Is it a great Autonomous Trading Strategy or you are just fooling yourself

Murilo Sibrão Bernardini and Paulo André Lima de Castro
Autonomous Computational Systems Lab - LABSCA
Aeronautics Institute of Technology (ITA)
São José dos Campos, São Paulo, Brazil
murilosibrao@gmail.com, pauloac@ita.br

Abstract—In this paper, we propose a method for evaluating autonomous trading strategies that provides realistic expectations, regarding the strategy’s long-term performance. This method addresses many pitfalls that currently fool even experienced software developers and researchers, not to mention the customers that purchase these products. We present the results of applying our method to several famous autonomous trading strategies, which are used to manage a diverse selection of financial assets. The results show that many of these published strategies are far from being reliable vehicles for financial investment. Our method exposes the difficulties involved in building a reliable, long-term strategy and provides a means to compare potential strategies and select the most promising one by establishing minimal periods and requirements for the test executions. There are many developers that create software to buy and sell financial assets autonomously and some of them present great performance when simulating with historical price series (commonly called backtests). Nevertheless, when these strategies are used in real markets (or data not used in their training or evaluation), quite often they perform very poorly. The proposed method can be used to evaluate potential strategies. In this way, the method helps to tell if you really have a great trading strategy or you are just fooling yourself.

I. INTRODUCTION

Building autonomous agents for trading, sometimes called Algorithmic trading or trading robots has been the focus of much research and development in recent years. There are many published papers relating remarkable performance of the strategies in backtests, which are simulations that use historical data. We discuss some of these results briefly in Section III. Nevertheless, when these strategies are used in actual markets (or data not used in their training or evaluation), quite often they perform very poorly. The average results vary considerably but are usually disappointing. Some analysts have even argued that a positive backtesting performance may not be taken to indicate a negative performance in real financial scenarios.

The goal of an investment manager, automated or human, is to find and acquire the most desirable set of assets that fits investor’s preference profile. The manager acquires these assets through the submission of buy and sell orders to the stock market and such decisions are based on analyzes of a set of possible interesting assets. Obviously, with so many assets to choose from, each with a wealth of data to review, a machine learning approach, if not required. It is important to note that there are also ethical and legal implications that should be taken into consideration in the process of building automated trading strategies. We recommend Wellman and Rajan’s paper [1] about ethical issues in Autonomous Trading Agents.

The broad spectrum of AI techniques used in finance includes reinforcement Learning [2], multiagent systems [4], complex networks [6], decision trees [7], genetic algorithms [8], random forests [9] to more recent approaches, like convolutional neural networks [10] and deep reinforcement learning [11]. Recent studies have combined several techniques as LSTM NN and reinforcement learning [12], while others have explored an ensemble of deep learning models in a multi-layer architecture [13]. There are many cases where the developers are successful in creating strategies with great performance results when backtesting with historical price series. In fact, some electronic platforms offer thousands of autonomous agents, or trading robots, which are also called bots or expert advisors. Their developers claim this software will be profitable in real financial markets [14]. These agents’ strategies may be created with a simple idea such as moving averages or use complex machine learning schemes. To date, however, these strategies have usually performed poorly with data not used in their training, despite all the hype.

The problem of creating a trading strategy with great performance in some previous known data set, but with bad performance in new data is well known in Machine Learning field and it is usually called overfitting. Nevertheless, this problem seems to be unknown or at least underestimated by many developers in autonomous trading. In truth, overfitting is more likely to occur in finance than in the traditional machine learning problems on which it is based, such as face recognition. Moreover, the consequences of a false positive can be far more disastrous in financial trading. False expectations are created that can mislead investors, putting their financial livelihood at stake. There is very little out there to protect investors from naive, incompetent, or even malicious software developers. Section I-A Overfitting in Finance, deals with this important issue.

In this paper, we propose a method for evaluating trading strategies that implements procedures that may applied regardless of the way the strategies were built or the technology
they use. This method is called the (statistically) Significant Trading-Strategy Evaluation, or STSE. It highlights many pitfalls that have been overlooked by software developers, including experienced machine learning experts, thus minimizing financial risks when using these strategies. We present the results of applying this method to several famous autonomous strategies used in foreign exchange assets and stock market. The results show that these strategies are far from being reliable vehicles of investment. Some periods of returns can be observed, but they are not predictable and probably happen by pure chance, rather than by the software having identified a real pattern in the input data. This fleeting return can be used maliciously by those selling the strategies, by their advertising great, but non-repetitive results, while obscuring negative results. The method we propose can be used to select among potential strategies, by establishing minimal periods and requirements as part of the test to avoid this seesaw in return results. Some autonomous approaches have been tried for dealing with derivative instruments [15]. We did not apply our method to any derivative instruments, which involve the future value of an underlying asset [16], because of their price volatility [16]. Instead, we focused on foreign exchanges and stock assets.

The remainder of this paper is organized as follows: In Section III we discuss some comparable proposals. In Section II, we present some distinct characteristics of financial markets that make them an unique challenge for machine learning and artificial intelligence approaches. In Section IV we propose our Significant Trading-Strategy Evaluation (STSE) method. We used it to evaluate several well known trading strategies currently on the market. The results are discussed in Section V. Finally, we discuss some possible extensions and future study in Section VI. We have also all necessary details to reproduce the experiments (Appendix A), which results are are shown in Section V.

We believe that reproducible results are important for research and that there is a significant reproducibility crisis in general. In fact, a recent survey [17] published in Nature journal showed that many researchers share that opinion. As pointed out by Stark [18], researchers should provide enough information for others to repeat their experiments. We do that in appendix A. We also provide the source code that implements STSE freely at github [19].

