Models of Sequential Data 2023

Final project

Adversarial Attack on Time Series

Alina Bogdanova, Nikita Ligostaev,
Anastasia Sozykina, Matvey Skripkin

TA:
Petr Sokerin

October 17, 2023

Problem Statement

Problem we solve: developing a concealed adversarial attack for time series classification models to evade easy detection by discriminators

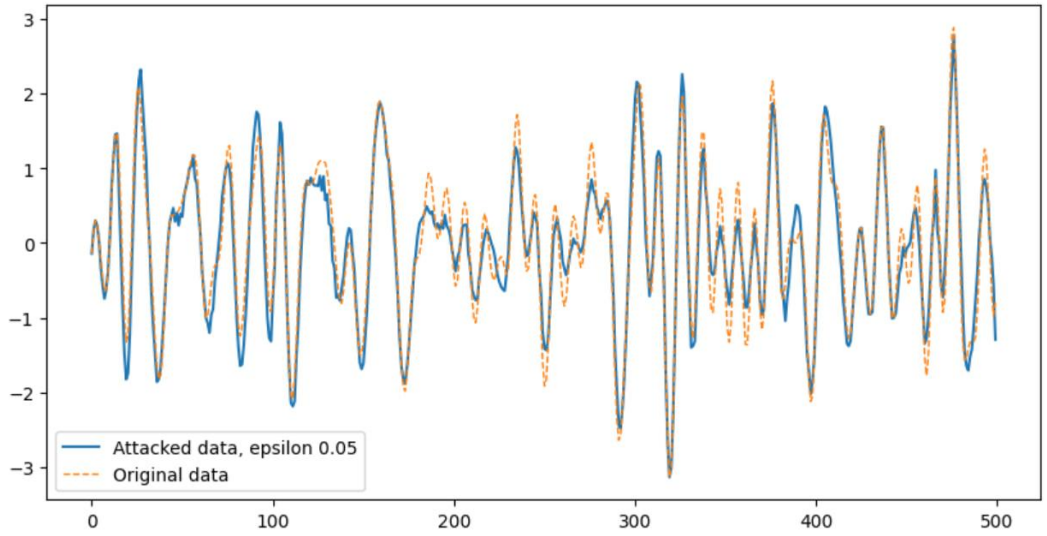


Fig 1. Example of adversarial attack that can be easily detected by discriminator

Adversarial attacks in different domains

- Adversarial attacks include a wide range of techniques that perturb input data in order to cause misclassification or degrade model performance.

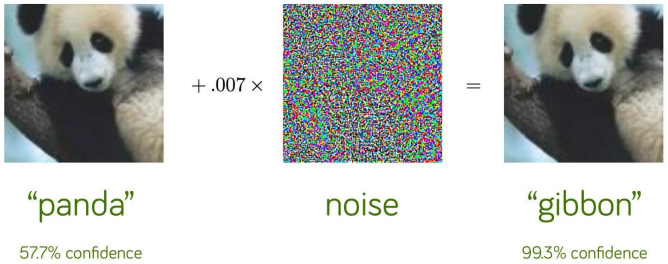


Fig 2. Adversarial attacks on image classification [1]

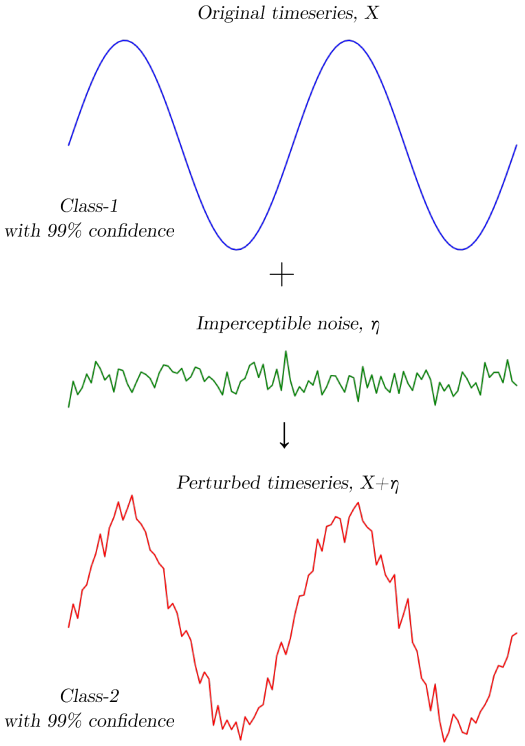


Fig 3. Adversarial attacks on time series classification [2]

1 https://arxiv.org/abs/1412.6572
2 https://link.springer.com/article/10.1007/s10618-019-00619-1

Motivation

- Adversarial attacks in time-series domain is not well-studied.
- Time series classification might have crucial impact in industry.

Idea description

- How to apply special hidden method of adversarial attacks to different time series Neural Networks that it could not be easily detected by special classificator model?

- **Goal of the project is to apply such hidden method of adversarial attacks to time series models and estimate the hiddenness and effectiveness of the attack.**

  - **Effectiveness:**

    1 - accuracy of attacked model on perturbed data

  - **Concealability:**

    1 - accuracy of discriminator

1 https://doi.org/10.48550/arXiv.2201.12220

Adversarial attack methods

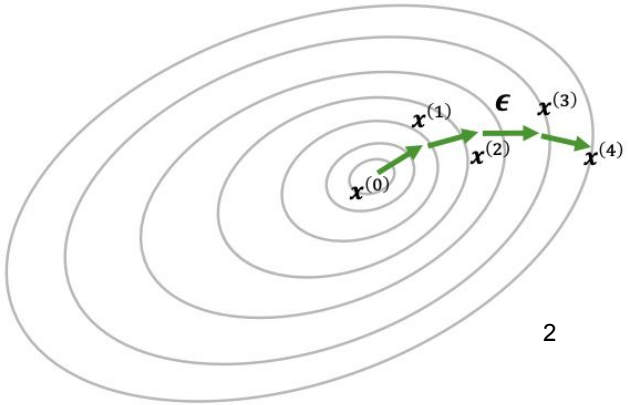
- **IFGSM**

Algorithm 1 Iterative Fast Gradient Sign Method (IFGSM) for Binary Classifier (Kurakin et al., 2018) [1]

Require: Time series X, classifier f, maximum iterations T, step size α.

