Adversarial Examples in Deep Learning for Multivariate Time Series Regression

Gautam Raj Mode and Khaza Anuarul Hoque
Department of Electrical Engineering & Computer Science
University of Missouri, Columbia, MO, USA
gmwyec@mail.missouri.edu, hoquek@missouri.edu

Abstract—Multivariate time series (MTS) regression tasks are common in many real-world data mining applications including finance, cybersecurity, energy, healthcare, prognostics, and many others. Due to the tremendous success of deep learning (DL) algorithms in various domains including image recognition and computer vision, researchers started adopting these techniques for solving MTS data mining problems, many of which are targeted for safety-critical and cost-critical applications. Unfortunately, DL algorithms are known for their susceptibility to adversarial examples which also makes the DL regression models for MTS forecasting also vulnerable to those attacks. To the best of our knowledge, no previous work has explored the vulnerability of DL MTS regression models to adversarial time series examples, which is an important step, specifically when the forecasting from such models is used in safety-critical and cost-critical applications. In this work, we leverage existing adversarial attack generation techniques from the image classification domain and craft adversarial multivariate time series examples for three state-of-the-art deep learning regression models, specifically Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU). We evaluate our study using Google stock and household power consumption dataset. The obtained results show that all the evaluated DL regression models are vulnerable to adversarial attacks, transferable, and thus can lead to catastrophic consequences in safety-critical and cost-critical domains, such as energy and finance.

Index Terms—Multivariate time series, Regression, Deep learning, Adversarial examples, FGSM, BIM.

I. INTRODUCTION

Time series forecasting is an important problem in data mining with many real-world applications including finance [1]–[4], weather forecasting [5], [6], power consumption monitoring [7], [8], industrial maintenance [9], [10], occupancy monitoring in smart buildings [11], [12], and many others. Recently, deep learning (DL) models showed tremendous success in analyzing time series data [1]–[15] when compared to the other traditional methods. This is due to the fact that DL models can automatically learn complex mappings from multiple inputs to outputs. Interestingly, DL models can be easily fooled by adversarial examples [15], [16]. From the perspective of image processing or computer vision, an adversarial example can be an image formed by making small perturbations (insignificant to the human eye) to an example image. Another interesting fact is that the adversarial examples can often transfer from one model to another model, known as black-box attacks, which means that it is possible to attack DL models to which the adversary does not have access [16], [17]. In recent years, many techniques have been proposed for increasing the robustness of DL algorithms against adversarial examples [18]–[23], however, most of them have been shown to be vulnerable to future attacks [24].

Adversarial attacks in deep learning have been extensively explored for image recognition and classification applications. However, their application to the non-image domain is vastly under-explored. This also includes the lack of studies applying adversarial examples to time series analysis despite the increasing popularity of DL models in time series analysis. Very recently, the authors in [25] showed that a deep neural network (DNN) univariate time series classifier (specifically ResNet [26]) are vulnerable to adversarial attacks. Unfortunately, to the knowledge of the authors, there exists no research work to date evaluating the impact of adversarial attacks on multivariate time series (MTS) deep learning regression models. This is indeed a major concern as potential adversarial attacks are present in many safety-critical applications that exploit DL models for time series forecasting. For instance, adding small perturbations to multivariate time series data (using false data injection methods [27]) that uses a DL regression model [28] for smart grid electric load forecasting can generate wrong predictions, thus may lead to a nationwide power outage.

In this paper, we apply and transfer adversarial attacks from the image domain to deep learning regression models for MTS forecasting. We present two experimental studies using two datasets from the finance and energy domain. The obtained results show that modern DL regression models are prone to adversarial attacks. We also show that adversarial time series examples crafted for one network architecture can be transferred to other architectures, thus holds their transferability property [16]. Therefore, this work highlights the importance of protecting against adversarial attacks in deep learning regression models for safety-critical MTS forecasting applications.