II. AI IN FINANCE

A financial agent is responsible for selecting, buying and selling financial assets in order to fulfill its investors requirements for risk and return, respecting the investor’s horizon of investment. A financial autonomous agent is a software program that implements a trading strategy able to perform such activities more efficiently.

The Financial Environment where these trading strategies are applied may be classified as: partially observable, sequential, stochastic, dynamic, continuous and multiagent, according to the Russell and Norvig taxonomy [20]. The environment is also strategic, in the sense that two active investors compete for a more accurate valuation of assets and their acts may change other economic agents’ behavior. Russel and Norvig consider this the most complex environment class [20].

This taxonomy, however, does not really represents the whole complexity of the problem, because financial markets are not an independent and identically distributed (i.i.d) environment, as assumed in most of machine learning problems. Furthermore, financial environment have a low signal to noise ratio, which makes overfitting even more likely in machine learning approaches. Another aspect that should be addressed by any trading strategy, but is often forgotten, is that people do not have the same level of acceptable risk. Some investors may present a much stronger risk aversion than others. A trading agent must be aware of its investor’s preferences, in order to trade appropriately. Investment errors caused by overfitting can also be very costly.

A. Overfitting in Finance

Financial environments are not only stochastic environments but usually non-stationary process where the probability distribution do change along the time. Due to this non-stationary feature, a specific strategy may perform extremely well for a certain period of time and then perform very poorly thereafter. Moreover, different assets may require different information and models. For instance, oil companies may be very sensitive to changes of petrol prices, while probably it is not the same case for banks. In other terms, a specific asset price can be very correlated to some time series, but that time series may have no relevant correlation with another asset price.

Markets conditions may change abruptly leading to structural breaks in financial time series, which are unexpected changes over time in the parameters of regression models. In fact, it can be checked by searching for structural breaks in financial time series. As stated by Andreou and Ghysels [21], there are abundant empirical evidence which shows that there are structural breaks in financial markets that may be caused by various economic events. When structural breaks occur the true model is changed; therefore, previous data does not describe the same process. It reduces the amount of useful data that can be used in fitting machine learning models. However, if that effect is not observed and outdated data is used, despite structural breaks, it may lead to models that do not generalize well in current market situations; or in other words, Overfitted models. Structural breaks do not usually occur in ML environments, which are traditionally considered to be independent and identically distributed (i.i.d). This hypothesis is so widely used in ML, that sometimes researchers forget that it is not valid for financial environments.

Moreover, models with high representational capacity, such as deep learning approaches [22], have higher risks of overfitting, if they are not properly used. The problem is that such models have many parameters and therefore they can fit many different solutions to the available data points. There is little chance that the learning algorithm will learn a solution.
that generalizes well, i.e. it is not overfitted, when so many wildly different solutions exist [22].

In financial environments, it is hard to separate signal (real information) from the noise, because they are almost at the same level and, if the signal is clear, it will not be for long. Arbitrage, exploiting price differences of similar financial instruments on different markets, further decreases the signal-to-noise ratio, making it easier to confuse signal and noise. This confusion may also lead to overfitted models.

Overfitting in finance leads to unrealistic expectation of the performance and, therefore, it can be used to mislead investors to bet on overfitted autonomous trading systems, which may be built by naive or even malicious developers [23].

In fact, it is not rare to see promises of high return by trading strategy vendors in digital media. As pointed out by Prado [23], without control for backtest overfitting, good backtest performance may be an indicator for negative future results.

Unfortunately, it is not easy to prove that a given model is not overfitted, despite the fact that there are many techniques in the ML field that help avoiding lack of generalization power [22], [24]. It is beyond the scope of this paper to address such general ML techniques to avoid overfitting and we are going to focus on specific features of financial environments. In fact, we believe it is fundamental to have a good understanding of the market faced by trading strategies in order to avoid problems. By that we mean the reasons that make overfitting more likely to happen in the financial environment and some common mistakes in building and evaluating trading strategies. We tackle these issues with the proposed method in Section IV.

III. RELATED WORK

There is a long history of studies related to evaluating portfolio management performance in the finance field [25]. However, there are some caveats when applying these methods to autonomous trading strategies. It is relatively simple to create new trading strategies using some machine learning technique and historical data or some model based approach with optimized parameters for a given market scenario. A developer needs to avoid some common pitfalls (1), select a meaningful measure (2) and be able to correctly pick among possible candidate strategies (3). We describe some related work that discuss these issues next.

A. Common pitfalls in evaluating trading strategies

The complexity of the financial environment for autonomous agents, as discussed in Section II, hides common pitfalls for the evaluation of trading strategies. We have noted that these problems have been overlooked in many studies as pointed out by other researchers [26]. The main pitfalls for evaluating trading strategies are the following:

