PROJECT PROPOSAL PRESENTATION

Handling Adversarial Attacks on Deep Neural Network through Change Point Detection Application to Smart Home Time Series Data

---

CSE 419 Section -02 Fall 2019

Team Members

Name	ID
Mahbuba Tasmin	1610064042
Sharif Uddin Ruman	1611557642
Sifat Jahan	1611702642
N.M. Shihab Islam	1420339042
Abdur Raufus Saleheen	1610472642
Taoseef Ishtiak	1612142042
Md. Sajid Ahmed	1610364042
Arif Suhan	1610437042
Shahnawaz Zulminan	1611708042

Adversarial Attacks on Deep Neural Networks for Time Series Classification

Authors - Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar and Pierre-Alain Muller

-IRIMAS, University Haute-Alsace, Mulhouse, France

Publication Journal - IEEE International Joint Conference on Neural Networks - (IJCNN) 2019{7.426}

Dataset Source:
https://www.cs.ucr.edu/~eamonn/time_series_data/UCR_TS_Archive_2015.zip

Dataset Name - **UCR_TS_Archive_2015**

Abstract

- Abstract— Time Series Classification (TSC) problems are contained in many real-life data mining activities ranging from medication and health to identification of human activity and food safety.
- They propose to leverage existing adversarial attack mechanisms to add a special noise to the input time series in order to decrease the network's confidence when classifying instances at test time.

Introduction

- An adversarial attack consists in modifying an original image so that the changes are almost undetectable by a human. The modified image is called an adversarial image, which will be misclassified by the neural network, while the original one is correctly classified.
- Another application is the alteration of illegal content to make it undetectable by automatic moderation algorithms.
- Transfer and adapt adversarial attacks that have been shown to work well on images to time series data.

Example

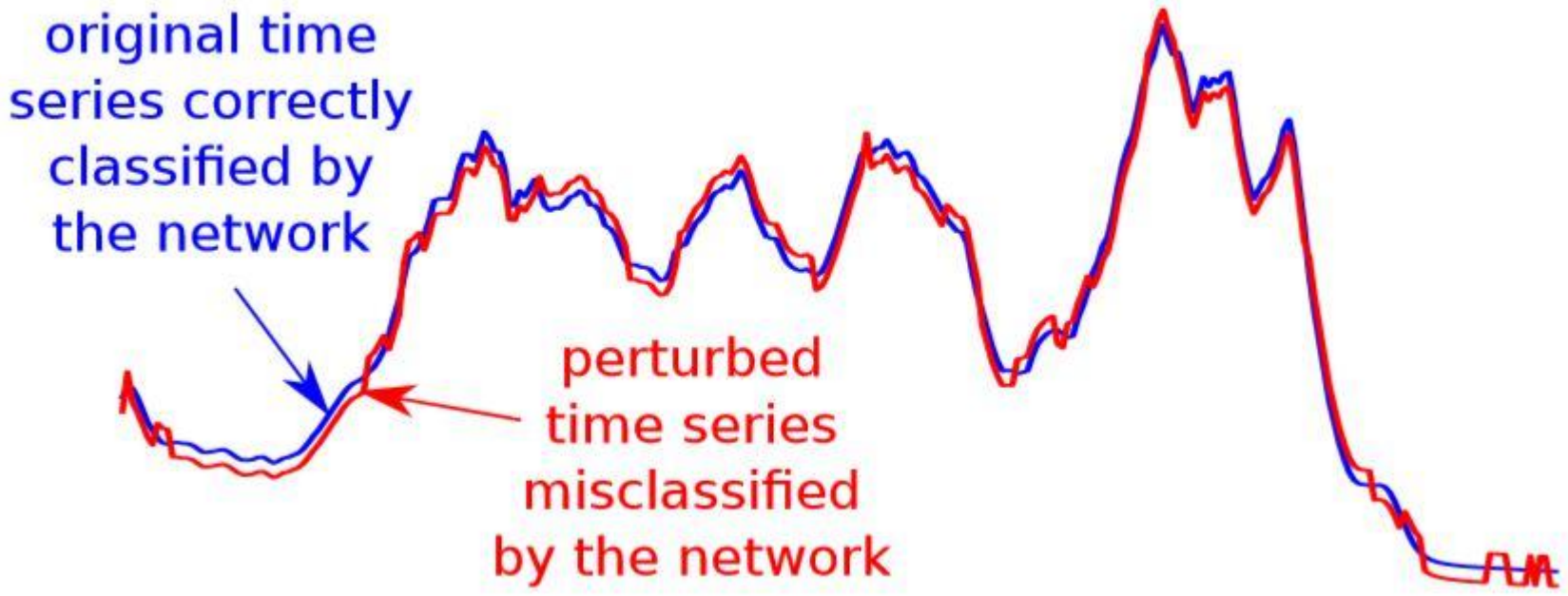


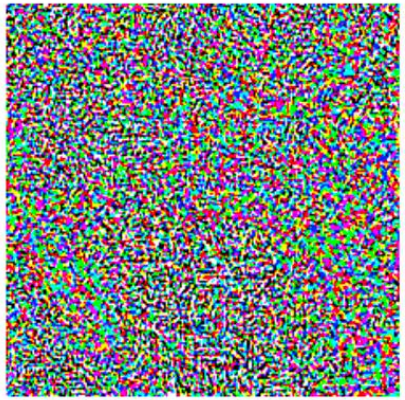
Fig: Example of a perturbed time series that is misclassified by a deep network. (Coffee)

Example

“panda”

57.7% confidence

$+ .007 \times$



noise

$=$

“gibbon”

99.3% confidence

Methodology

Residual Network (ResNet) was used for adversarial attacks for **time series.**

Two attack method was used to generate adversarial time series.