To summarize, the main contributions of this paper are:

- Formalize adversarial attacks on DL regression models for MTS forecasting.
- Crafting adversarial attacks for MTS DL regression models using methods that are popular in the image
Adversarial attack: small perturbation

\[ X' = X + h \]

In this work, we focus on the time/cost-sensitive and safety-critical applications of deep learning time series forecasting, which motivates us for investigating the impact of adversarial attacks on them. Specifically, we explore the impact of adversarial attacks on LSTM, CNN, and GRU. All of these models are known for their effectiveness in time series forecasting. LSTM is capable of learning long-term dependencies using several gates and thus suits well the time series forecasting problems. In [40], authors employ an LSTM model for predicting the traffic flow with missing data. The other successful applications of LSTM in time series forecasting includes petroleum production forecasting [41], financial time series forecasting [42], solar radiation forecasting [43], and remaining useful life prediction of aircraft engines [44]. GRU is an improvised version of Recurrent Neural Network (RNN) [45], and also effective in time series forecasting [46]. For instance, in [47], authors employ 1D convnets and bidirectional GRUs for air pollution forecasting in Beijing, China. The other applications of GRU models in time series forecasting include personalized healthcare and climate forecasting [48], mine gas concentration forecasting [49], smart grid bus load forecasting [50]. In [51], authors present a CNN-based bagging model for forecasting hourly loads in a smart grid. Apart from the energy domain, CNNs are also useful for financial time series forecasting [1], [52].

In [53], time-series data from smart grids are analyzed for the detection of electricity theft. In such use cases, perturbed data can help thieves to avoid being detected. Using adversarial attacks, a hacker might generate such perturbed synthetic data to bypass the system’s attack detection techniques without even having access or knowledge about the DL model used for decision making [16], [17]. Perturbing the data recorded by sensors placed on safety-critical applications (using false data injection techniques [27], [54]) such as aircraft engines, smart grids, gas pipeline, etc. have a catastrophic impact on human lives and productivity, whereas attacks on financial data [55]–[57] has a direct impact on the economy. Indeed, the list of potential attacks presented in this section is not exhaustive due to the space limitation.

\[ \text{Adversary} \]

\[ \text{Time series length } T \]

\[ \text{Dimension } N \]

**B. Adversarial attacks**

The concept of adversarial attack was proposed by Szegedy et al. [58] at first for image recognition. The main idea is to add a small perturbation to the input images which is insignificant to human eyes, but as a result, the target model misclassifies the input images with high confidence. The severity of such attacks is shown by the researchers in a recent experiment where a strip of tape on a 35 mph limit sign was added which tricked a self-driving car into acceleration to 85 mph [59]. Based on this idea proposed in [58], hundreds of methods have been proposed for their analysis [34]. With the success of machine learning (ML) algorithms in different domains, ML techniques for time series forecasting is also popular [35], [36]. However, among these methods, only a few (when compared to the non-DL methods) have considered DL methods for time series forecasting [37]–[39].

The concept of adversarial attack was proposed by Szegedy et al. [58] at first for image recognition. The main idea is to add a small perturbation to the input images which is insignificant to human eyes, but as a result, the target model misclassifies the input images with high confidence. The severity of such attacks is shown by the researchers in a recent experiment where a strip of tape on a 35 mph limit sign was added which tricked a self-driving car into acceleration to 85 mph [59]. Based on this idea proposed in [58], hundreds of methods have been proposed for their analysis [34]. With the success of machine learning (ML) algorithms in different domains, ML techniques for time series forecasting is also popular [35], [36]. However, among these methods, only a few (when compared to the non-DL methods) have considered DL methods for time series forecasting [37]–[39].
many researchers have developed algorithms [14], [15], [58], [60] for constructing such adversarial examples relying on the architecture and parameters of the DL model. Most of these adversarial attacks are proposed for image recognition tasks. A fast gradient sign method (FGSM) [14] attack was introduced in the year 2014 which signifies the presence of adversarial examples in image recognition tasks. Followed by FGSM, an iterative version of it, known as the basic iterative method (BIM) [15] was proposed in the year 2016. BIM showed more effectiveness in crafting a stealthier adversarial example, however, it comes with a higher computational cost. Comprehensive reviews of adversarial attacks in DL models in different applications can be found in [61]–[64].