- **Overfitting** allow bad trading strategies present good performance for certain periods of time, because they exploit false patterns that are unlikely to repeat in real markets and it can be easily happen in trading strategies.
- **Data leakage** can occur by using information that was not public at the moment the simulated decision was made. The time stamp for each data point must be observed and take into account release dates, distribution delays, and back-fill corrections [26]. One should also avoid training the model with testing data, or allow the leak of testing data to the training process. Some kind of leakage (or data-snooping) may also happen when a given set of data is used more than once for purposes of inference or model selection [27]. Data leakage may also happen using data points with very close time stamps in training and testing sets. See discussion in [28] about overlapping and embargo (pgs. 105-109).
- **Survivor bias** may occur by selecting to simulate only securities and companies that are still listed; therefore, ignoring all bankrupt and delisted securities in the process [26].
- **Evaluating strategies without caring about risk:** In finance theory and finance institutions, a great deal of effort is devoted to to measure and control risk. Since the seminal work of Markowitz [29] that established modern portfolio theory, building diversified portfolio, in order to achieve high returns and mitigate risk is the norm in the finance field. It is surprising to observe that creators of financial autonomous strategies often underestimate the relevance of risk measurement and control and care only about return [4] (pg. 213).
- **The Disregard transaction costs** produces strategies that look great in backtests but may bring negative returns systematically in real operation [26].
- **The lack of control for the impact of their own trading** is based on the assumption that the autonomous trading strategy will not have significant effect in the price or on the behaviour of others traders. This is really reasonable as long the capital controlled by the strategy is significantly smaller than the total volume traded. However, it is relevant to note in this context, that autonomous trading strategies are easy to copy usually. In fact, there are platforms that sell or rent autonomous strategies to third parties, which magnifies its impact. Furthermore, copying trading strategies or using other trades as primary information on other strategies are not new [30]. The idea of investors building and sharing investment strategies with other investors has been called social trading by some [31] (pg. 127) and pointed out as a key trend in asset management [31].
- **Oversimplifying shorting** may falsely enhance performance results. In a short position, the investor borrows stock and sells it with plans to buy it later. Taking a short position on spot markets requires finding a lender. The cost of lending and the amount available is generally unknown, and depends on several factors like relations, inventory and others [26], but assuming it is zero is certainly a mistake.
B. Sharpe ratio as Investment Skill measure

Most of the work related to defining a meaningful performance evaluation for autonomous trading strategies rely on the Sharpe ratio. That includes those cited in this Section. In fact, the Sharpe ratio is well known and often used in evaluating fund performance. However, there are significant criticisms about it. One criticism is that it does not distinguish between upside and downside volatility. According to Rollinger and Hoffman [32], high outlier returns could increase the value of the Sharpe ratio’s denominator (standard deviation) more than the value of the numerator (return above the risk free return), thereby lowering the value of the ratio. Another criticism is that the assumption of normality in return distributions have been relaxed and the Sharpe ratio thus becomes a questionable tool for selecting optimal portfolios [35]. In fact, some indexes have been proposed as alternatives to the Sharpe ration, including the Sortino ratio [34] and Modigliani index [35]. Despite the quality of these proposals, the Sharpe ratio “...definitely is the most widely used.” measure of risk-adjusted performance [32](pg.42). Furthermore, as stated by Bailey and others [36](pg. 15) the Sharpe ratio, despite its known deficiencies, can provide evidence of investment skill since a proper track record length is observed.

C. Picking among trading strategies

Another notable related work defines a minimum track record length (or minimum backtest length) needed to avoid selecting the trading strategy with highest in-sample Sharpe Ratio among N independent strategies, but that has an expected out-of-sample Sharpe Ratio equal to zero [23]. The focus is avoiding selecting an overfitted strategy that performs well in the training (or in-sample) data, but performs poorly in real operations. Bailey and others [23] demonstrated the Theorem 1. It stresses that the backtest period must be longer as the number of independent strategies (N) grows in order to avoid the risk of selecting an overfitted strategy.

**Theorem 1.** The Minimum Backtest Period (MinBTL, in years) needed to avoid selecting a strategy with an in-sample Sharpe ratio of \( E[\max_N] \) among \( N \) independent strategies and with an expected out-of-sample Sharpe ratio of zero is given by Equation 1 (For demonstration, see [23]):

\[
\text{MinBTL} < \frac{2 \times \ln[N]}{E[\max_N]^2}.
\]

In Section IV we propose a method that incorporates such concerns when allows the evaluation of autonomous trading strategies, regardless of the way they were built.

IV. THE SIGNIFICANT TRADING-STRATEGY EVALUATION (STSE) METHOD - HOW TO EVALUATE TRADING STRATEGY PERFORMANCE

We propose a method called Significant Trading-Strategy Evaluation (STSE), that evaluates trading strategies in a statistically significant way regardless of the technologies used for building them. To be meaningful, an evaluation needs to cover a long enough time to avoid false positive results, which are unreplicable in the long term, as well as prevent data leakage from evaluation to training data and other pitfalls (Section III-A). There are four steps to our method.

1) Define Parameters:
   a) Financial parameters: set of assets, horizon of investment, and amount of capital
   b) Strategy parameters: according to picked approach and use available training data

2) Verify conditions:
   a) Verify if the traded volume may lead to an underestimate of the strategy’s impact on the market
   b) Verify if the strategy uses shorting, check if transaction costs have been tallied sufficiently

3) Refine parameters using historical prices (backtest) until a reasonable performance has been achieved

4) Execute strategy using real time simulation until achieving meaningful results (detailed in Section IV-A)

The first two phases, which define parameters and verify conditions, are quite direct, especially the first one. There are few details to keep in mind: Underestimating impact of own trading. That pitfall is perhaps the least likely to happen, it is not even listed by [26]. If your maximum traded volume is much smaller less than 10 base points than the average traded volume of the asset and the strategy is not used by others the impact of one’s own trading is probably small enough to be unconsidered. To account for Survivor bias, you need to state your criteria for selecting the target assets in a clear way and verify that if using that criteria and only information available in the beginning of the simulated time, you would select a problematic asset (delisting, bankruptcy, etc). To ensure that Short operation costs have been covered sufficiently you may use the maximum costs observed in the market for short operations similar for the short operations executed by your strategy.