Ensure: Perturbation $\hat{r}$.

1: Initialize $X_0 \leftarrow X$, $i \leftarrow 0$.
2: **while** $i < T$ **do**
3: Compute gradient: $\nabla J(X_i, Y)$ where J is the loss function and Y is the target class.
4: Perturb time series: $X_{i+1} \leftarrow X_i + \alpha \cdot \text{sign}(\nabla J(X_i, Y))$.
5: Clip perturbation: $X_{i+1} \leftarrow \text{clip}(X_{i+1}, X-\epsilon, X+\epsilon)$, where ϵ is a bound on the perturbation.
6: $i \leftarrow i + 1$
7: **end while**
8: **return:** $\hat{r} = X_T - X$.



2

$$x^{(t+1)} = x^{(t)} + \epsilon \, \text{sign}\left(\nabla_x J(\theta, x^{(t)}, y)\right)$$

1 https://arxiv.org/pdf/1607.02533.pdf
2 Illustration provided by P. Sokerin

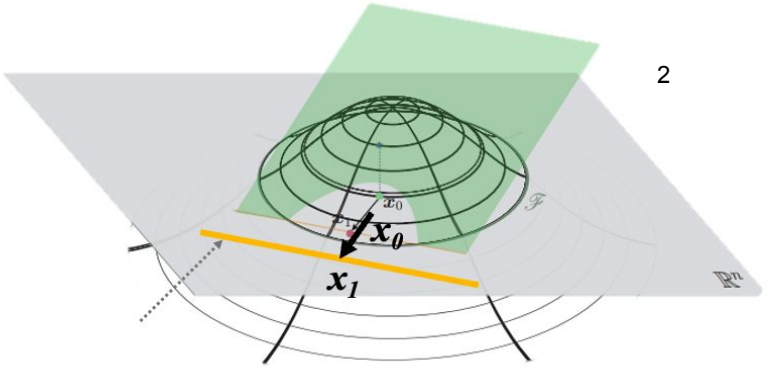
Adversarial attack methods

- **DeepFool**

Algorithm 2 DeepFool for Binary Classifier (Moosavi-Dezfooli et al., 2016) [1]

Require: Time series X, classifier f.
Ensure: Perturbation $\hat{r}$.

1: Initialize $X_0 \leftarrow X, i \leftarrow 0$.
2: **while** $\text{sign}(f(X_i)) = \text{sign}(f(X_0))$ **do**
3: $\quad \| \nabla f(X_i) \|^2$
4: $\quad X_{i+1} \leftarrow X_i + r_i$
5: $\quad i \leftarrow i + 1$
6: **end while**
7: **return:** $\hat{r} = \Pi r_i$.



2

$$f(\boldsymbol{x}_0)+\nabla f(\boldsymbol{x}_0)^T(\boldsymbol{x}-\boldsymbol{x}_0)$$

1 https://arxiv.org/pdf/1511.04599.pdf
2 Illustration provided by P. Sokerin

Proposed method

Add regularization

- IFGSM with regularization:

Regularization with **discriminator models**

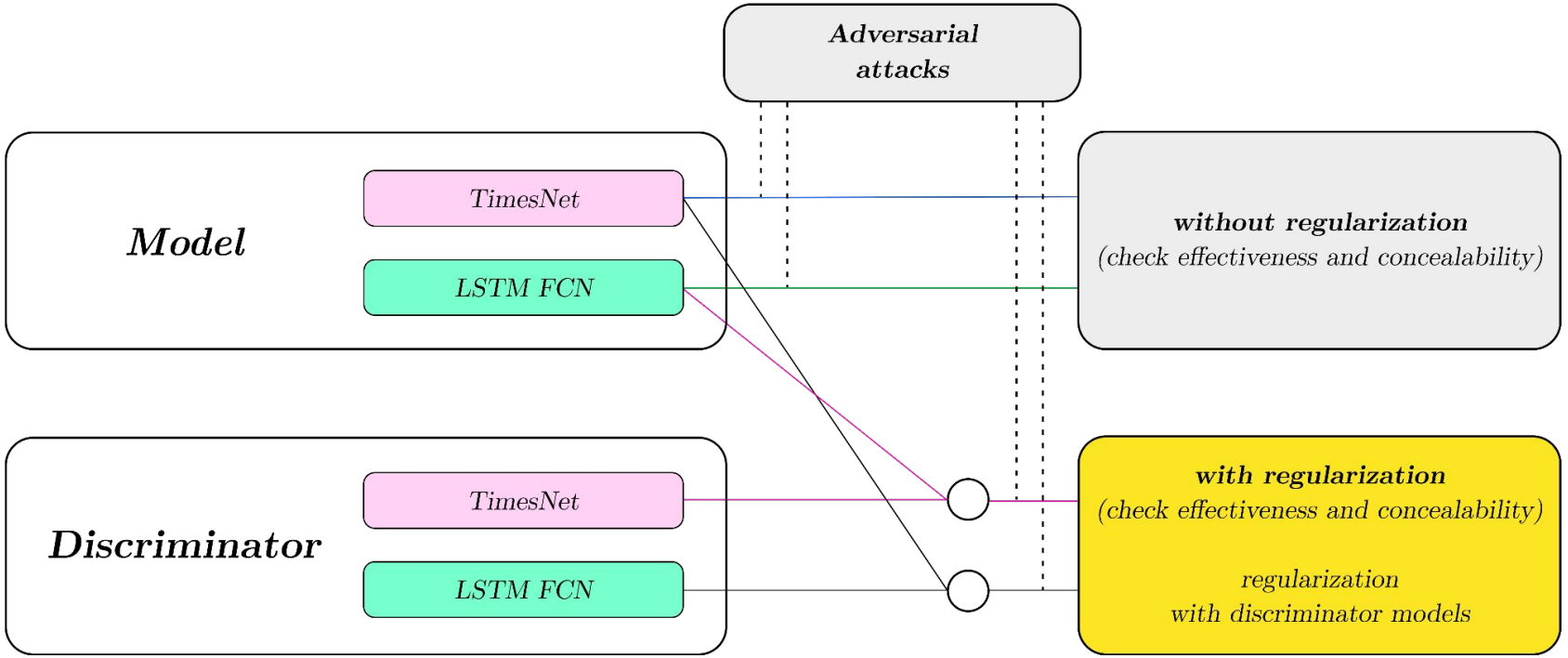
$$\boldsymbol{x}^{(t+1)} = \boldsymbol{x}^{(t)} + \epsilon \operatorname{sign}\left(\nabla_x J\left(\theta, \boldsymbol{x}^{(t)}, y\right) - \frac{\lambda}{k}\sum_{i=1}^{k} \log\left(D_i\left(\boldsymbol{x}^{(t)}\right)\right)\right)$$