Interestingly, the adversarial attack approaches for multivariate time series DL regression models have been ignored by the community. There are only two previous works that consider adversarial attacks on time series. In [65], the authors adopt a soft K-Nearest-Neighbours (KNN) coupled with Dynamic Time Warping (DTW) and show that the adversarial examples can fool the proposed classifier on a simulated dataset. Unfortunately, the KNN classifier is no longer considered the state-of-art classifier for time series data [66]. The authors in [25], utilize the FGSM and BIM attacks to fool Residual network (ResNet) classifiers for univariate time series classification tasks. In our work, we also employ the FGSM and BIM attacks, however, we apply and evaluate their impacts on DL regression models for multivariate time series forecasting.

In summary, our work sheds light on the resiliency of DL regression models for multivariate time series forecasting in real-world safety-critical and cost-critical applications (as explained in section II-A). This will guide the data mining, data science, and machine learning researchers to develop techniques for detecting and mitigating adversarial attacks in time series data.

III. ADVERSARIAL EXAMPLES FOR MULTIVARIATE TIME SERIES

In this section, we formalization of the problem, and present the FGSM and BIM attack algorithms that we use to generate adversarial MTS examples for the DL models.

A. Formalization of MTS regression

Definition 1: Let X be a multivariate time series (MTS). X can be defined as a sequence such that \( X = [x_1, x_2, ..., x_T] \), \( T = |X| \) is the length of X, and \( xi \in \mathbb{R}^N \) is a N dimension data point at time \( i \in [1, T] \).

Definition 2: \( D = (x_1, F_1), (x_2, F_2), ..., (x_T, F_T) \) is the dataset of pairs \( (xi, Fi) \) where \( Fi \) is a label corresponding to \( xi \).

Definition 3: Time series regression task consists of training the model on \( D \) in order to predict \( \hat{F} \) from the possible inputs. Let \( f(\cdot) : \mathbb{R}^{N \times T} \rightarrow \hat{F} \) represent a DL model for regression.

Definition 4: \( J_f(\cdot, \cdot) \) denotes the cost function (e.g. mean squared error) of the model \( f \).

Algorithm 1: FGSM attack on multivariate time series

Input : Original MTS \( X \) and its \( \hat{F} \)
Output : Perturbed MTS \( X' \)
Parameter: \( \epsilon \)

\[
\eta = \epsilon \cdot \text{sign}(\nabla_x J_f(X, \hat{F})); \\
X' = X + \eta.
\]

Definition 5: \( X' \) denotes the adversarial example, a perturbed version of \( X \) such that \( \hat{F} \neq \hat{F}' \) and \( \|X - X'\| \leq \epsilon \) where \( \epsilon \geq 0 \in \mathbb{R} \) is a maximum perturbation magnitude.

Given a trained deep learning model \( f \) and an input MTS \( X \), crafting an adversarial example \( X' \) can be described as a box-constrained optimization problem [64].

\[
\min_{X'} \|X' - X\| \quad \text{s.t.} \quad f(X') = \hat{F}' \quad \text{and} \quad \hat{F} \neq \hat{F}'
\]

Let \( \eta = X' - X \) be the perturbation added to \( X \). Fig[1] shows the process where a perturbation \( \eta \) is added to the original MTS \( X \) for crafting an adversarial example \( X' \).