The third phase, refine parameters, uses the so-called backtest to fit parameters in order to improve performance. It should be carried out without any data leakage from test to training set by including an embargo period as discussed in Section III-A. However, it is still possible to have some data leakage due to back-fill corrections or from avoiding assets that presented bad performance or negotiation interruptions (survivor bias). In fact, the best way to avoid data leakage of any kind and verify that the model represents patterns that are still valid in the market is by using real time simulation. That is the fourth step of STSE method.

Executing the strategy using real time simulation prevents data leakage, since each data point is out-of-sample and used for evaluation as soon as available. One could argue that a ‘reasonable amount of time’ is subjective. However, STSE uses a probabilistic approach that establishes the minimum time for passing a performance threshold with a given level of certainty. That amount of time is highly dependent on the strategy’s observed performance. Great strategies may show skill briefly, while others require much more time and some others will never pass a given performance threshold even after many
months or even years. In that case, the strategy is not a good option for the current market situation. Nevertheless since the financial environment is non-stationary, it is not possible to assure that such strategy will never present good performance in the future.

A. STSE Fundamentals

STSE is based on the idea that you need to observe a trading strategy avoiding data leakage and mitigating over-fitting risk for a long enough time. The amount of time is highly dependent on the trading strategy performance itself and the target performance threshold. This threshold is defined in terms of a minimum desired Sharpe Ratio for the trading strategy. Naturally, the observed Sharpe Ratio for any given trading strategy is highly dependent on market conditions, but it is possible to define confidence levels where the observed SR is equal or above the threshold even for non-Normal returns. In fact, Bailey and Prado [36] showed that the probability that the observed Sharpe Ratio (SR) will be greater than a given Sharpe Ratio threshold (SR*), can be given by Equation 2. In Equation 2, Z is the CDF (cumulative distribution function) of the Standard Normal distribution and γ₁ and γ₄ are the skewness and kurtosis, respectively.

\[
\text{Prob}[\hat{SR} > \text{SR}^*] = Z \left[ \frac{(\text{SR} - \text{SR}^*) \sqrt{n - 1}}{\sqrt{1 - \gamma_3 \text{SR} + \frac{\gamma_4 - 1}{4} \text{SR}^2}} \right] \tag{2}
\]

Naturally, the estimation of the Sharpe Ratio is subject to significant errors, so the question becomes: "how long should a track record be in order to have statistical confidence that its Sharpe Ratio is above a given threshold?". For a level of certainty α, it is shown that the number of observations \(n^*\) in the track record should be no less than:

\[
n^* = 1 + \left[ 1 - \frac{1}{4} \gamma_3 \text{SR} + \frac{\gamma_4 - 1}{4} \text{SR}^2 \right] \left( \frac{Z_\alpha}{\text{SR} - \text{SR}^*} \right)^2 \tag{3}
\]

Equation 3 shows that the exact number of observations will increase as the Sharpe Ratio threshold (\(\text{SR}^*\)) gets closer to the observed Sharpe Ratio (\(\hat{\text{SR}}\)). The more skewed the returns are, or, in other words, the greater the fat tails, the longer the track records must be. An increase in our required level of confidence also results in the need for longer track records. The authors also point that, given the assumptions, the number of required observations (\(n^*\)) should never be less than 30, which means 2.5 years of monthly data, or 0.5769 years of weekly data (52 observations/year), or 0.119 years of daily data (252 observations/year) [36]. In Appendix A, we provide an implementation of Equation 2 that calculates the probability of the Sharpe Ratio (PSR) being greater than a given threshold and another implementation of Equation 3 given the observed time series of returns of the trading algorithm and the desired SR threshold (\(\text{SR}^*\)). Both implementation are in MQL 5, the programming language used in the Metatrader platform, which is very similar to C++. In order to illustrate the use of Equations 2 and 3 we adapted Figure 1 from the original paper [36] and present the results achieved by the authors. Observe that these equations were applied to 33 HFR indices. HFR indices which are made up of funds managed by human experts, not by autonomous trading strategies. In Figure 1, note that only 9 indices presented investment skill over an annualized Sharpe Ratio of 0.5 with a 95% confidence level (cells in black, PSR(0.5) column), but almost all (29 out of 33) substantiated investment skill over an annualized Sharpe Ratio of zero (PSR(0)). It is also possible to observe the minimum track record (mTRL) required for each case, considering a Sharpe Ratio of 0.5 (mTRL(0.5)) or zero (mTRL(0)). These mTRL are reported in years, considering monthly observations.

It is important to note that the duration of the track record is dependent on strategy performance itself. So, a fixed duration cannot be stated for any trading strategy. However, it should never shorter than 30 time slices, in our implementation that means at least 30 business days.

We use the Sharpe Ratio as a measure for strategy performance adjusted to risk, despite the criticism, as we discussed in Section III-B. It is possible to estimate the probability that the Sharpe Ratio of an autonomous trading strategy is above a given threshold using Equation 2. It is important to note that the higher the threshold the longer the required track record may be. The required track record can be estimated using Equation 3. Furthermore, a higher level of certainty also may require a longer track record. Both equations are implemented in the STSE module, which is open source code and publicly available.
bet sizing is a complex question in itself. It may be called should be allocated to an autonomous trading strategy is a problem of estimating the optimal amount of capital that promises safety of capital and adequate return, while the second lacks these features. Investors need to be sure which investment should be reduced if there is a performance degradation. Even with those procedures, it is still true that past returns do not guarantee future ones. As stated by Graham and co-authors of his famous book, investment and speculative operations are different because the first are based on thorough analysis that promises safety of capital and adequate return, while the second lacks these features. Investors need to be sure which operation they are implementing through their trading robots. The problem of estimating the optimal amount of capital that should be allocated to an autonomous trading strategy is a complex question in itself. It may be called bet sizing. We intend to address this in the future, but it is out of the scope of this paper.