Where:
λ – regularization parameter
Di - Discriminator i
k – number of discriminators

- Deepfool with regularization:

Iteration of Deepfool attack -> Step of gradient descent for discriminator loss

Pipeline


Adversarial attacks
Model
TimesNet
LSTM FCN
without regularization
(check effectiveness and concealability)
Discriminator
TimesNet
LSTM FCN
with regularization
(check effectiveness and concealability)
regularization
with discriminator models

Conducted experiments

Ford-A dataset

Task: binary classification

Background: sounds of automobiles

Class balanced: balanced

Number of dims: single dimension time-series

Number objects: 3601 train, 1320 test

Length of sequence: 500

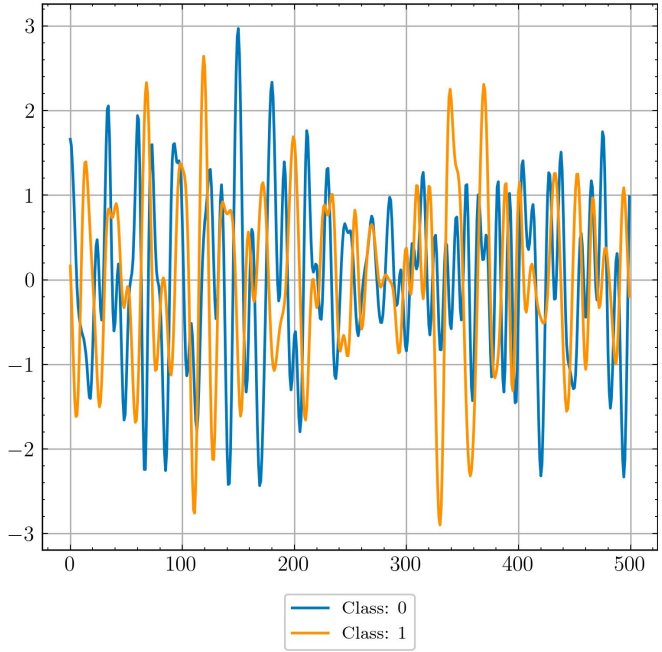
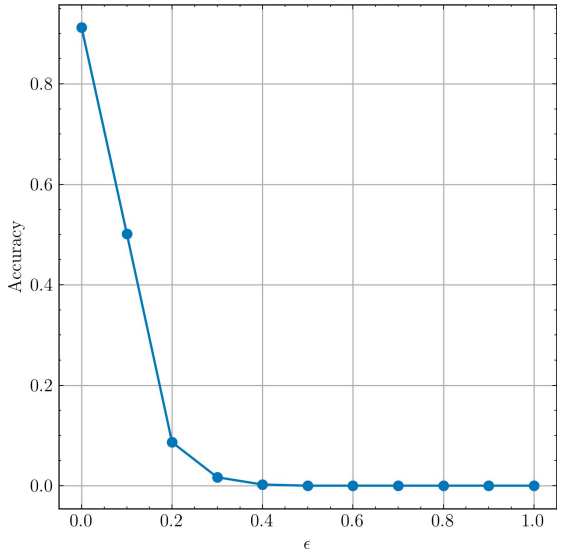


Fig 4. Sample of FordA dataset.
Orange time series - 1 class, dark blue time series - 0 class.

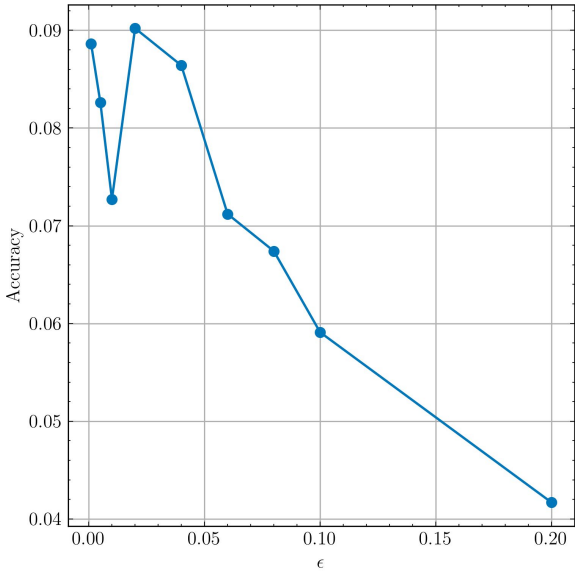
Obtained Results. Adversarial attack methods