B. Fast gradient sign method

The FGSM was first proposed in [14] where it was able to fool the GoogLeNet model by generating stealthy adversarial images. FGSM calculates the gradient of the cost function relative to the neural network input. This attack is also known as the one-shot method as the adversarial perturbation is generated by a single-step computation. Note, FGSM is an approximate solution based on linear hypothesis [61]. Adversarial examples are produced by the following formula:

\[
\eta = \epsilon \cdot \text{sign}(\nabla_x J_f(X, \hat{F}))
\]

\[
X' = X + \eta
\]

Here, \( J_f \) is the cost function of model \( f \), \( \nabla_x \) indicates the gradient of the model with respect to the original MTS \( X \) with the correct label \( \hat{F} \). \( \epsilon \) denotes the hyper-parameter which controls the amplitude of the perturbation and \( X' \) is adversarial MTS. Algorithm[1] shows different steps of the FGSM attack.

C. Basic iterative method

The BIM [15] is an extension of FGSM. In BIM, FGSM is applied multiple times with small step size and clipping is performed after each step to ensure that they are in the range [\( X - \epsilon \), \( X + \epsilon \)] i.e. \( \epsilon \)-neighbourhood of the original MTS \( X \). BIM is also known as Iterative-FGSM as FGSM is iterated with smaller step sizes. Algorithm[2] shows different steps of the BIM attack, where it requires three hyperparameters: the per-step small perturbation \( \alpha \), the amount of maximum perturbation \( \epsilon \), and the number of iterations \( I \). Note, BIM does not rely on the approximation of the model, and the adversarial examples crafted through BIM are closer to the original samples when compared to FGSM. This is because the perturbations are added iteratively and hence have a better
The dataset is comprised of seven variables (besides the rate for almost 4 years (December 2006 to November 2010)) consumption in one household with a one-minute sampling series dataset that includes the measurements of electric power. Household power consumption dataset is a multivariate time series dataset that includes the measurements of electric power. Household electric power consumption dataset is a multivariate time series dataset [33]. The attacks on household energy forecasting using the individual household electric power consumption dataset is shown in Fig. 3. For the FGSM attack (with $\epsilon = 0.2$), we observe that the RMSE for the CNN, LSTM, and GRU models (under attack) are increased by 19.9%, 12.3%, and 11%, respectively, when compared to the models without attack. For the BIM attack (with $\alpha = 0.001$, $\epsilon = 0.2$, and $I = 200$), we also observe the similar trend, that is the RMSE of the CNN, LSTM, and GRU models increased in a similar fashion, specifically by 25.9%, 22.9%, and 21.7%, respectively for the household electric power consumption dataset. We observe that for both FGSM and BIM attacks, it is evident that the CNN model is more sensitive to adversarial attacks when compared to the other DL models. Also, BIM results in a larger RMSE when compared to the FGSM. This means BIM is not only stealthier that FGSM, but also has a stronger impact on DL regression models for the this dataset.

IV. RESULTS

In this section, we evaluate the crafted adversarial examples on two datasets (from the finance and energy domain) and present the obtained results. We also provide a brief discussion on potential defense mechanism for detecting the adversarial MTS examples in DL regression models. For the sake of reproducibility and to allow the research community to build on our findings, the artifacts (source code, datasets, etc.) of the following experiments are publicly available on our GitHub repository[7].

A. Attacks on household power consumption

Due to the increase in demand for efficient energy needs, there is a need for a smart infrastructure to meet the growing demands and to generate energy more efficiently. Recently, deep learning [67]–[69] has shown tremendous success in forecasting the energy demands by training on the past power consumption data and forecasting the energy consumption in the future. This indeed helps in making an informed decision of how much energy should be generated for a given day in the recent future, avoids the excessive generation of surplus energy, and thus helps in reducing the loss of resources, manpower, and cost. In this context, an adversarial attack could result in incorrect predictions of global active power, which is the power consumed by electrical appliances other than the sub-metered appliances. Such an incorrect forecast may lead to either excessive surplus or inadequate generation of energy—both of which have a direct impact on cost, productivity, available resources, and environment.

