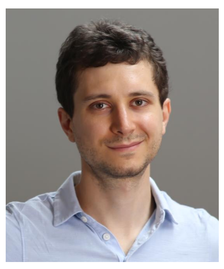
UNIVERSITÄT LIECHTENSTEIN

Liechtenstein Business School

SAMLAF: Security Assessment of Machine Learning Applications in Finance

Reviewing the State of the Art, developing realistic Algorithmic Trading Systems, and exposing their vulnerabilities to 'Ephemeral Perturbations' (open-source artifact)

Giovanni Apruzzese, Advije Rizvani
Liechtenstein Business School, Universität Liechtenstein



Abstract

We study the security of stock price forecasting using Deep Learning (DL) in *computational finance*. Despite abundant prior research on vulnerability of DL to adversarial perturbations, such work has hitherto hardly addressed practical adversarial threat models in the context of DL-powered *algorithmic trading systems* (ATS). Specifically, we investigate the vulnerability of ATS to adversarial perturbations launched by a realistically constrained attacker. We first show that existing literature has paid limited attention to DL security in the financial domain—which is naturally attractive for adversaries. Then, we formalize the concept of *ephemeral perturbations* (EP), which can be used to stage a novel type of attack tailored for DL-based ATS. Finally, we carry out an end-to-end evaluation of our EP against a profitable ATS. Our results reveal that the introduction of small changes to the input stock-prices not only *(i)* induces the DL model to behave incorrectly but also *(ii)* leads to the whole ATS to make suboptimal buy/sell decisions, resulting in a worse financial performance of the targeted ATS.

Literature Review

Background

Algorithmic Trading Systems (ATS) are tools that automatically make trading decisions (e.g., buy or sell stocks). Advances in machine and deep learning (ML and DL) led to DL-based techniques suited for time-series forecasting (such as LSTM) to be deployed in ATS. In this way, ATS can quickly analyze large amounts of historical data and, by accurately predicting future values, make sensible market decisions according to the available resources and trading strategies. DL-powered ATS are widely used today, due to their potential of "automatically" generating a profit for their owners. However, such widespread deployment naturally begs the question: *how reliable are these systems in adversarial environments?*

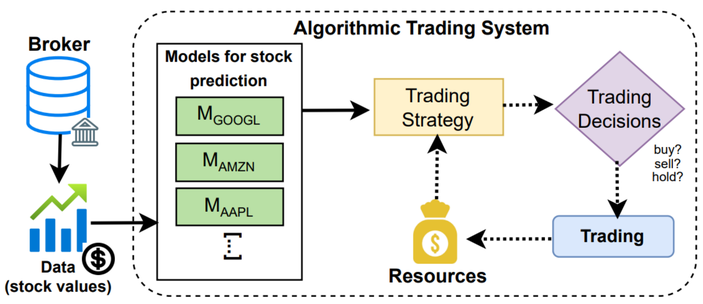


Fig. 1: Schema of an Algorithmic Trading System (ATS). The *broker* (e.g., a bank) sends stock-related *data* to a given organization which owns an ATS (dotted box). The ATS includes various DL models, used to make predictions on the basis of the input data. Such predictions are then used by the ATS to enact a given *trading strategy*, which must account for the available *resources* and decide what to do (i.e., buy/hold/sell). After making a decision, the resources are updated.

Related Work and Research Gap

The security vulnerabilities of DL/ML have been widely investigated by prior work on 'adversarial ML'. Similarly, a large body of research tackled various security aspects of computational finance (e.g., fraud detection). However, we wondered: *"to what extent has prior literature considered the DL-specific vulnerabilities of DL-powered ATS?"* We carried out a systematic literature review encompassing over 7000 peer-reviewed works accepted in top-tier venues, complemented by a broad search on Google Scholar on an additional set of 390 papers. We found that only 5 papers can be considered to address security aspects that are specific of ML/DL deployments in ATS. However, further examination revealed that such works are affected by shortcomings such as: the assumption of unrealistic threat models; lack of source-code disclosure; and a predominant focus on the effects of the attack on the DL model—instead of scrutinizing its system-wide effects on the overarching ATS. This finding drove our second research question: *"in what way can DL models of ATS be misled so as to induce the ATS to make decisions that negatively affect its real-world profitability?"*

Algorithmic Trading System Security Framework (ATS-SF)