1. **Fast Gradient Sign Method (FGSM)**

1. **Basic Iterative Method (BIM)**

Residual Network

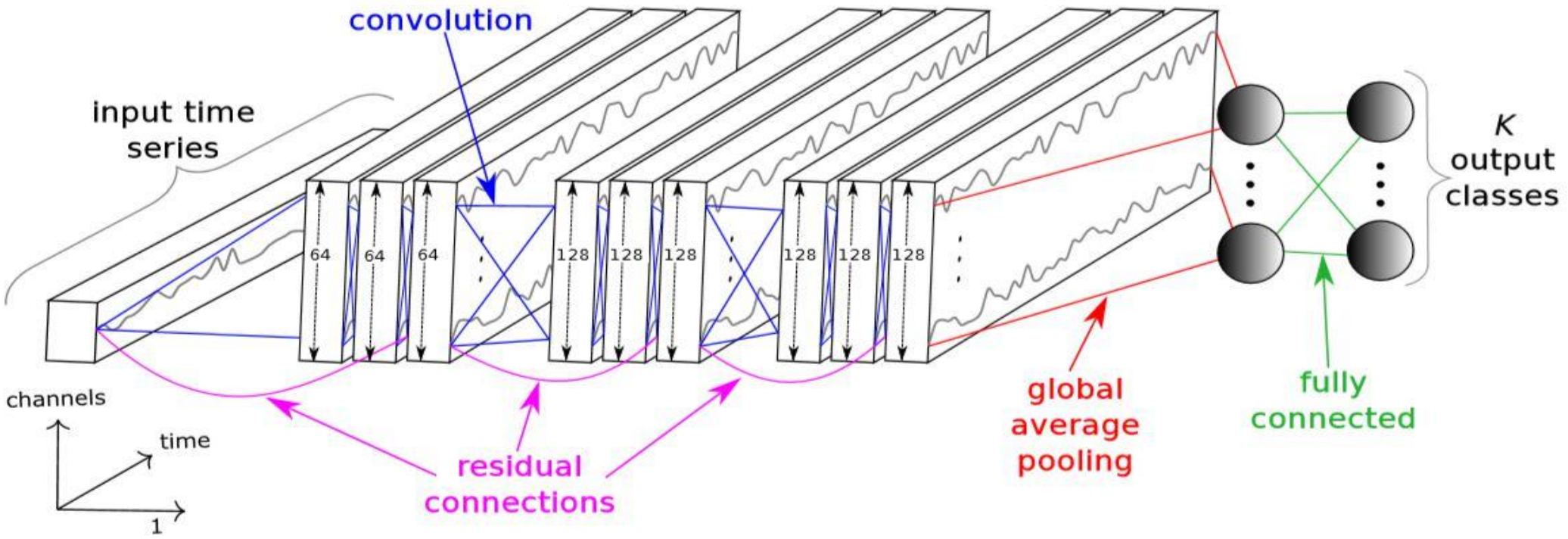


Fig: The deep Residual Network architecture for Time Series Classification.

Fast Gradient Sign Method (FGSM)

This adversarial method fooled GoogLeNet model.

This attack is based on a one step gradient update along the direction of the gradient's sign at each time stamp.

$$\eta = \epsilon \cdot sign(\nabla_x J(X, \hat{Y}))$$

Approach with FGSM

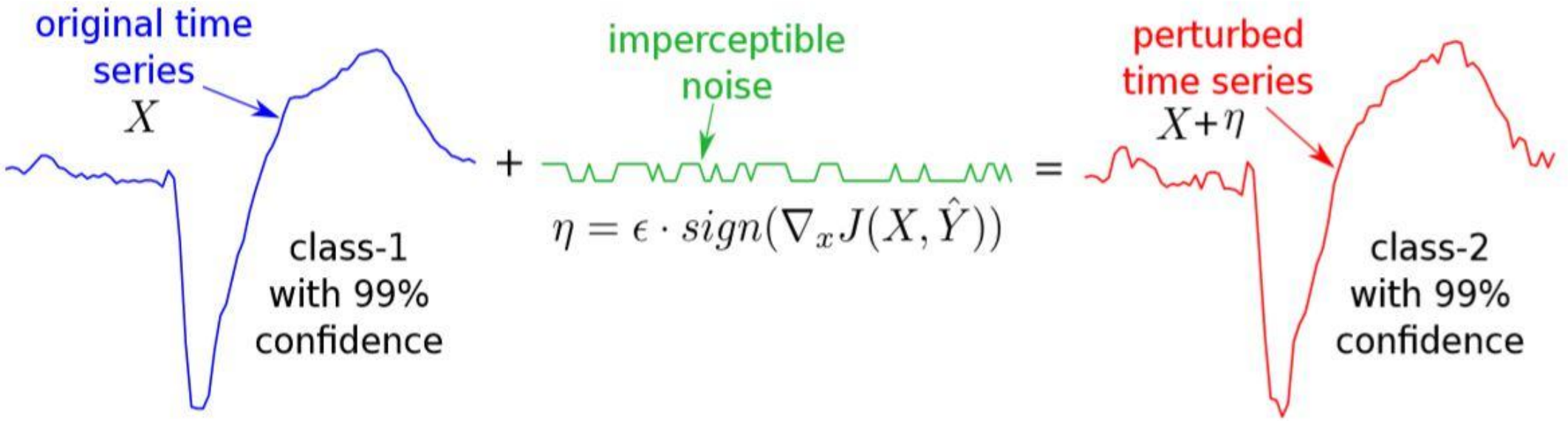


Fig : Example of perturbing the classification of an input time series from a dataset (TwoLeadECG) by adding an imperceptible noise computed using the Fast Gradient Sign Method (FGSM).