The STSE method may be used in strategies by yourself or by third parties. However, if dealing with a strategy created by a third party, there may be additional problems that we address in Section IV-C.

C. Third-party Strategy Evaluation

The evaluation of a strategy developed by a third-party presents the same challenges a strategy developed personally but with some additional pitfalls. As the trading strategy and its parameters were defined by others, the investor does not know if it suffers from some kind of data leakage (intentional or not) from evaluating data to training data nor does she know how many configurations were tried in order to define the parameter. As pointed out by Bailey and others, if one does not know how many trials or parameter configurations were tested, one cannot assess the risk of overfitting and that risk grows with the number of trials. Additionally, any developer is tempted to try the maximum possible number of configurations in order to find the one with the best performance. In fact, even if the developer reveals the number of trials and declares that all required precautions against data leakage were observed, there is still an incentive for him to lie, because the reported performance would be more credible.

However, even if the number of tested configurations is unknown and the robot’s strategy is a ‘black box’, i.e., there is no access to strategy’s source code it is still possible to evaluate the strategy’s performance using STSE method. To do so, there must be access to the strategy’s trading decisions using reliable live data. This can be done using a trading platform that allows simulated operation with real time data. There are several available like Metatrader, Interactive Brokers, and Alpaca or others. It is beyond the scope of this paper to discuss these platforms, but one should be able to use STSE with any platform, since you can extract the equity time series marked to market. The STSE method can also be used no matter the number of assets managed by the strategy, as long as it is possible to obtain the equity time series reliably.

In Section V we apply the STSE method to several autonomous trading strategies with different financial parameters. We have used the Metatrader platform to perform many experiments. The evaluated strategies include five for which we had access to the source code and five for which we did not have access to source code. They were downloaded from the Metatrader platform and tested. The tested scenarios and results are discussed in the next Section.

V. STSE Results and Discussion for Several Strategies

Most of the autonomous trading strategies that we read about and almost all trading strategies available in the Metatrader platform deal with only one asset at time. This is true, despite the fact there are very good reasons for creating strategies that deal with several assets and try to explore their complementarity in terms of risk and return, which is widely used in non-autonomous trading strategies. We evaluate autonomous trading strategies using one asset per scenario in this Section to reflect the literature. However, that is not a limitation of STSE method, which can be used for evaluating portfolio strategies.

We present the results achieved by applying the proposed method, Significant Trading-Strategy Evaluation (STSE), to many different trading strategies and assets in this Section. A scenario is defined by an autonomous strategy, a set of assets and a time period. We tested more than 300 scenarios. These many scenarios were split into two groups. The first group (FX) focused on Foreign Exchange assets, while the second group (SX) dealt with Stock Exchange assets. We used two different thresholds for the minimum Sharpe Ratio: zero and 0.1, both with a level of confidence of 95%. These values seem very reasonable but one may set the values in STSE setup that best represents individual preferences. In FX scenarios, we
used five different strategies, four based on technical signals and one based on Machine Learning. We had access to the Source code. However, in SX scenarios we evaluated nine strategies and five of them were developed by third-parties and we did not have access to their source code. We were still able to evaluate them using STSE method as discussed in Section [IV-C] The two groups of scenarios (FX and SX) and their respective results are described next.

A. Foreign Exchange (FX) Scenarios

In the context of Foreign Exchange assets (FX), we executed twenty scenarios composed of five strategies and five different assets, over the period of three years of trading data (Jan-1-2017 to Dec-31-2019). The assets were CFD (Contract for Difference) instruments [40]. CFD is a contract that enables two parties to trade based on the difference between the entry and closing prices. In the case of foreign exchange pairs, a trader may bet that one currency will rise (or fall) in comparison with on other. If she is right, she will profit, and suffer a loss in she is wrong [41]. The foreign exchange pairs are listed in Table I and one of them (BTCUSD) is based on a cryptocurrency (Bitcoin) against the US dollar.

We evaluated four well-known strategies: moving average convergence/divergence (MACD) [42] (pg. 134,150), Moving Average (MAMA) [43] (pg. 109-132), Moving Average Parabolic Stop and Reverse (MAPS) [43] and another version of MAPS with size optimization, that we will call MAPS2 from now on. We also used one Trading Strategy based on Random Forest [44], that we refer as RFOR from now on. All of them use technical indicators and have some different implementations, each with some parameters. We have used the implementation available in the Metatrader5 platform [14] with the default parameters defined in Appendix A. We present each one of the scenarios results next and discuss them in the following subsection.

| #  | Symbol | Description |
|----|--------|-------------|
| 1  | EURUSD | Euro vs US dollar |
| 2  | USDJPY | US dollar vs Japan Yen |
| 3  | EURJPY | Euro vs Japan Yen |
| 4  | GBPUSD | Great Britain Pound vs US dollar |
| 5  | BTCUSD | Bitcoin vs US dollar |

**TABLE I**

**SELECTED ASSETS FROM THE MOST TRADED ASSETS IN FOREIGN EXCHANGE (FOREX) AND CRYPTOCURRENCY (BITCOIN) MARKET.**

The results show that none of the tested autonomous traders was able to surpass the 0.5 threshold, but some of them were able to surpass the zero threshold. However, as discussed in Section [III] and presented in Figure 1 some human asset management professionals are able to overcome the 0.5 threshold. In fact, nine out of 33 HFR indices surpassed the 0.5 Sharpe Ratio threshold as presented in [36]. We also tested an easier threshold (0.1) for the selected autonomous, but once again none of them were able to surpass it.