In this work, we evaluate the impact of adversarial attacks on household energy forecasting using the individual household electric power consumption dataset [33]. The household power consumption dataset is a multivariate time series dataset that includes the measurements of electric power consumption in one household with a one-minute sampling rate for almost 4 years (December 2006 to November 2010) and collected via sub-meters placed in three distinct areas. The dataset is comprised of seven variables (besides the date and time) which includes global active power, global reactive power, voltage, global intensity, and sub-metering (1 to 3). We re-sample the dataset from minutes to hours and then predict global active power using seven variables or input features (global active power, global reactive power, voltage, global intensity, and sub-metering (1 to 3)). Then we use the first three years (2006 to 2009) for training our three DL models (LSTM, GRU, and CNN), and last year’s data to test our models. The DL architecture of the DL models can be represented as LSTM(100,100,100) lh(14), GRU(100,100,100) lh(14), and CNN(60,60,60) lh(14). The notation LSTM(100,100,100) lh(14) refers to a network that has 100 nodes in the hidden layers of the first LSTM layer, 100 nodes in the hidden layers of the second LSTM layer, and a sequence length of 14. In the end, there is a 1-dimensional output layer. In Fig. 2 we compare the performance of these three DL architectures in terms of their root mean squared error (RMSE) [70]. From Fig. 2 it is evident that the LSTM(100, 100, 100) has the best performance (with least RMSE) when predicting the global active power (without attack) which was trained with 250 epochs using Adam optimizer [71] and grid search [72] for hyperparameter optimization to minimize the objective cost function: mean squared error (MSE). The hyperparameter settings for the evaluated DL models are shown in Table I.

Fig. 3 shows an example of the normalized FGSM and BIM attack signatures (adversarial examples) generated for the global reactive power variable (an input feature in the form of a time series). Similar adversarial examples are generated for the remaining five input features to evaluate their impact on the LSTM, GRU and CNN models for energy consumption prediction (global active power prediction). As shown in Fig. 3, the adversarial attack generated using BIM is close to the original time series data which makes such attack stealthy, hard to detect and often bypass the attack detection algorithms. The impact of the generated adversarial examples on the household electric power consumption dataset is shown in Fig. 4. For the FGSM attack (with $\epsilon = 0.2$), we observe that the RMSE for the CNN, LSTM, and GRU model (under attack) are increased by 19.9%, 12.3%, and 11%, respectively, when compared to the models without attack. For the BIM attack (with $\alpha = 0.001$, $\epsilon = 0.2$, and $I = 200$), we also observe the similar trend, that is the RMSE of the CNN, LSTM, and GRU models increased in a similar fashion, specifically by 25.9%, 22.9%, and 21.7%, respectively for the household electric power consumption dataset. We observe that for both FGSM and BIM attacks, it is evident that the CNN model is more sensitive to adversarial attacks when compared to the other DL models. Also, BIM results in a larger RMSE when compared to the FGSM. This means BIM is not only stealthier that FGSM, but also has a stronger impact on DL regression models for the this dataset.

For instance, as shown in Fig. 4a, the CNN MTS regression model forecasts the global active power (without attack) to be 2.10 kW and 4.51 kW on 161st hour and 219th hour,

### Algorithm 2: BIM attack on multivariate time series

| Input : Original MTS $X$ and its $\hat{F}$ |
| Output : Perturbed MTS $X'$ |
| Parameter: $I$, $\epsilon$, $\alpha$ |

$X' \leftarrow X$;

while $i = 1 \leq I$ do

$\eta = \alpha \cdot \text{sign}(\nabla_x J_f(X', \hat{F}))$;

$X' = X + \eta$;

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

$i ++$;

end

1. https://github.com/dependable-cps/adversarial-MTSR
respectively. After performing the FGSM and BIM attack, the same CNN MTS regression model forecasts the global active power to be 1.36 kW and 0.37 kW on 161st hour, and 5.24 kW and 6.94 kW on 219th hour, respectively. This represents a 35.2% and 82.3% decrease, and a 16% and 53.8% increase in the predicted values on the 161st and 219th hour respectively (when compared to the without attack situation). Such an under-prediction as a consequence of attack may result in the inadequate generation of energy, thus leading to a failure of meeting the future energy demands with a potential power outage. In contrast, over-prediction may result in the surplus generation of energy leading to increased cost and waste of resources.