Basic Iterative Method

Basic Iterative Methods (BIM) extends FGSM by applying it multiple times with a small step size.

By adding smaller changes in an iterative meaner generates adversarial time series.

Generated series examples was closer to original time series.

Iterative Adversarial Attack

Parameter: I, ϵ, α

Input: original time series X & its label $\hat{Y}$

Output: perturbed time series $X^{'}$

$X^{'} \leftarrow X$

for $i = 1$ to I **do**

$\eta = \alpha \cdot sign(\nabla_x J(X^{'}, \hat{Y}))$

$X^{'} = X^{'} + \eta$

$X^{'} = min\{X + \epsilon, max\{X - \epsilon, X^{'}\}\}$

end for

UCR Archive

Total 85 different Dataset on Time Series Classification.

1. Ham
2. Coffee
3. FordA
4. SmallKitchenAppliances

   ……..……..

   ……..

84. LargeKitchenAppliances
85. ItalyPowerDemand

Result and Finding

Result on whole UCR archive:

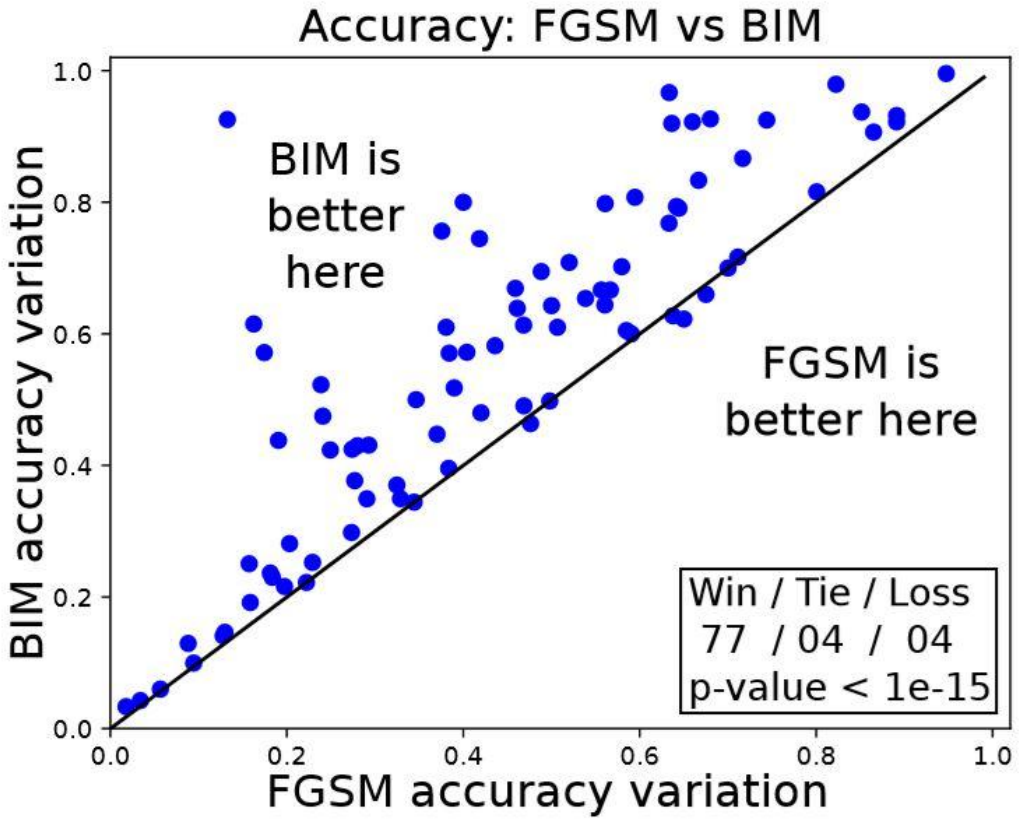


Fig: Accuracy variation of two attack.

Attacks on food quality and safety(HAM)