Therefore, we can state that none of the tested strategies showed skill in any of the assets comparable to good hu-
man experts. By significant skill, we mean that the strategy performance were above the threshold (0 or 0.1) according to the case, with 95% level of confidence and mTRL equal or shorter than the simulated period (3 years, in this case). That indicates that creating successful trading strategies may be a really hard task. However, two strategies (MAMA and MAPS) were able to surpass the zero threshold. Even though it is the easiest barrier, it is not completely meaningless. As shown in Figure 1, four out of the 33 HFR indices were not able to surpass the zero threshold. RFOR performance was particularly bad. That fact should not be interpreted as a signal that AI-based strategies are not promising, because we used the standard implementation presented in [24], which was focused on stock markets rather than FX markets. Furthermore, if we had changed some hyper-parameters or included more information, such as simple moving averages, it is likely its performance would have been enhanced. In fact, we included RFOR to show that the STSE method could also be used with AI-based strategies.

We also executed the tests for a much shorter period of time (only six months, from Jul-1, 2019 to Dec-31, 2019). The results are shown below in Tables VII, VIII, IX, X, and XI. It is possible to observe that MAPS and MAMA strategies performed above the threshold (0 or 0.1) according to the case, with 95% level of confidence and mTRL equal or shorter than the simulated period (3 years, in this case). That indicates that creating successful trading strategies may be a really hard task. However, two strategies (MAMA and MAPS) were able to surpass the zero threshold. Even though it is the easiest barrier, it is not completely meaningless. As shown in Figure 1, four out of the 33 HFR indices were not able to surpass the zero threshold. RFOR performance was particularly bad. That fact should not be interpreted as a signal that AI-based strategies are not promising, because we used the standard implementation presented in [24], which was focused on stock markets rather than FX markets. Furthermore, if we had changed some hyper-parameters or included more information, such as simple moving averages, it is likely its performance would have been enhanced. In fact, we included RFOR to show that the STSE method could also be used with AI-based strategies.

D. Stock Exchange (SX) Results

The SX results are presented in Tables XIV and XV. The black cells in these tables show the situations where some strategy failed to pass the zero threshold for a given asset. We checked if those strategies with performance above zero, were able to pass the 0.1 threshold and we also checked if some of them could pass an easier limit (0.05). The results are presented in Table XVI. Just as was observed in FX scenarios, we executed 270 scenarios using these nine strategies and thirty assets. They were executed using real data from Jan 01, 2018 to Dec 31, 2019.

C. Stock Exchange (SX) Scenarios

In addition to crypto and traditional currency scenarios, we have also used STSE to evaluate autonomous trading strategies operating in stock exchange markets. We selected thirty (30) stocks from 3 different stock exchanges, that are listed in Table XIV. Nine autonomous trading strategies were used; four of them were also used FX Scenarios (MACD, MAMA, MAPS and MAPS2) and the five others developed by third parties and available for free download in the Metatrader platform in binary distributions without access to their source codes. The complete list is available in Table XII. In the stock exchange scenarios, we executed 270 scenarios using these nine strategies and thirty assets. They were executed using real data from Jan 01, 2018 to Dec 31, 2019.