IFGSM



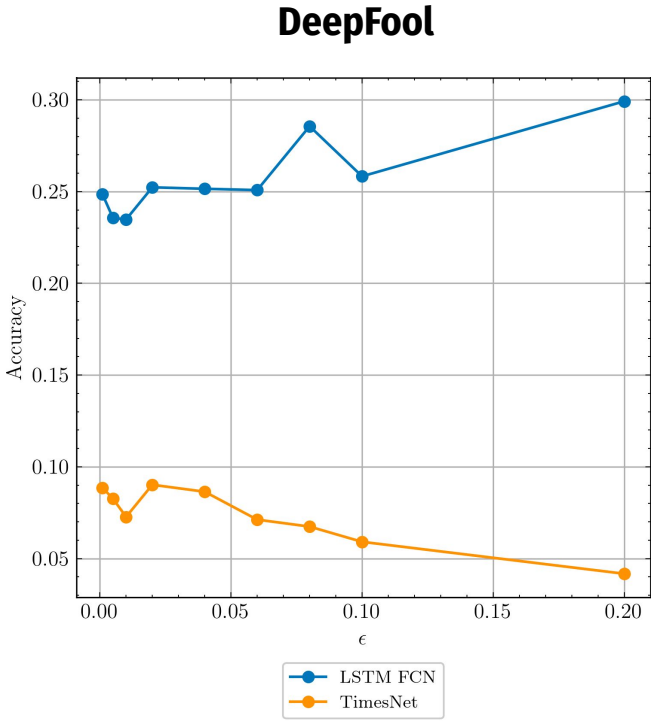
The relationship between accuracy and the epsilon parameter in TimesNet model with applied IFGSM adversarial attack. As epsilon increases, more perturbations are introduced into the data. Consequently, the quality of the model decreases due to the greater impact of these perturbations on its performance.

DeepFool



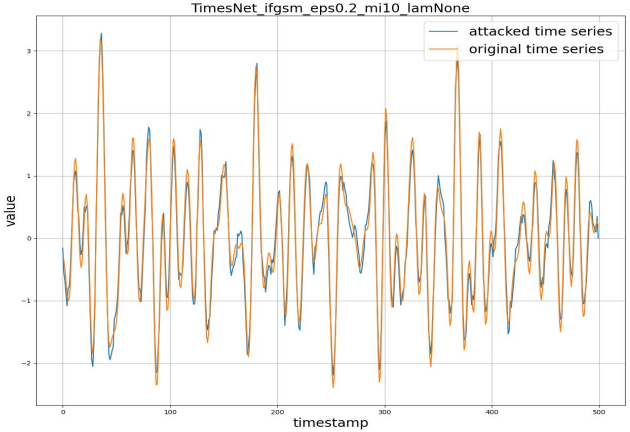
The relationship between accuracy and the epsilon parameter in TimesNet model with applied DeepFool adversarial attack. Increasing the value of epsilon in DeepFool does not necessarily make the adversarial attack more potent as in IFGSM.

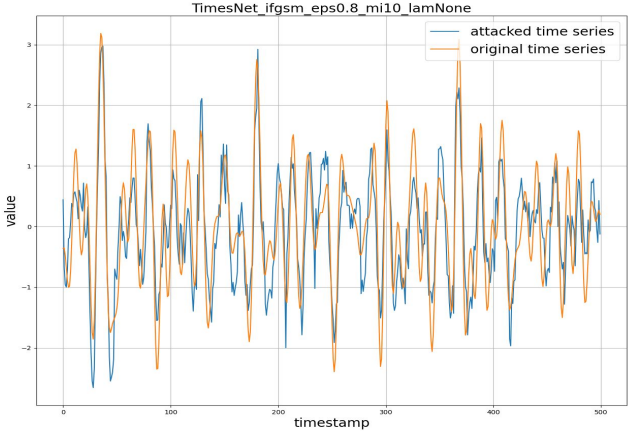
Obtained Results. Adversarial attack methods

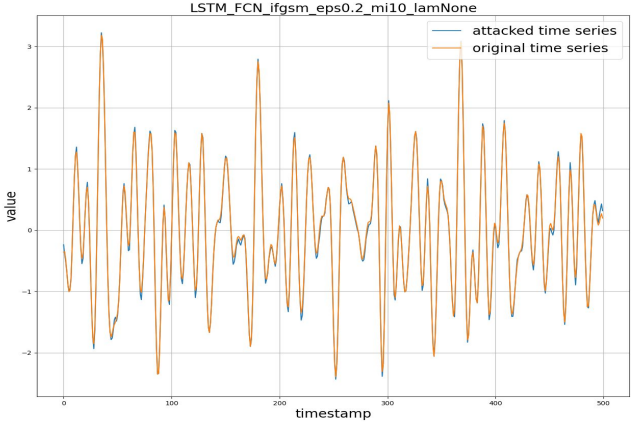


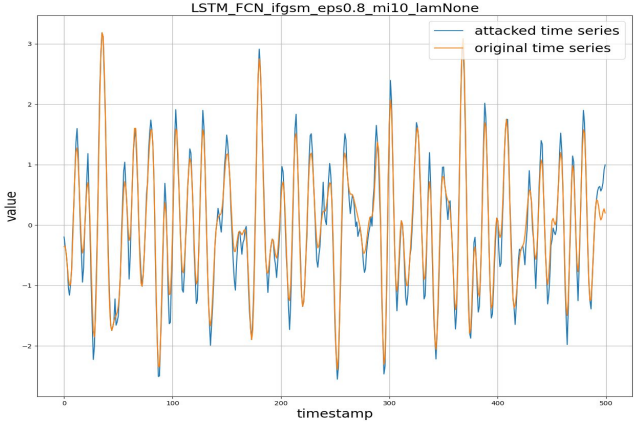
The relationship between accuracy and the epsilon parameter in TimesNet and LSTM FCN models with applied DeepFool adversarial attack.