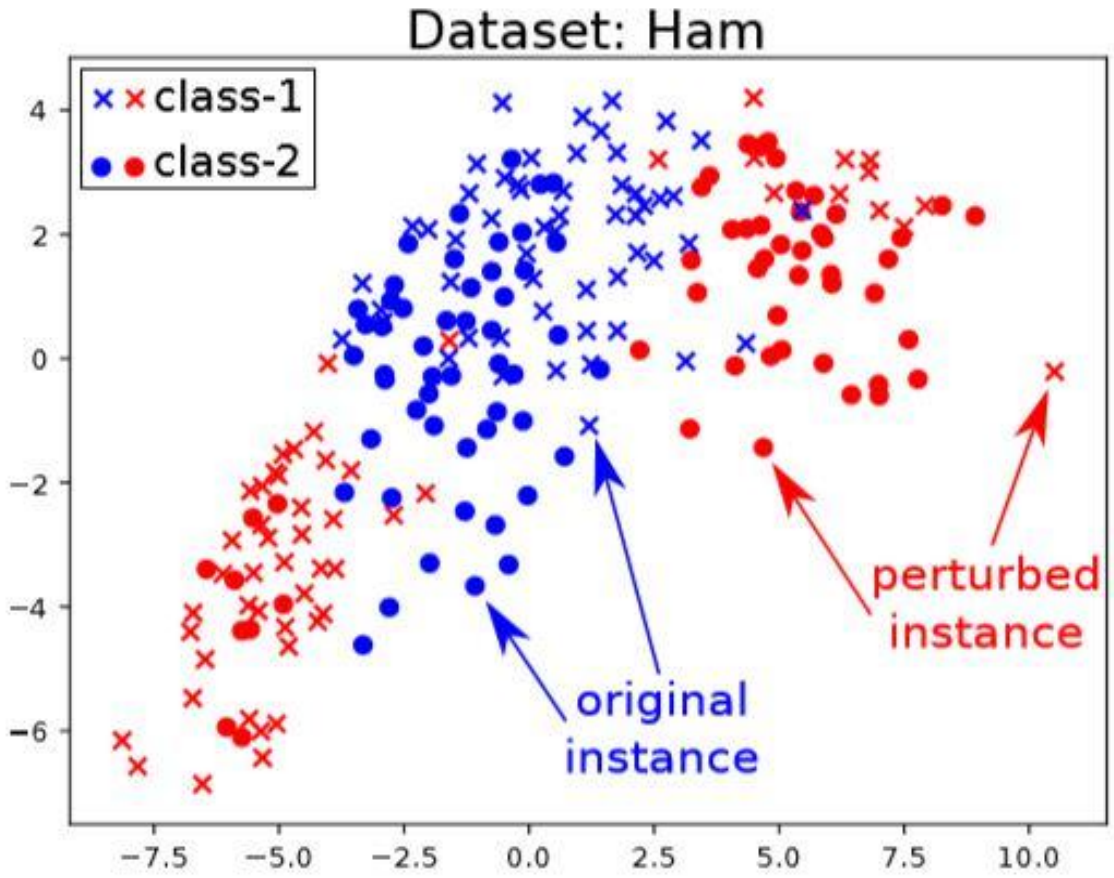


Fig: Multidimensional Scaling (MDS) showing the distribution of perturbed time series

Attacks on food quality and safety(Coffee)

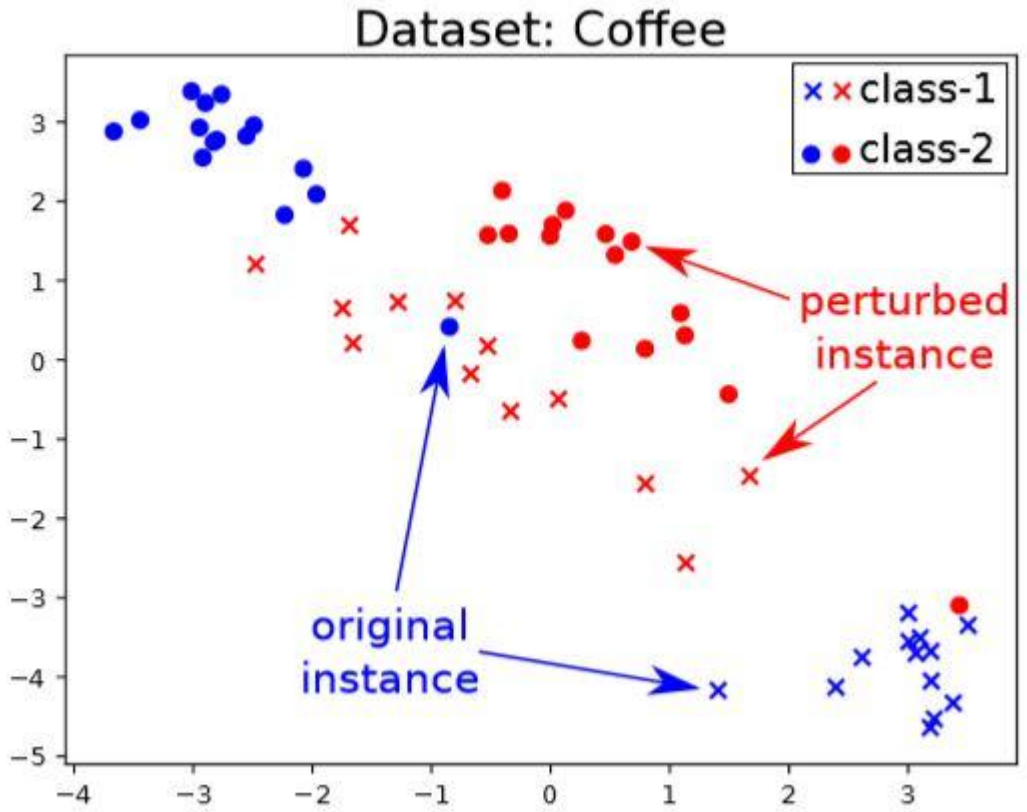


Fig: Multidimensional Scaling (MDS) showing the distribution of perturbed time series

Attack on vehicle sensors(FordA)