Our extensive literature review highlighted a worrying lack of open-source resources that could be used to address our second research question. Without a valid baseline (i.e., a DL-powered ATS whose performance would justify its real-world deployment), it is just impossible to carry out any meaningful security assessment. Hence, as a starting point, but also to provide a constructive foundation for future work, we have developed the "Algorithmic Trading System Security Framework" (ATS-SF), for which we provide an open source repository (https://github.com/AdvijeR/ep-ats/). We used ATS-SF to develop our baseline ATS, which leverages LSTM neural networks to predict the next-day stock prices of a portfolio encompassing 38 stocks. Through a 2-year simulation on real stock-market data, we demonstrate that our ATS would net a profit to its owners (as measured by a positive Sharpe Ratio, and Cumulative Returns denoting a 25% increase of the initial capital).

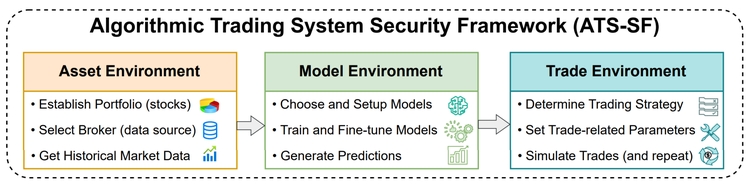


Fig. 2: Architecture of our ATS-SF. Our framework has three environments that allow fine-grained control on the entire management pipeline of an ATS, thereby enabling security assessments.

Original threat model: Ephemeral Perturbations

It is well-known that DL models can be "attacked" by manipulating the data received as input via so-called "adversarial perturbations". However, the impact of perturbations stemming from an attacker who is subject to the real-world constraints defining an (arguably, highly secure) ATS context are currently unknown. We thus hypothesize an original attack based on the concept of "ephemeral perturbations" (EP). The guiding principle of EP is that they should not just be "small" but they can affect only a single time unit analysed by the DL-based model integrated in the ATS. In other words, an EP should look like a natural occurrence---an "error" that may be due to communication issues between the ATS and the broker used to receive the up-to-date market data. Indeed, we assume an attacker who, potentially via a successful man-in-the-middle attempt, can somehow manipulate the data sent by the broker and received by the ATS: however, if such manipulations are conspicuous, the owners of the ATS would notice this and potentially opt to change broker (which would be bad for the attacker). This is why the changes that the attacker can introduce should be small. Moreover, the attacker has no access to the internal pipeline of the ATS, preventing fine-grained changes (so called "white-box attacks") that are guaranteed to dramatically affect the ATS. Put simply, the attacker is subject to hard constraints, preventing to accurately determine when to introduce an EP, or how large such an EP should be to achieve the intended goal (i.e., inducing the targeted ATS to yield a lower profit). Such a threat model is hence very realistic—but the effects of tiny EP on an exemplary ATS are unpredictable. For instance, due to an EP, the ATS may make a "wrong" decision on the following day, but in the long run the very same ATS may make trades thae ultimately lead to a higher profit (thereby making the corresponding EP detrimental for the attack). At the same time, a different EP may lead to an unrecoverable, but imperceptible, reduction in the net profits of the ATS for its owners (which is the attacker's true goal).

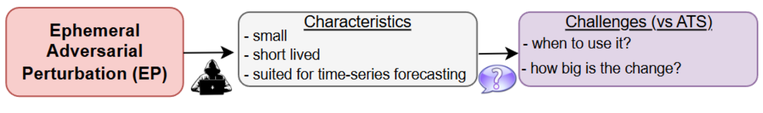
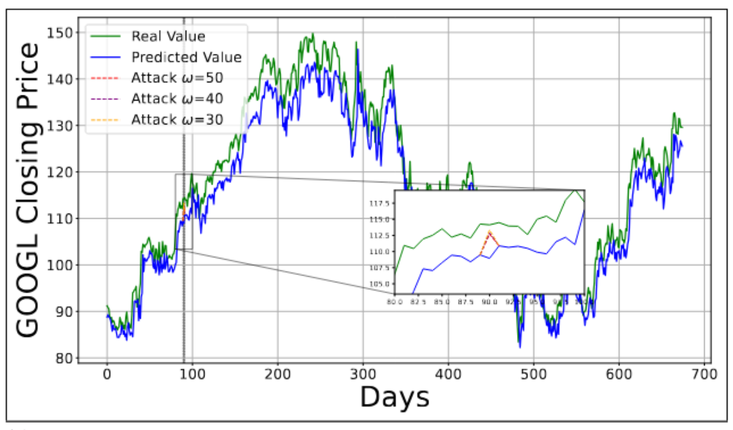