Obtained Results. IFGSM, different epsilon

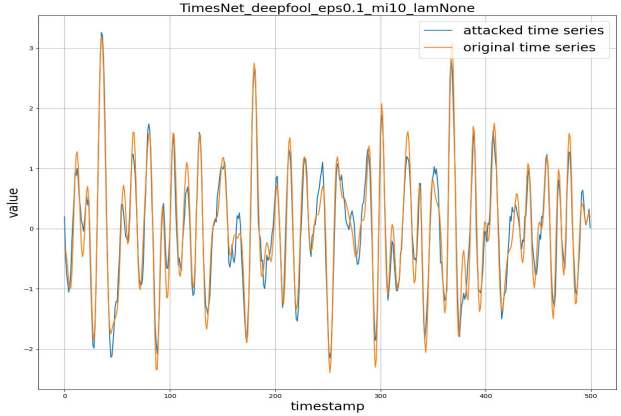


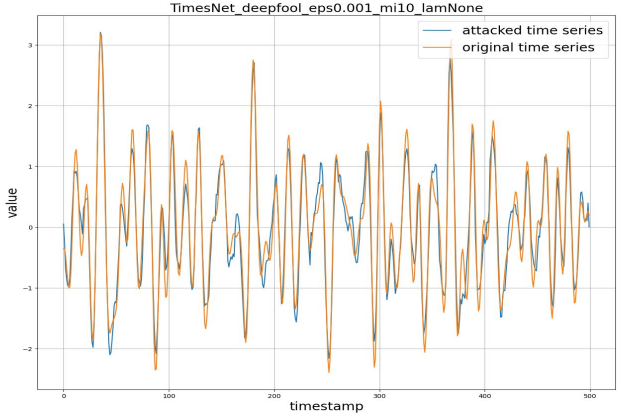


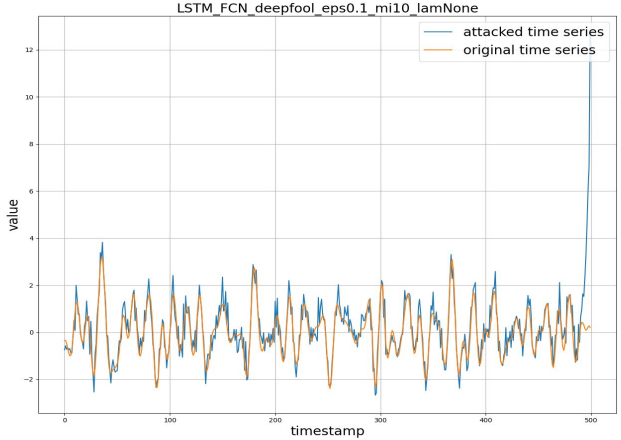


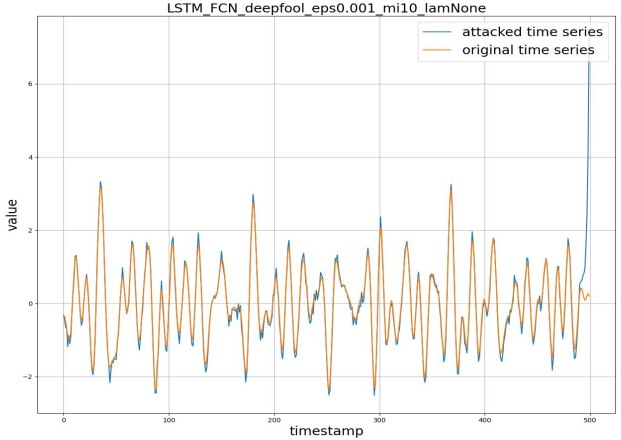


Obtained Results. DeepFool, different epsilon









Obtained Results. IFGSM, no regularization

Table 1. Series of experiments on IFGSM adversarial attack without regularization.

MODEL	DISCRIMINATOR	ϵ	MAX_ITER	CONCEALABILITY	EFFECTIVENESS
TIMESNET	TIMESNET	0.9	10	0.008	1.0000
TIMESNET	TIMESNET	0.8	10	0.0174	1.0000
TIMESNET	TIMESNET	0.7	10	0.1030	1.0000
TIMESNET	TIMESNET	0.6	10	0.4197	1.0000
TIMESNET	TIMESNET	0.5	10	0.8750	1.0000
LSTM FCN	LSTM FCN	0.9	10	0.000	0.9871
LSTM FCN	LSTM FCN	0.8	10	0.000	0.9864
LSTM FCN	LSTM FCN	0.7	10	0.0182	0.9856
LSTM FCN	LSTM FCN	0.6	10	0.1515	0.9848
LSTM FCN	LSTM FCN	0.5	10	0.3212	0.9818

Obtained Results. IFGSM, with regularization

Table 2. Series of experiments on IFGSM adversarial attack with regularization.

MODEL	DISCRIMINATOR	ϵ	MAX_ITER	λ	CONCEALABILITY	EFFECTIVENESS
TIMESNET	TIMESNET	0.9	10	0.005	0.7402	0.7409
TIMESNET	TIMESNET	0.8	10	0.005	0.8780	**0.7591**
TIMESNET	TIMESNET	0.7	10	0.005	0.9432	0.7530
TIMESNET	TIMESNET	0.6	10	0.005	0.9909	0.7455
TIMESNET	TIMESNET	0.5	10	0.005	**0.9992**	0.7182
LSTM FCN	LSTM FCN	0.9	10	0.01	0.7750	0.9061
LSTM FCN	LSTM FCN	0.8	10	0.01	0.7962	0.9061
LSTM FCN	LSTM FCN	0.7	10	0.01	0.8008	0.9197
LSTM FCN	LSTM FCN	0.6	10	0.01	0.8061	0.9288
LSTM FCN	LSTM FCN	0.5	10	0.01	**0.8447**	**0.9326**

- Proposed regularization allowed to **increase concealability** with a slight decrease in effectiveness.

Obtained Results. IFGSM, with regularization, different number of discriminators

Table 3. Series of experiments on IFGSM adversarial attack with regularization using different number of discriminators.

MODEL	DISCRIMINATOR	ϵ	NUMBER OF DISCRIMI-NATORS	λ	CONCEALABILITY	EFFECTIVENESS
TIMESNET	TIMESNET	0.9	5	0.005	0.7492	0.7553
TIMESNET	TIMESNET	0.8	5	0.005	**0.8386**	**0.7576**
TIMESNET	TIMESNET	0.9	4	0.005	0.7515	0.7492
TIMESNET	TIMESNET	0.8	4	0.005	0.8439	0.7508
TIMESNET	TIMESNET	0.9	3	0.005	0.7394	0.7409
TIMESNET	TIMESNET	0.8	3	0.005	0.8235	0.7402
TIMESNET	TIMESNET	0.9	2	0.005	0.7720	0.7583
TIMESNET	TIMESNET	0.8	2	0.005	**0.8508**	**0.7583**
TIMESNET	TIMESNET	0.9	1	0.005	0.7894	0.7568
TIMESNET	TIMESNET	0.8	1	0.005	**0.8780**	**0.7591**
LSTM FCN	LSTM FCN	0.9	3	0.01	0.7750	0.9061
LSTM FCN	LSTM FCN	0.9	2	0.03	**0.8841**	**0.9227**
LSTM FCN	LSTM FCN	0.9	1	0.05	0.8417	0.8894

Obtained Results. DeepFool, no regularization

Table 4. Series of experiments on DeepFool adversarial attack without regularization.