Attack	Accuracy
ResNet_Original	91.8
ResNet_FGSM	33.9
ResNet_BIM	21.6

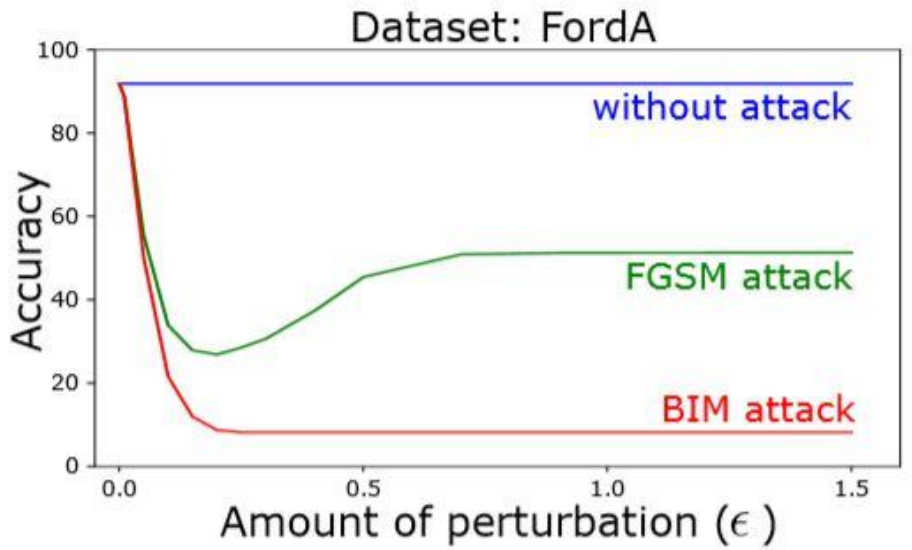


Fig : FGSM and BIM attack on FordA

Attack on electricity consumption (ItalyPowerDemand)

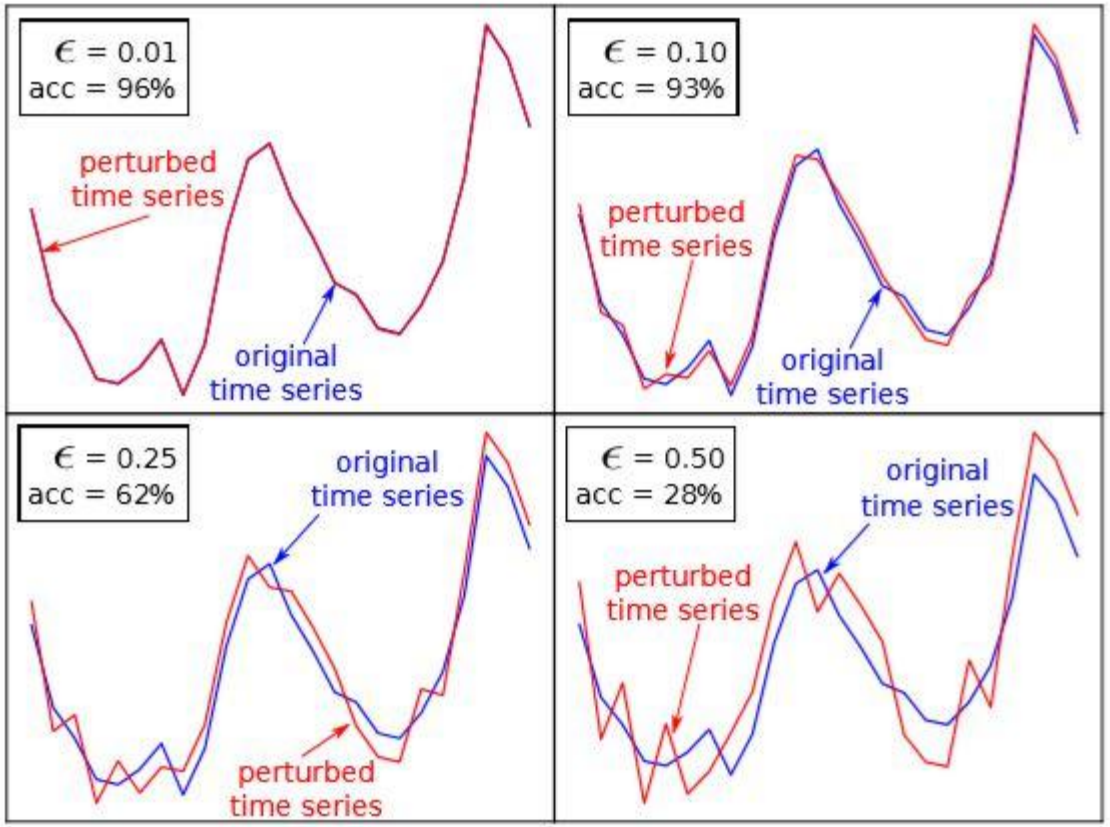


Fig: Accuracy variation for ItalyPowerDemand

Evaluation against FCN

Dataset	ResNet Original	ResNet FGSM	ResNet BIM	FCN Original	FCN FGSM	FCN BIM
Ham	80.0	21.0	**20.0**	71.4	**27.6**	**27.6**
Coffee	100.0	50.0	**35.7**	100.0	75.0	**64.3**
FordA	91.8	33.9	**21.6**	90.1	59.6	**57.3**
SmallKitchen Appliances	78.9	40.5	**21.9**	78.7	47.5	**28.8**
LargeKitchen Appliances	90.4	74.7	**65.3**	89.6	66.4	**63.5**
yoga	87.2	45.4	**12.8**	84.1	44.9	**19.2**
ItalyPowerDemand	95.9	92.5	**91.6**	96.1	89.8	**89.6**

Fig: Evaluation of ResNet against FCN (7/85 Dataset)

with application to Smart Home Time Series Data

Authors - Samaneh Aminikhanghahi, Tinghui Wang, and Diane J. Cook, *IEEE Fellow*

Publication Journal - **IEEE Transactions on Knowledge and Data Engineering (3.865)**

Dataset Source- CASAS (Center of Advanced Studies in Adaptive System)

Dataset Name - Human Activity Recognition from Continuous Ambient Sensor Data Dataset

Abstract

Change Point Detection (CPD) is the problem of discovering time points at which the behavior of a time series changes abruptly.

