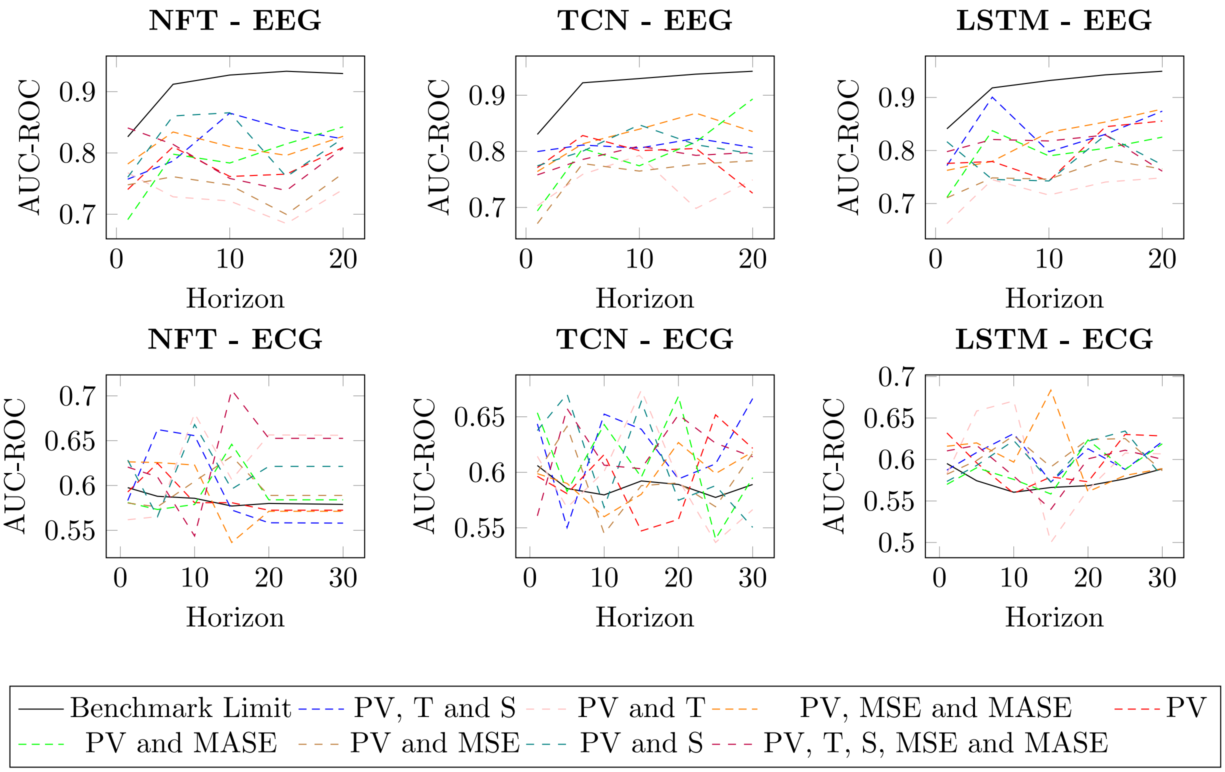

Figure 5: AUC-ROC Performance Across Different Models on the EEG and ECG datasets. The Predicted Values feature was added to each feature combination

6.2. True Positive Rate at 1% False Positive Rate Results

In addition to AUC-ROC, we measured the attack's True Positive Rate at a False Positive Rate of 1% for a fixed horizon of 5. The results are presented in Tables 4 and 5.

For the EEG dataset, using the Seasonality (S) feature with the PatchTST model resulted in a 9x increase in TPR compared to using the MSE feature alone. Significant improvements were observed across the NFT, TimesNet, and LSTM models when using the Trend (T) feature, with gains ranging from 2.1x to 4x over the MSE-only baseline. The TCN model demonstrated a 2.1x improvement with the combination of Trend and Seasonality, while DLinear showed similar gains with both combined and standalone Seasonality features. In the ECG dataset, the Trend feature and its combination with Seasonality in the NFT model led to a 3x improvement in TPR compared to the MSE-only baseline. The Seasonality feature led to notable improvements for the TimesNet and PatchTST models, with up to a 12x increase.