MODEL	DISCRIMINATOR	ϵ	MAX_ITER	CONCEALABILITY	EFFECTIVENESS
TIMESNET	TIMESNET	0.001	10	0.112	0.9114
TIMESNET	TIMESNET	0.005	10	0.205	0.9174
TIMESNET	TIMESNET	0.01	10	0.172	0.9273
TIMESNET	TIMESNET	0.02	10	0.210	0.9098
TIMESNET	TIMESNET	0.04	10	0.263	0.9136
LSTM FCN	LSTM FCN	0.001	10	0.1424	0.7515
LSTM FCN	LSTM FCN	0.005	10	0.1523	0.7644
LSTM FCN	LSTM FCN	0.01	10	0.1598	0.7652
LSTM FCN	LSTM FCN	0.02	10	0.1341	0.7477
LSTM FCN	LSTM FCN	0.04	10	0.1371	0.7485

Obtained Results. DeepFool, with regularization

Table 5. Series of experiments on DeepFool adversarial attack with regularization.

MODEL	DISCRIMINATOR	ϵ	MAX_ITER	λ	CONCEALABILITY	EFFECTIVENESS
TIMESNET	TIMESNET	0.001	10	0.005	0.6349	**0.7342**
TIMESNET	TIMESNET	0.005	10	0.005	0.6389	0.6293
TIMESNET	TIMESNET	0.01	10	0.005	**0.8255**	0.6365
TIMESNET	TIMESNET	0.02	10	0.005	0.6010	0.5237
TIMESNET	TIMESNET	0.04	10	0.005	0.5290	0.5382
LSTM FCN	LSTM FCN	0.001	10	0.01	0.5382	0.4299
LSTM FCN	LSTM FCN	0.005	10	0.01	**0.6200**	0.3726
LSTM FCN	LSTM FCN	0.01	10	0.01	0.5038	0.3845
LSTM FCN	LSTM FCN	0.02	10	0.01	0.4638	**0.4583**
LSTM FCN	LSTM FCN	0.04	10	0.01	0.4227	0.3673

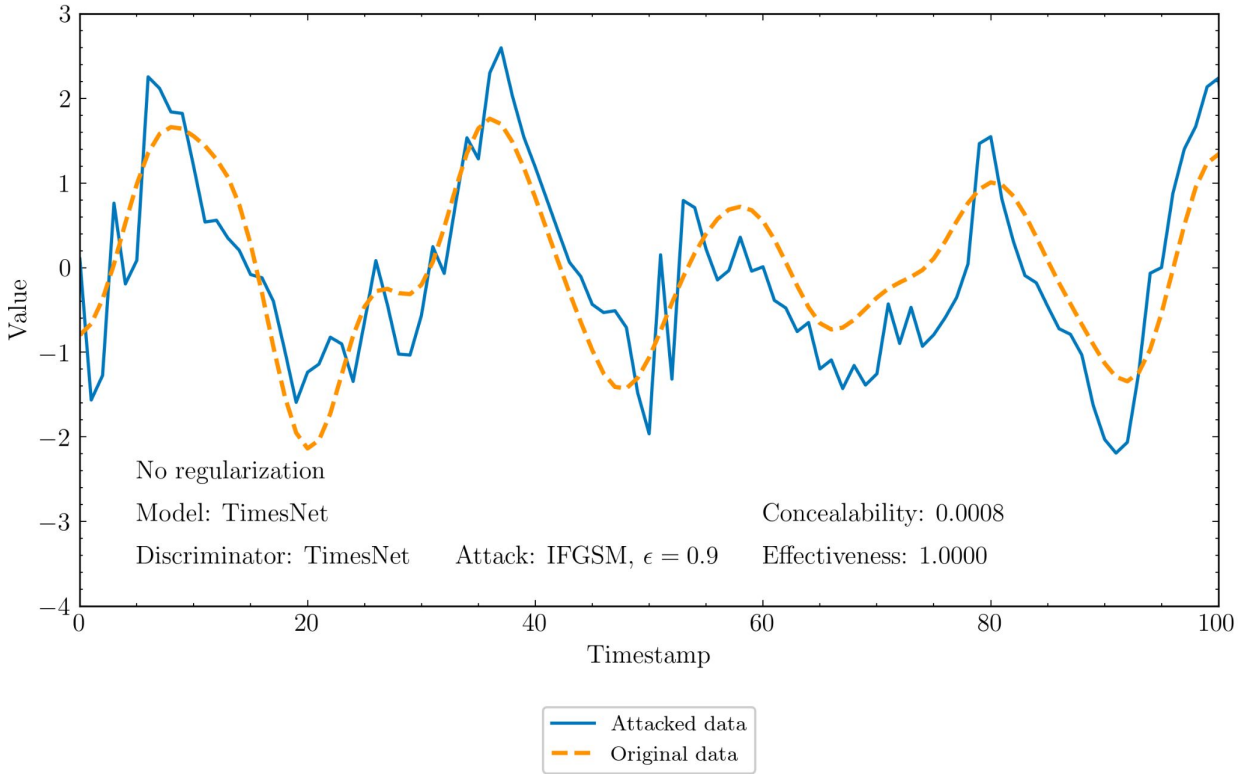
Obtained Results. DeepFool, with regularization, different number of discriminators

Table 6. Series of experiments on DeepFool adversarial attack with regularization using different number of discriminators.