A novel real-time nonparametric change point detection algorithm called *SEP* (Separation distance as a divergence measure to detect change points in high-dimensional time series).

Index Terms—Activity transition detection, change detection algorithms, Separation distance, smart homes, time series data

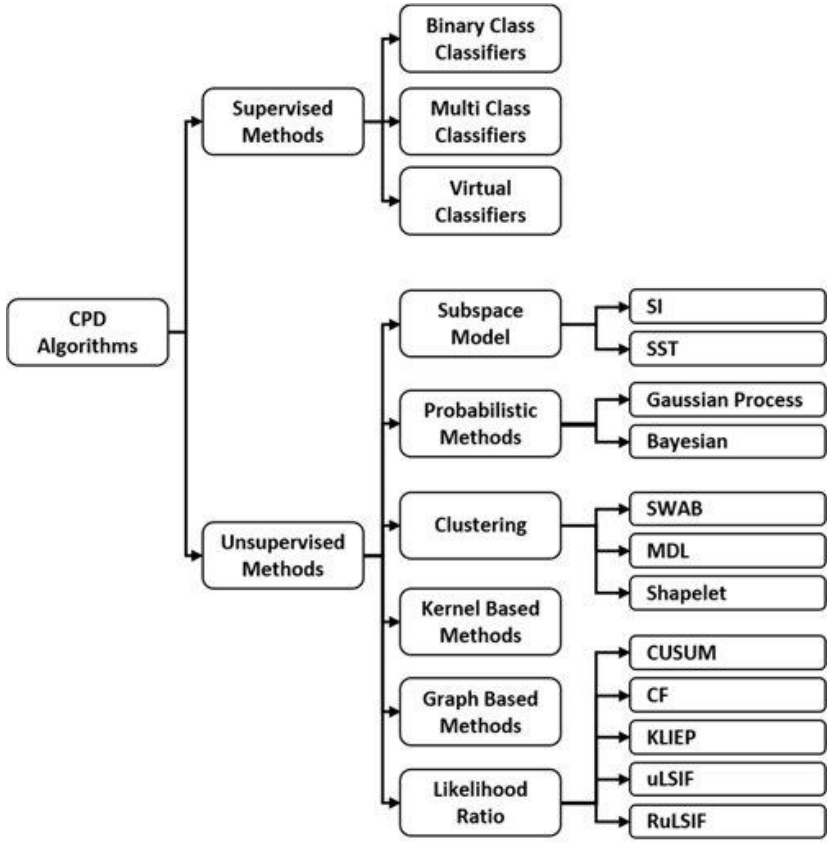
Introduction

Direct density ratio change point detection algorithms

Detect change points between two consecutive windows of data by estimating their probability density ratio based on the assumption that the probability density of two consecutive windows are same if they belong to the same state.

Introduction of new CPD method built on the notion of SEParation distance

Overview of CPD Algorithms



Methodology: SEP Change Point Detection

Two probability densities, $f_t(x)$ and $f_{t-1}(x)$, corresponding to two consecutive windows, each with length n, density ratio-based CPD methods use "*dissimilarity measures*" as a measure of difference between them to determine whether or not there exists a change point between these two windo

$$g_t(x) = \frac{f_{t-1}(x)}{f_t(x)} = \sum_{i=1}^{n} \theta_i \prod_{j=1}^{n} K\left(x_t^i, x_{t-1}^j\right)$$

$$K(x_1, x_2) = exp\left(-\frac{\|x_1 - x_2\|^2}{2\sigma^2}\right)$$

Parameters Description

$\Theta=(\Theta_1,\ldots,\Theta_n)^T$ represents the set of parameters for the ratio function to be learned from existing data points in the current windows, and $\sigma>0$ represents the kernel parameter.

In the training phase, the parameters Θ are determined for each window so that a chosen dissimilarity measure is minimized.

Given a density-ratio estimator $g_t(x)$, a dissimilarity measure between windows is calculated during the test phase as a change point score.

Kullback-Leibler Importance Estimation Procedure (KLIEP)

One of the first direct density ratio CPD methods, the Kullback-Leibler importance estimation procedure (KLIEP) [35], estimates the density ratio using KullbackLeibler (KL) divergence. KL divergence, defined in Equation 5, is a popular choice for the dissimilarity measure

$$KL = -\int f_t(x) log \frac{f_{t-1}(x)}{f_t(x)} dx$$

Convex optimization problem

Unique global optimal solution θ can be obtained by a gradient projection method.

$$\widehat{KL} = \frac{1}{n}\sum_{i=1}^{n} log \hat{g}(x_i)$$

Unconstrained Least-Squares Importance Fitting (uLSIF)

Pearson (PE) divergence as a similarity measure

$$PE = \int f_t(x) \left(\frac{f_{t-1}(x)}{f_t(x)} - 1 \right)^2 dx$$

The density-ratio model is fitted to the true density ratio under the squared loss.