B. Attacks on stock prices

Data scientists and financial theorists have been employed for the past 50 years to make sense of the market by increasing the return on the investment. However, due to the multidimensional nature, the scale of the problem, and its inherent variation with time makes it an overwhelming task. Advancements in DL algorithms and their application to finance [1]–[4] has shown tremendous prospect to revolutionize this domain including stock market analysis and prediction. DL algorithms can learn the multivariate nature of the stocks and can make more accurate predictions [73], [74]. In this context, an adversarial attack could result in incorrect stock price predictions, which may, in turn, result in a diminishing return of the investment, and have a significant impact on the stock market.

In this work, we evaluate the impact of adversarial attacks on Google stock prediction using the Google stock dataset [32]. The Google stock dataset contains Google stock prices for the past 5 years. This multivariate time series dataset has

| DL models | Power consumption dataset | Google stock dataset |
|-----------|--------------------------|----------------------|
|           | Hidden neurons | Batch size | Epochs | Hidden neurons | Batch size | Epochs |
| CNN       | 60,60,60        | 512       | 200    | 60,60,60        | 14        | 250    |
| LSTM      | 30,30,30        | 32        | 250    | 100,100,100    | 14        | 300    |
| GRU       | 30,30,30        | 32        | 250    | 100,100,100    | 14        | 300    |

Fig. 2: Comparison of deep learning algorithms for power consumption dataset

Fig. 3: Attack signatures for power consumption dataset; FGSM ($\epsilon = 0.2$) and BIM ($\alpha = 0.001$, $\epsilon = 0.2$, and $I = 200$)
Similar adversarial examples are also generated using the volume of stocks traded (an input feature in the form of a time series). BIM attack signatures (adversarial examples) generated for the BIM attack are shown in Table I. The hyperparameter settings for the evaluated DL models are shown in Table II.

The objective cost function: mean squared error (MSE). The grid search [72] for hyperparameter optimization to minimize MSE was trained with 300 epochs using Adam optimizer [71] and when predicting stock opening prices (without attack) which GRU(30, 30, 30) has the best performance (with least RMSE) and CNN(60,60,60) lh(60). From Fig. 5, it is evident that the CNN MTS regression model forecasts the normalized stock opening price (without attack) which is more sensitive to adversarial attacks when compared to the LSTM, GRU and CNN models for the Google stock prediction (stock opening price). From Fig.6 we observe that the adversarial attack generated using BIM is close to the original time series opening price. From Fig.6, we observe that the adversarial attacks are more sensitive to adversarial attacks when compared to the models without attack. For the BIM attack (with $\alpha = 0.001$, $\epsilon = 0.2$, and $I = 200$), we also observe the similar trend, that is the RMSE for the CNN, LSTM and GRU model (under attack) are increased by 35.2%, 27.2% and 28.9%, respectively. Similar to our observation on the power consumption dataset, we notice that the CNN model has a larger RMSE when compared to the FGSM. Moreover, we also observe the similar trend, that is the RMSE for the CNN, LSTM and GRU model (under attack) are increased by 16%, 12.9%, and 13.1%, respectively, when compared to the models without attack. For instance, as shown in Fig. 7a, the CNN MTS regression model forecasts the normalized stock opening price (without attack) to be $0.781$ on day 11 and $0.662$ on day 297. After performing the FGSM and BIM attack, the same CNN MTS

![Fig. 4: Power consumption prediction after FGSM ($\epsilon = 0.2$) and BIM ($\alpha = 0.001$, $\epsilon = 0.2$, and $I = 200$)](image)