A. Black Cell Indicates a Situation Where the Strategy Surpasses the Proposed Sharpe Ratio Threshold: 0 or 0.1

B. Black Cell Indicates a Situation Where the Strategy Surpasses the Proposed Sharpe Ratio Threshold: 0 or 0.1

C. Black Cell Indicates a Situation Where the Strategy Surpasses the Proposed Sharpe Ratio Threshold: 0 or 0.1

D. Black Cell Indicates a Situation Where the Strategy Surpasses the Proposed Sharpe Ratio Threshold: 0 or 0.1

E. Black Cell Indicates a Situation Where the Strategy Surpasses the Proposed Sharpe Ratio Threshold: 0 or 0.1

F. Black Cell Indicates a Situation Where the Strategy Surpasses the Proposed Sharpe Ratio Threshold: 0 or 0.1

G. Black Cell Indicates a Situation Where the Strategy Surpasses the Proposed Sharpe Ratio Threshold: 0 or 0.1

H. Black Cell Indicates a Situation Where the Strategy Surpasses the Proposed Sharpe Ratio Threshold: 0 or 0.1

I. Black Cell Indicates a Situation Where the Strategy Surpasses the Proposed Sharpe Ratio Threshold: 0 or 0.1

J. Black Cell Indicates a Situation Where the Strategy Surpasses the Proposed Sharpe Ratio Threshold: 0 or 0.1

K. Black Cell Indicates a Situation Where the Strategy Surpasses the Proposed Sharpe Ratio Threshold: 0 or 0.1

L. Black Cell Indicates a Situation Where the Strategy Surpasses the Proposed Sharpe Ratio Threshold: 0 or 0.1

M. Black Cell Indicates a Situation Where the Strategy Surpasses the Proposed Sharpe Ratio Threshold: 0 or 0.1

N. Black Cell Indicates a Situation Where the Strategy Surpasses the Proposed Sharpe Ratio Threshold: 0 or 0.1

O. Black Cell Indicates a Situation Where the Strategy Surpasses the Proposed Sharpe Ratio Threshold: 0 or 0.1

P. Black Cell Indicates a Situation Where the Strategy Surpasses the Proposed Sharpe Ratio Threshold: 0 or 0.1

Q. Black Cell Indicates a Situation Where the Strategy Surpasses the Proposed Sharpe Ratio Threshold: 0 or 0.1

R. Black Cell Indicates a Situation Where the Strategy Surpasses the Proposed Sharpe Ratio Threshold: 0 or 0.1

S. Black Cell Indicates a Situation Where the Strategy Surpasses the Proposed Sharpe Ratio Threshold: 0 or 0.1

T. Black Cell Indicates a Situation Where the Strategy Surpasses the Proposed Sharpe Ratio Threshold: 0 or 0.1

U. Black Cell Indicates a Situation Where the Strategy Surpasses the Proposed Sharpe Ratio Threshold: 0 or 0.1

V. Black Cell Indicates a Situation Where the Strategy Surpasses the Proposed Sharpe Ratio Threshold: 0 or 0.1

W. Black Cell Indicates a Situation Where the Strategy Surpasses the Proposed Sharpe Ratio Threshold: 0 or 0.1

X. Black Cell Indicates a Situation Where the Strategy Surpasses the Proposed Sharpe Ratio Threshold: 0 or 0.1

Y. Black Cell Indicates a Situation Where the Strategy Surpasses the Proposed Sharpe Ratio Threshold: 0 or 0.1

Z. Black Cell Indicates a Situation Where the Strategy Surpasses the Proposed Sharpe Ratio Threshold: 0 or 0.1

Table V

Results for the five strategies for asset GBPUSD in three years. A black cell indicates a situation where the strategy surpasses the proposed Sharpe Ratio threshold: 0 or 0.1

Table VI

Results for the five strategies for asset BTCUSD in three years. A black cell indicates a situation where the strategy surpasses the proposed Sharpe Ratio threshold: 0 or 0.1

Table VII

Results for the five strategies for asset EURUSD in six months. A black cell indicates a situation where the strategy surpasses the proposed Sharpe Ratio threshold: 0 or 0.1

Table VIII

Results for the five strategies for asset EURJPY in six months. A black cell indicates a situation where the strategy surpasses the proposed Sharpe Ratio threshold: 0 or 0.1

Table IX

Results for the five strategies for asset USDJPY in six months. A black cell indicates a situation where the strategy surpasses the proposed Sharpe Ratio threshold: 0 or 0.1

Table X

Results for the five strategies for asset JPYEUR in six months. A black cell indicates a situation where the strategy surpasses the proposed Sharpe Ratio threshold: 0 or 0.1

Table XI

Results for the five strategies for asset JPYUSD in six months. A black cell indicates a situation where the strategy surpasses the proposed Sharpe Ratio threshold: 0 or 0.1

Table XII

Complete list of strategies.
### TABLE IX
RESULTS FOR THE FIVE STRATEGIES FOR ASSET USDJPY IN SIX MONTHS. A BLACK CELL INDICATES A SITUATION WHERE THE STRATEGY SURPASSES THE PROPOSED SHARPE RATIO THRESHOLD: 0 OR 0.1.

| Strategy | MACD | MAMA | MAPS | MAPS2 | RFOR |
|----------|------|------|------|-------|------|
| Return   | -0.000 | 0.000 | 0.000 | 0.000 | -98.774 |
| Sharpe Ratio | -0.014 | 0.051 | 0.051 | 0.025 | 0.003 |
| PSR(0)   | 0.214 | 0.999 | 0.999 | 0.920 | 0.553 |
| mTRL(0)  | 13.581 | 0.149 | 0.149 | 4.697 | -98.774 |
| PSR(0.1) | 0.000 | 0.000 | 0.000 | 0.000 | 0.553 |
| mTRL(0.1) | 0.225 | 0.207 | 0.214 | 0.542 | 0.287 |

### TABLE X
RESULTS FOR THE FIVE STRATEGIES FOR ASSET GBPUSD IN SIX MONTHS. A BLACK CELL INDICATES A SITUATION WHERE THE STRATEGY SURPASSES THE PROPOSED SHARPE RATIO THRESHOLD: 0 OR 0.1.

| Strategy | MACD | MAMA | MAPS | MAPS2 | RFOR |
|----------|------|------|------|-------|------|
| Return   | -0.000 | 0.000 | 4.139 | -0.637 | 130.384 |
| Sharpe Ratio | -0.004 | 0.027 | 0.017 | -0.022 | 0.003 |
| PSR(0)   | 0.407 | 0.999 | 0.977 | 0.072 | 0.934 |
| mTRL(0)  | 155.402 | 0.791 | 2.124 | 4.490 | 0.287 |
| PSR(0.1) | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| mTRL(0.1) | 0.225 | 0.207 | 0.214 | 0.542 | 0.287 |

### TABLE XI
RESULTS FOR THE FIVE STRATEGIES FOR ASSET BTCUSD IN SIX MONTHS. A BLACK CELL INDICATES A SITUATION WHERE THE STRATEGY SURPASSES THE PROPOSED SHARPE RATIO THRESHOLD: 0 OR 0.1.

| Strategy | MACD | MAMA | MAPS | MAPS2 | RFOR |
|----------|------|------|------|-------|------|
| Return   | 0.070 | 0.010 | -0.010 | 0.000 | -192.8 |
| Sharpe Ratio | 0.090 | 0.000 | 0.000 | -1.00 | -0.34 |
| PSR(0)   | 0.947 | 0.514 | 0.491 | NaN | 0.0008 |
| mTRL(0)  | 0.100 | 268.800 | 733.800 | NaN | 1.5 |
| PSR(0.1) | 0.394 | 0.137 | 0.125 | NaN | 0.000 |
| mTRL(0.1) | 4.900 | 0.300 | 0.300 | NaN | 0.1 |