$$\widehat{PE} = -\frac{1}{2n} \sum_{i=1}^{n} \hat{g}(x_i)^2 + \frac{1}{n} \sum_{i=1}^{n} \hat{g}(x_i) - \frac{1}{2}$$

Relative uLSIF (RuLSIF)

The problem of unbounded value of density-ratio balue

α - relative PE divergence for $0 \leq \alpha < 1$ is used as Dissimilarity measure

$$PE_\alpha = PE(f_{t-1}(\mathrm{x}), \alpha f_{t-1}(\mathrm{x}) + (1-\alpha) f_t(\mathrm{x}))$$

Result and Findings

In this section, we evaluate our proposed SEP change point detection and compare results with other popular CPD methods

1. Performance Measures:

G-mean : A supervised learning algorithm that attempts to perform change point detection.

Detection Delay. This directly measures how close the time value of each correctly-predicted CP

is to the actual CP time value in the series.

1. Artificial Dataset
   - They have used the three artificial time-series datasets that contain manually inserted change points to show the effectiveness of SEP method in detecting different changes and compare the performance to existing similar method
   - Dataset 1(Jumping mean)
   - Dataset 2 (Scaling variance).
   - Dataset 3 (Changing frequency)

Result and Findings

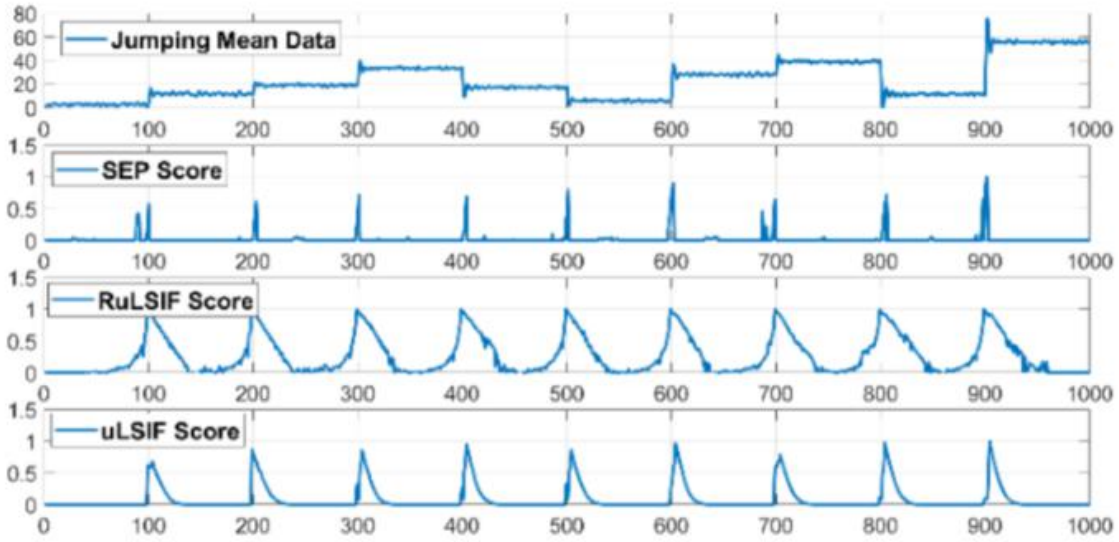


Fig. 7. Jumping mean time-series samples and the change-point scor obtained by different methods.

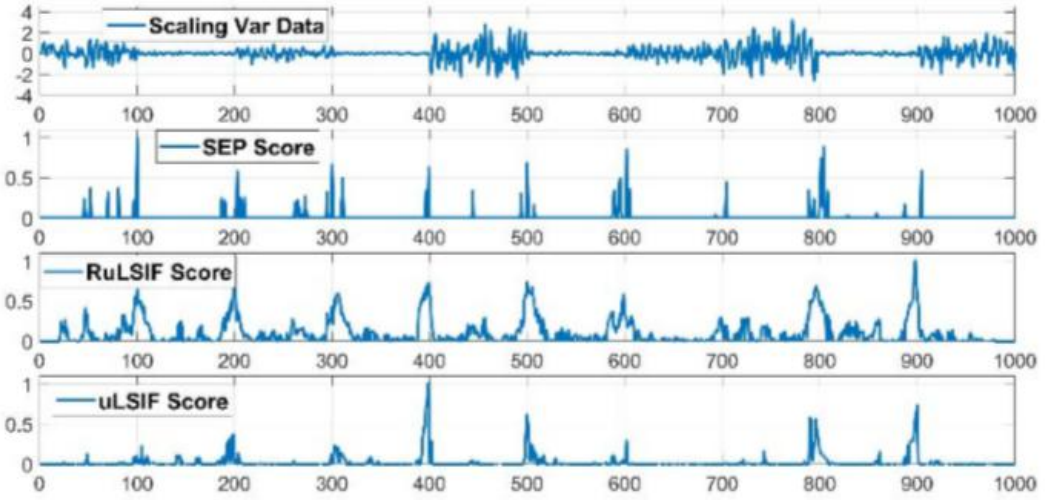


Fig. 8. Scaling variance time-series samples and the change-point score obtained by different methods.

Dataset 1(Jumping mean). The following 1-dimensional auto-regressive model is used to generate 1000 samples:

$$y(t) = 0.6y(t-1) - 0.5y(t-2) + \epsilon_t \quad (25)$$

Dataset 2 (Scaling variance). The same auto-regressive model as Dataset 1 is used, but a change point is inserted at every 100 time steps by infusing origin-centered noise with a random standard deviation between 0.01 and 1.