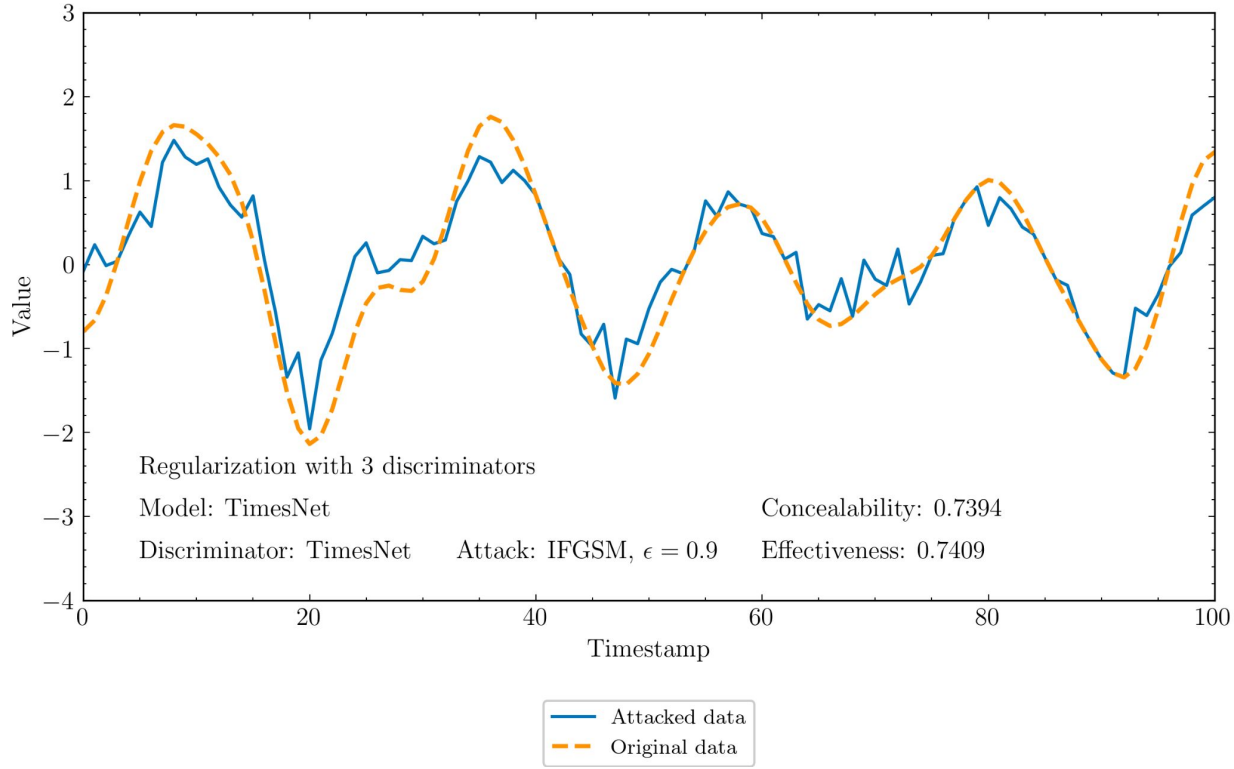
MODEL	DISCRIMINATOR	ϵ	NUMBER OF DISCRIMI-NATORS	λ	CONCEALABILITY	EFFECTIVENESS
TIMESNET	TIMESNET	0.01	5	0.005	**0.6273**	0.4280
TIMESNET	TIMESNET	0.01	4	0.005	0.5923	0.4733
TIMESNET	TIMESNET	0.01	3	0.005	0.5977	**0.4956**
LSTM FCN	LSTM FCN	0.01	5	0.01	**0.6378**	**0.4037**
LSTM FCN	LSTM FCN	0.01	4	0.01	0.6239	0.3929
LSTM FCN	LSTM FCN	0.01	3	0.01	0.6328	0.3446