Table 4: Average TPR at 1% FPR for EEG Dataset at Horizon 5

Model	MSE	MASE	MSE and MASE	T	S	T and S	T, S, MASE, MSE, and PV	PV
DLinear	0.03	0.02	0.04	0.06	**0.07**	**0.07**	0.06	0.06
TCN	0.23	0.04	0.23	0.40	0.42	**0.43**	0.35	0.31
LSTM	0.18	0.10	0.20	**0.44**	0.39	0.38	0.39	0.27
NFT	0.20	0.06	0.21	**0.42**	0.37	0.41	0.38	0.30
TimesNet	0.02	0.01	0.03	**0.08**	0.06	0.05	0.06	0.06
PatchTST	0.07	0.02	0.06	0.28	**0.63**	0.39	0.35	0.13

Table 5: Average TPR at 1% FPR for ECG Dataset at Horizon 5

Model	MSE	MASE	MSE and MASE	T	S	T and S	T, S, MASE, MSE, and PV	PV
DLinear	0.01	0.01	0.02	**0.03**	**0.03**	**0.03**	**0.03**	**0.03**
TCN	0.01	0.01	0.01	**0.02**	**0.02**	**0.02**	**0.02**	**0.02**
LSTM	0.02	0.02	0.01	0.02	0.02	0.01	0.02	**0.04**
NFT	0.01	0.02	0.02	**0.03**	0.02	**0.03**	0.02	0.02
TimesNet	0.03	0.05	0.03	0.14	**0.36**	0.35	0.36	0.03
PatchTST	0.34	0.03	0.32	0.52	**0.99**	0.51	0.66	0.03

7. Conclusion and Future Work

This paper is the first to explore membership inference attacks on numeric time-series machine-learning models. Our primary contribution is the introduction of two novel features, Trend and Seasonality, derived from low-degree polynomial fitting and Multivariate Fourier Transform, respectively. These features enhance MIA models by leveraging the inherent characteristics of time-series data, improving the identification of member samples.

Various state-of-the-art forecasting models incorporate those components into their design. Consequently, when targeting a time-series prediction model, there is a significant likelihood that the model will precisely estimate the series' Seasonality and Trend of its training data, providing a strategic advantage in MIAs.

The effectiveness of these features was tested on six diverse models using two medical datasets. The Trend and Seasonality features showed superior accuracy, with a 3% to 26% improvement in attack AUC-ROC scores and up to 12x improvement in TPR at low FPR over traditional features across various horizons.

We plan to explore additional MIA scenarios, including models pre-trained on many patients and fine-tuned for home monitoring of a specific patient, to see if such an attack can expose the original training data. Additionally, we aim to investigate user-level attacks that exploit the fact that multiple samples in the training set belong to the same person.

Acknowledgment

This work was performed as part of the NEMECYS project, which is co-funded by the European Union under grant agreement ID 101094323, by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee grant numbers 10065802, 10050933, and 10061304, and by the Swiss State Secretariat for Education, Research and Innovation (SERI).

References

Laith Alzubaidi, Jinglan Zhang, Amjad J Humaidi, Ayad Al-Dujaili, Ye Duan, Omran Al-Shamma, José Santamaría, Mohammed A Fadhel, Muthana Al-Amidie, and Laith Farhan. Review of deep learning: concepts, cnn architectures, challenges, applications, future directions. *Journal of big Data*, 8:1–74, 2021.

Guy Amit, Abigail Goldsteen, and Ariel Farkash. Sok: Reducing the vulnerability of fine-tuned language models to membership inference attacks. 2024.

Maya Anderson, Guy Amit, and Abigail Goldsteen. Is my data in your retrieval database? membership inference attacks against retrieval augmented generation. *arXiv preprint arXiv:2405.20446*, 2024.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1897–1914. IEEE, 2022a.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1897–1914. IEEE, 2022b.

Kalyan Das, Jiming Jiang, and JNK Rao. Mean squared error of empirical predictor. 2004.

Rahul Dey and Fathi M Salem. Gate-variants of gated recurrent unit (gru) neural networks. In *2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS)*, pages 1597–1600. IEEE, 2017.

Everette S Gardner Jr. Exponential smoothing: The state of the art. *Journal of forecasting*, 4(1):1–28, 1985.

A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. Ch. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C. K. Peng, and H. E. Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000.

Umang Gupta, Dimitris Stripelis, Pradeep K Lam, Paul Thompson, Jose Luis Ambite, and Greg Ver Steeg. Membership inference attacks on deep regression models for neuroimaging. pages 228–251, 2021.

Jamie Hayes, Luca Melis, George Danezis, and Emiliano De Cristofaro. Logan: Membership inference attacks against generative models. *arXiv preprint arXiv:1705.07663*, 2017.