Result and Findings

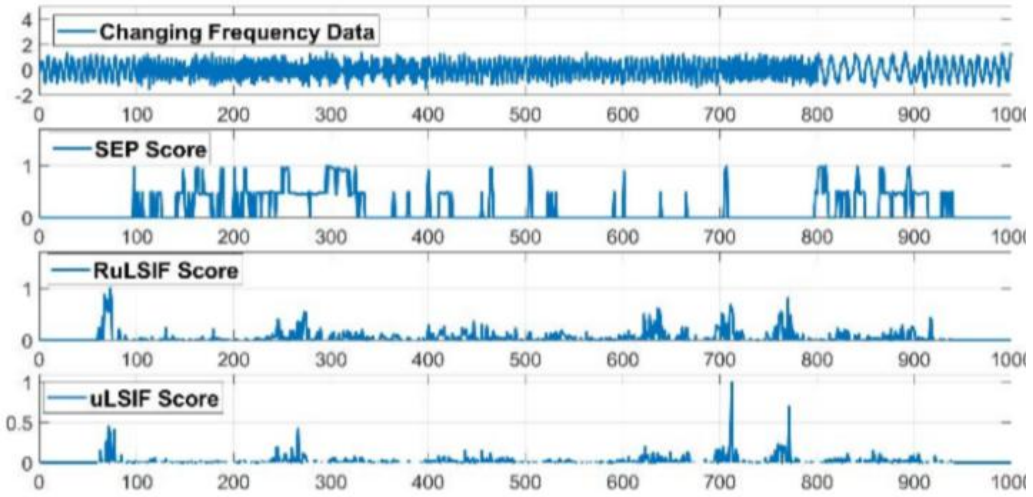


Fig. 9. Changing frequency time-series samples and the change-po score obtained by different methods.

The results show, in the case of the Jumping Mean dataset, the SEP, RuLSIF, and t-test methods successfully detect all change points although SEP and RuLSIF reduce the false alarms in comparison to a basic t-test.

PERFORMANCE OF CPD ALGORITHM FOR ARTIFICIAL DATASETS.

	SEP	RuLSIF	uLSIF	T test
Dataset 1 – Jumping Mean				
TPR	1.00	1.00	0.56	1.00
FPR	0.03	0.05	0.01	0.14
G-mean	0.99	0.98	0.74	0.93
Dataset 2 – Scaling Variance				
TPR	1.00	1.00	1.00	0.11
FPR	0.14	0.14	0.15	0.01
G-mean	0.93	0.93	0.92	0.33
Dataset 3 – Changing Frequency				
TPR	1.00	1.00	0.44	0.22
FPR	0.13	0.25	0.01	0.06
G-mean	0.93	0.87	0.66	0.46

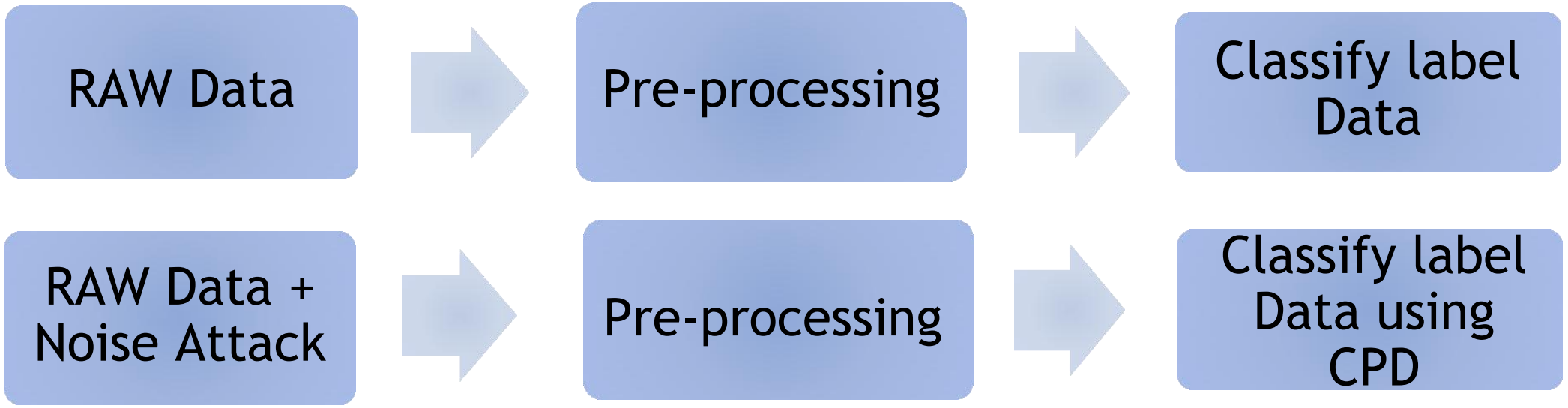
Project Outline

Title-: Handling Adversarial Attacks on Deep Neural Network through Change Point Detection Application to Smart Home Time Series Data

Key Features –

1. Change Point Detection Application
2. Handling Adversarial Attacks
3. Classification of Activities from Smart Home Time Series Data

Work Flow


RAW Data
Pre-processing
Classify label Data
RAW Data + Noise Attack
Pre-processing
Classify label Data using CPD