![Fig. 5: Comparison of deep learning algorithms for Google stock dataset](image)
thus RMSE follows an increasing trend. This is due to the
MTS examples for fooling both the regression models and
$\epsilon$ larger values of $\alpha$ helpful in generating adversarial MTS examples for fooling Fig. 5. We observe that for larger values of $\alpha$ best performance for the MTS regression task in Fig. 2 and examples. We pick the LSTM and GRU as they showed the
of perturbations allowed for crafting the adversarial MTS
prices are decreasing and increasing, respectively, thus leading
on day 11 and 297, respectively. This represents a 10.6% and 24.8%
regression model forecasts the normalized stock opening price
to be $0.864$ and $0.975$ on day 11, and $0.607$ and $0.556$ on
day 297, respectively. This represents a 10.6% and 24.8%
increase, and an 8.3% and 16% decrease in the predicted stock
prices on day 11 and 297, respectively (when compared to the
without attack situation). Such an over-prediction and under-
prediction in stock prices may result in investors investing
more and investing less in a particular stock whereas the stock
prices are decreasing and increasing, respectively, thus leading
to a loss in the return of investment in both cases.

### C. Performance variation vs. the amount of perturbation

In Fig. 8 we evaluate the LSTM and GRU regression
model’s performance with respect to the different amount
of perturbations allowed for crafting the adversarial MTS
examples. We pick the LSTM and GRU as they showed the
best performance for the MTS regression task in Fig. 2 and
Fig. 5. We observe that for larger values of $\epsilon$, FGSM is not very
helpful in generating adversarial MTS examples for fooling
the LSTM and GRU regression model. In comparison, with
larger values of $\epsilon$, BIM crafts more devastating adversarial
MTS examples for fooling both the regression models and
thus RMSE follows an increasing trend. This is due to the

- **Fig. 6: Attack signatures for google stock dataset; FGSM ($\epsilon = 0.2$) and BIM ($\alpha = 0.001$, $\epsilon = 0.2$, and $I = 200$)**
- **Fig. 7: Stock price prediction after FGSM ($\epsilon = 0.2$) and BIM ($\alpha = 0.001$, $\epsilon = 0.2$, and $I = 200$)**
- **Fig. 8: RMSE variation with respect to the amount of perturbation ($\epsilon$) for FGSM and BIM attacks**
D. Transferability of adversarial examples

To evaluate the transferability of adversarial attacks, we apply the adversarial examples crafted for a DL MTS regression model on the other DL models. Table II summarizes the obtained results on transferability. We observe that for both datasets, the adversarial examples crafted for CNN are the most transferable. This means a higher RMSE is observed when adversarial examples crafted for the CNN model are transferred to other models. For instance, adversarial MTS examples crafted using BIM for the CNN regression model (Google stock dataset) causes a 23.4% increase when transferred to the GRU regression model. A similar trend is also observed, however, with a lower percentage increases, when adversarial examples crafted for GRU and LSTM regression models are transferred to the other DL regression models. In addition, the obtained results also show that BIM is better than FGSM in fooling (even when they are transferred) the DL models for MTS regression tasks, e.g. BIM increases the RMSE more when compared to the FGSM. Overall, the results show that the adversarial examples are capable of generalizing to a different DL network architecture. This type of attack is known as black box attacks, where the attackers do not have access to the target models internal parameters, yet they are able to generate perturbed time series that fool the DL models for MTSR tasks.