Xinlei He, Rui Wen, Yixin Wu, Michael Backes, Yun Shen, and Yang Zhang. Node-level membership inference attacks against graph neural networks. *arXiv preprint arXiv:2102.05429*, 2021.

Pradeep Hewage, Ardhendu Behera, Marcello Trovati, Ella Pereira, Morteza Ghahremani, Francesco Palmieri, and Yonghuai Liu. Temporal convolutional neural (tcn) network for an effective weather forecasting using time-series data from the local weather station. *Soft Computing*, 24:16453–16482, 2020.

Sorami Hisamoto, Matt Post, and Kevin Duh. Membership Inference Attacks on Sequence-to-Sequence Models: Is My Data In Your Machine Translation System? *Transactions of the Association for Computational Linguistics*, 8:49–63, 01 2020. ISSN 2307-387X.

Rob J Hyndman and Anne B Koehler. Another look at measures of forecast accuracy. *International journal of forecasting*, 22(4):679–688, 2006.

Noam Koren and Kira Radinsky. Interpretable multivariate time series forecasting using neural fourier transform. *arXiv preprint arXiv:2405.13812*, 2024.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. *arXiv preprint arXiv:2010.08895*, 2020.

Elias Masry. Multivariate regression estimation local polynomial fitting for time series. *Stochastic Processes and their Applications*, 65(1):81–101, 1996.

Hmeda Musbah and Mo El-Hawary. SARIMA model forecasting of short-term electrical load data augmented by fast fourier transform seasonality detection. In *2019 IEEE Canadian Conference of Electrical and Computer Engineering (CCECE)*, pages 1–4. IEEE, 2019.

Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*, 2022.

Henri J Nussbaumer. *The fast Fourier transform*. Springer, 1982.

Iyad Obeid and Joseph Picone. The Temple University Hospital EEG data corpus. *Frontiers in neuroscience*, 10:196, 2016.

Boris N Oreshkin, Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio. N-beats: Neural basis expansion analysis for interpretable time series forecasting. *arXiv preprint arXiv:1905.10437*, 2019.

Apostolos Pyrgelis, Carmela Troncoso, and Emiliano De Cristofaro. Knock knock, who's there? membership inference on aggregate location data. *arXiv preprint arXiv:1708.06145*, 2017.

Shlomit Shachor, Natalia Razinkov, and Abigail Goldsteen. Improved membership inference attacks against language classification models. In *arXiv:2310.07219*, 2023.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models (s&p'17). 2016.

Richard Tolimieri, Myoung An, and Chao Lu. *Mathematics of multidimensional Fourier transform algorithms*. Springer Science & Business Media, 2012.

Oskar Triebe, Hansika Hewamalage, Polina Pilyugina, Nikolay Laptev, Christoph Bergmeir, and Ram Rajagopal. Neuralprophet: Explainable forecasting at scale. *arXiv preprint arXiv:2111.15397*, 2021.

Stacey Truex, Ling Liu, Mehmet Emre Gursoy, Lei Yu, and Wenqi Wei. Demystifying membership inference attacks in machine learning as a service. *IEEE transactions on services computing*, 14(6):2073–2089, 2019.

Antonin Voyez, Tristan Allard, Gildas Avoine, Pierre Cauchois, Elisa Fromont, and Matthieu Simonin. Membership inference attacks on aggregated time series with linear programming. In *SECRYPT 2022 - 19th International Conference on Security and Cryptography*, pages 193–204. Springer, 2022.

Qingsong Wen, Tian Zhou, Chaoli Zhang, Weiqi Chen, Ziqing Ma, Junchi Yan, and Liang Sun. Transformers in time series: A survey. *arXiv preprint arXiv:2202.07125*, 2022.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in neural information processing systems*, 34:22419–22430, 2021.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. In *The eleventh international conference on learning representations*, 2023.

Kun Yi, Qi Zhang, Longbing Cao, Shoujin Wang, Guodong Long, Liang Hu, Hui He, Zhendong Niu, Wei Fan, and Hui Xiong. A survey on deep learning based time series analysis with frequency transformation. *arXiv preprint arXiv:2302.02173*, 2023.

Yong Yu, Xiaosheng Si, Changhua Hu, and Jianxun Zhang. A review of recurrent neural networks: Lstm cells and network architectures. *Neural computation*, 31(7):1235–1270, 2019.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 11121–11128, 2022.

G Peter Zhang. Time series forecasting using a hybrid arima and neural network model. *Neurocomputing*, 50:159–175, 2003.

Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *International conference on machine learning*, pages 27268–27286. PMLR, 2022.