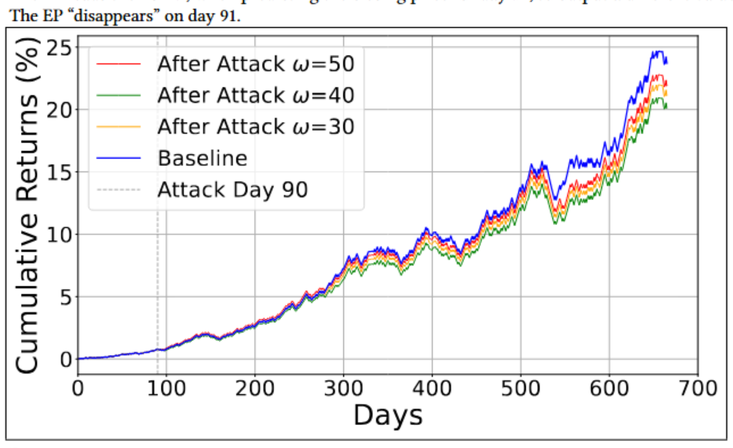
Fig. 3: Ephemeral Adversarial Perturbations (EP). Our EP must avoid raising suspicion. It is particularly applicable to DL for time-series: EP last for a single time-point—thereby affecting only one prediction.

Evaluation: what can EP do?

We tested our baseline ATS against exemplary EP. For a comprehensive evaluation, we investigated the effects that EP would have if they were launched on each day of the 2-year simulation. For instance, we show in Fig. 4 the effects of three EP (each having a different "magnitude", denoted with ω) when used to slightly change the data of a single stock (out of 38) included in the portfolio and received by the ATS on a single day (out of the 666 days of our simulation). Intriguingly, we see that these EP have an almost negligible effect on the predictions of the underlying DL-based predictor (i.e., an LSTM network); however, such an EP led to the overarching ATS to make a trading decision that significantly affected all the future decisions---ultimately leading to the ATS yielding a lower (but still positive) profit to its owners. Such an effect (i.e., lower cumulative returns than the baseline) appeared for over 60% of the EP we tested across the 2-year simulation.



(a) DL predictor. We introduce a tiny EP (dotted line) on the closing price of GOOGL on day 90. The EP leads the LSTM, when predicting the closing price for day 91, to output a different value. The EP "disappears" on day 91.



(b) Whole ATS. Effects of the EP introduced on day 90. Starting from the following day (day 91), the Cumulative Returns of the ATS drops (inducing a monetary loss) w.r.t. the baseline—despite the EP affecting only one day.

Fig. 4: Exemplary results of an EP. We showcase what happens if a DL predictor and overarching ATS are targeted by some of our proposed EP. The blue line represents the baseline performance (y-axis), whereas the others represent the effects of various EPs (targeting the same day, but with different m) over our test timeframe (x-axis).

Conclusions

In summary, the SAMLAF research project highlighted that:

- Limited attention has been given to the DL-specific security vulnerabilities of (DL-powered) ATS, demanding more research.
- There was a lack of open-source tools for out security assessments of DL-driven ATS (such a lack was filled by ATS-SF)
- DL-powered ATS can be significantly affected by (imperceptible) EP, calling for future research to develop ad-hoc defenses.

These takeaways should be heard by researchers and practitioners alike, due to the increasing reliance of ATS on DL methods---the security of which must be appropriately scrutinized. For instance, recent developments in large-language models beg the question: are such models affected by EP, too? Our reproducible evaluation enables the investigation of this (and similar) research question.

Te findings stemming from the SAMLAF project have been covered in a peer-reviewed publication that has just been accepted to ACM CODASPY [1], which also includes all experimental details.

Reference Publication

Rizvani, A., Apruzzese, G., & Laskov, P., "The Ephemeral Threat: Assessing the Security of Algorithmic Trading Systems powered by Deep Learning" ACM Conference on Data and Application Security and Privacy (CODASPY), 2025