E. Defense against adversarial attacks

Researchers have proposed different types of adversarial attack defense strategies so far [61] most of which are applicable to the image domain. The existing adversarial attack defense strategies can be divided into three categories: modifying data, modifying models, and using auxiliary tools. Modifying data refers to modifying the training dataset in the training stage, or changing the input data in the testing stage. It also includes adversarial training [14], blocking the transferability [75], data compression [76], gradient hiding [77], and data randomization [78]. In contrast, modifying models refer to the modification of DL models, such as defensive distillation [79], feature squeezing [80], regularization [81], deep contractive network [82] and mask defense [83]. Using additional tools to the DL models is referred to as using an auxiliary tool which includes the use of defense-GAN [84], MagNet [85] and high-level representation guided denoiser [86]. Unfortunately, most of these detectors are prone to adversarial attacks due to the fact that these attacks are designed specifically to fool such detectors [64]. Hence, the time series, data mining and machine learning need to pay special attention to this area as DL MTS regression models are gaining popularity in the safety and cost-critical application domains. A potential idea for the detection of adversarial examples in MTS DL regression models can be the use of inductive conformal anomaly detection method [87], [88]. Another potential idea is to leverage the decades of research into non-probabilistic classifiers, such as the nearest neighbor coupled with DTW [64].

V. Conclusion

In this paper, we introduced the concept of adversarial attacks on deep learning (DL) regression models for multivariate time series (MTS) regression. We formalized and evaluated two adversarial example generation techniques, originally proposed for the image domain for the MTS regression task. The obtained results showed how adversarial attacks can induce inaccurate forecasting when evaluated on the household power consumption and the Google stock dataset. We also observed that BIM is not only a more stealthy attack but also causes higher damage in DL MTS regression models. Finally, among the three evaluated DL regression models, the obtained results revealed that the adversarial examples crafted for CNN are more transferable when compared to the others. Through our work, we shed light on the importance of acknowledging adversarial attacks as one of the prominent threats to the DL MTS regression models as they find their applications in safety-critical and cost-critical domains.

In the future, we would like to extend our work by adapting other adversarial attacks for the image domain and evaluate them for MTS DL regression. In addition, we also plan to investigate defense strategies to detect and mitigate adversarial threats in DL regression models.

TABLE II: Transferability of FGSM and BIM attacks for power Consumption and Google stock datasets. The notation X/Y represents the percentage of RMSE increase using FGSM/BIM

| DL models | Transferability (% increase of RMSE) |
|-----------|-------------------------------------|
|            | Power consumption dataset | Google stock dataset |
| CNN       | LSTM | GRU | CNN | LSTM | GRU |
| LSTM      | 8.3/16.9 | 7.5/11.2 | 13.8/19.7 | 11.6/16.3 |
| GRU       | 9.2/16.5 | 6.6/11.7 | - | - |
| LSTM      | 10.8/18.1 | - | - | - |
| GRU       | 11.6/16.3 | - | - | - |
| LSTM      | 10.2/18.1 | 7.5/11.2 | 13.8/19.7 | 11.6/16.3 |
| GRU       | 11.6/16.3 | - | - | - |
| LSTM      | 16.2/23.4 | - | - | - |
| GRU       | 16.2/23.4 | - | - | - |

References

[1] O. B. Sezer, M. U. Gudelek, and A. M. Ozbayoglu, “Financial time series forecasting with deep learning: A systematic literature review: 2005–2019,” Applied Soft Computing, vol. 90, p. 106181, 2020.
[2] L. Gan, H. Wang, and Z. Yang, “Machine learning solutions to challenges in finance: An application to the pricing of financial products,” Technological Forecasting and Social Change, vol. 153, p. 119928, 2020.
[3] J. Sirignano and R. Cont, “Universal features of price formation in financial markets: perspectives from deep learning,” Quantitative Finance, vol. 19, no. 9, pp. 1449–1459, 2019.
[4] S. I. Lee and S. J. Yoo, “Multimodal deep learning for finance: integrating and forecasting international stock markets,” The Journal of Supercomputing, pp. 1–19, 2019.
[5] A. G. Salmon, B. Kanigor, and Y. Heryadi, “Weather forecasting using deep learning techniques,” in 2015 international conference on advanced computer science and information systems (ICACSIS). IEEE, 2015, pp. 281–285.