Obtained Results. Comparison of data samples

TimesNet - TimesNet without regularization

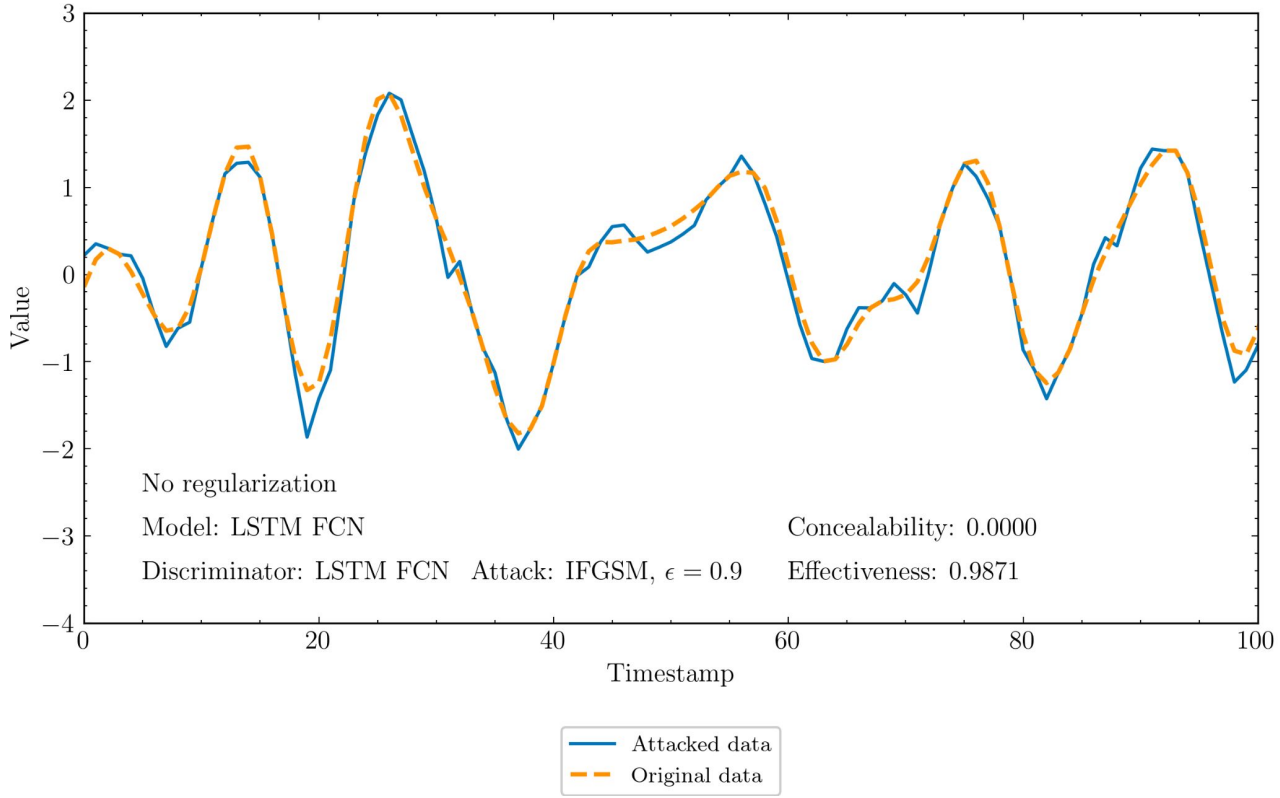


Obtained Results. Comparison of data samples
TimesNet - TimesNet with regularization



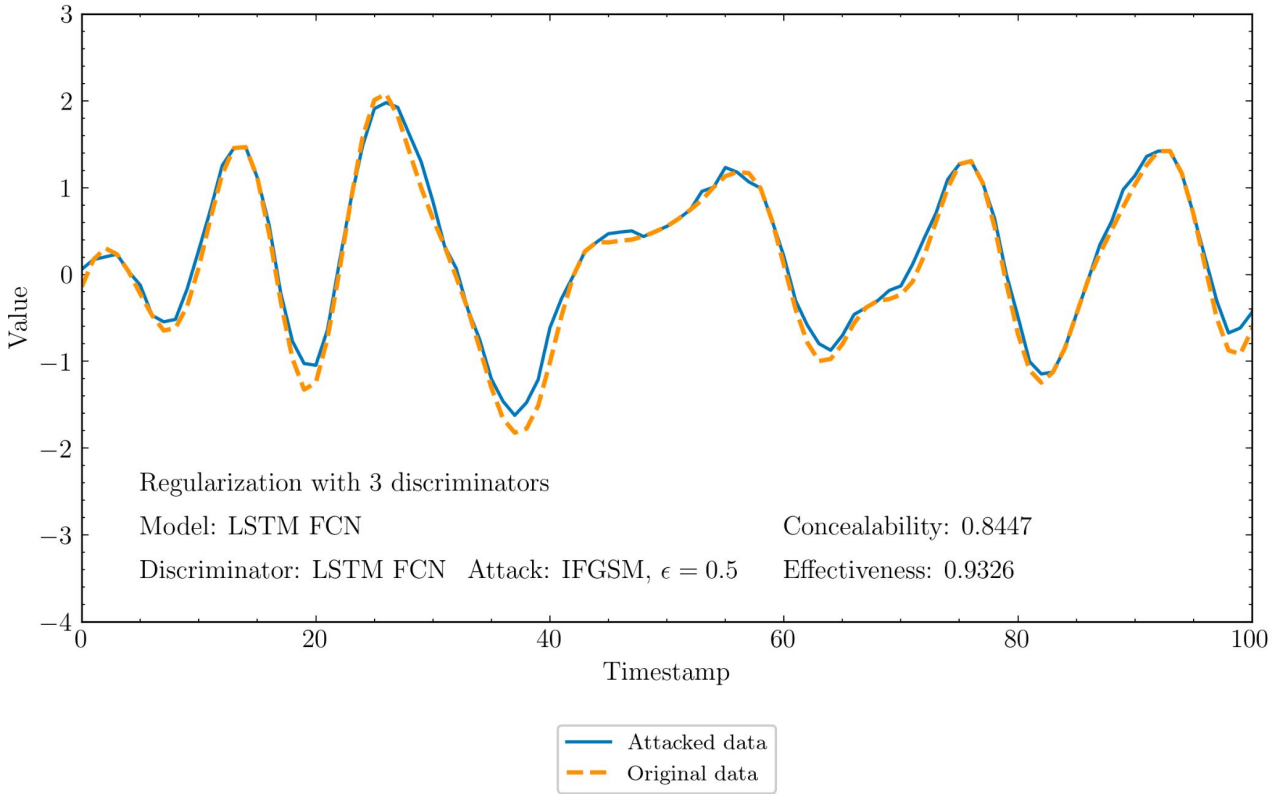
Obtained Results. Comparison of data samples

LSTM FCN - LSTM FCN without regularization



Obtained Results. Comparison of data samples

LSTM FCN - LSTM FCN with regularization



Conclusion

- A method of adversarial attack on time series classification was developed,
- Adversarial attack on time series models with regularization improved concealability with remaining high effectiveness,
- Deepfool performs better with TimesNet architecture than LSTM-FCN (with and without regularization),
- IFGSM performs better on TimesNet without regularization and on LSTM-FCN with regularization,
- Extensive simulations on real-world dataset confirm effectiveness of proposed method,
- To achieve good quality extensive search of hyperparameters is required

Contributions

- **Alina Bogdanova**
  - Implemented DeepFool on Pytorch
  - Implemented discriminator regularization procedure for DeepFool on Pytorch
  - Prepared the Final Report
  - Prepared the presentation
  - Prepared GitHub repository

- **Nikita Ligostaev**
  - Implemented FGSM, IFGSM on PyTorch
  - Prepared the Final Report
  - Prepared the presentation
  - Prepared GitHub repository

- **Anastasia Sozykina**

  - Conducting experiments with TimeNet and LSTM-FCN models
  -Preparing the Final Report
  - Preparing the presentation
  - Preparing GitHub repository

- **Matvey Skripkin**
  - Implemented pipeline for experiments on PyTorch
  - Implemented LSTM FCN on PyTorch
  - Implemented discriminator regularization procedure for IFGSM on Pytorch
  - Preparing GitHub repository