### TABLE XII
SELECTED TRADING STRATEGIES FOR STOCK EXCHANGE SCENARIOS. THE STRATEGIES MARKED (*) WERE DOWNLOADED AS BINARY CODE FROM METATRADER PLATFORM, WITHOUT ACCESS TO SOURCE CODE.

| # | Strategy |
|---|----------|
| 1 | MACD     |
| 2 | MAMA     |
| 3 | MAPSR    |
| 4 | MAPS2    |
| 5 | Dark Venus MT5 (*) |
| 6 | Neural Networks 2 Moving Averages (*) |
| 7 | Dark Moon MT5 (*) |
| 8 | Hyper Trade Global (*) |
| 9 | CAP Random Trader EA MT5 (*) |

### TABLE XIII
LIST OF ASSETS USED IN SX SCENARIOS

| # | Stock Ex. | Asset ID | Company |
|---|-----------|----------|---------|
| 1 | Nasdaq    | AAPL     | Apple Inc. |
| 2 | Nasdaq    | AMD      | Advanced Micro Devices, Inc. |
| 3 | Nasdaq    | MSFT     | Microsoft |
| 4 | Nasdaq    | FB       | Facebook |
| 5 | Nasdaq    | AAL      | American Airlines Group, Inc. |
| 6 | Nasdaq    | TSL      | Tesla, Inc. |
| 7 | Nasdaq    | TLRY     | Tilray, Inc. |
| 8 | Nasdaq    | V        | Visa Inc. |
| 9 | Nasdaq    | INJ      | Johnson & Johnson |
| 10 | Nasdaq    | MU       | Micron Technology, Inc. |
| 11 | NYSE      | BABA     | Alibaba Group |
| 12 | NYSE      | SQ       | Square Inc. Cl A |
| 13 | NYSE      | BA       | Boeing Co. |
| 14 | NYSE      | NIO      | NIO Inc. |
| 15 | NYSE      | AMC      | AMC Entertainment Inc. |
| 16 | NYSE      | JPM      | JPMorgan Chase |
| 17 | NYSE      | ABEV     | Ambev |
| 18 | NYSE      | SHOP     | Shopify Inc. Cl A |
| 19 | NYSE      | AA       | Alcoa corp. |
| 20 | NYSE      | DI       | Didi Inc. |
| 21 | B3        | VALE3    | Vale |
| 22 | B3        | ITUB4    | Itaú |
| 23 | B3        | PETR4    | Petrobras |
| 24 | B3        | BBDC4    | Bradesco |
| 25 | B3        | B3SA3    | B3 |
| 26 | B3        | PETR3    | Petrobras |
| 27 | B3        | ABEV3    | Ambev |
| 28 | B3        | OIBR3    | Ot |
| 29 | B3        | ITSA4    | Itausa |
| 30 | B3        | BBAS3    | Banco do Brasil |
| Asset ID | MACD | MAMA | MAPS | MAPS2 |
|----------|------|------|------|-------|
|          | PSR(0) | mTRL(0) | PSR(0) | mTRL(0) | PSR(0) | mTRL(0) | PSR(0) | mTRL(0) |
| AAPL     | 0.362 | 8.8  | 0     | 0.1    | 0      | 0.1     | 0.004  | 0.2     |
| AMD      | 0.218 | 1.9  | 0     | 0.1    | 0.218  | 1.9     | 0.033  | 0.3     |
| MSFT     | 0.199 | 1.6  | 0     | 0.1    | 0      | 0.1     | 0      | 0.1     |
| FB       | 0.375 | 11.2 | 0.012 | 0.2    | 0.012  | 0.2     | NaN    | NaN     |
| AAL      | 0.455 | 88.8 | 0.434 | 40.8   | 0.434  | 40.8    | 0.044  | 0.4     |
| TSLA     | 0.136 | 0.5  | 0.988 | 0.100  | 0.981  | 0.3     | 0.001  | 0.1     |
| TLYR     | 0.129 | 0.5  | 0.918 | 0.3    | 0.918  | 0.3     | 0.953  | 0.200   |
| V        | 0.308 | 4.2  | 0     | 0.1    | 0      | 0.1     | 0.004  | 0.2     |
| JNJ      | 0.668 | 6.1  | 0.013 | 0.2    | 0.013  | 0.2     | 0      | 0.1     |
| MU       | 0.422 | 29.3 | 0.106 | 0.7    | 0.106  | 0.7     | 0.001  | 0.1     |
| BABA     | 0.348 | 3.9  | 0.009 | 0.1    | 0.009  | 0.1     | 0      | 0       |
| SQ       | 0.725 | 1.7  | 0.144 | 0.5    | 0.144  | 0.5     | 0.001  | 0.1     |
| BA       | 0.625 | 11.2 | 0.075 | 0.6    | 0.075  | 0.6     | 0.002  | 0.1     |
| NIO      | 0.024 | 0.2  | 0.473 | 125.5  | 0.473  | 125.5   | 0.575  | 16.7    |
| AMC      | 0.491 | 1154.9 | 0.505 | 3426.3 | 0.505 | 3426.3 | 0.002 | 0.1     |
| JPM      | 0.695 | 4.4  | 0.003 | 0.2    | 0.003  | 0.2     | 0      | 0.1     |
| ABEV     | 0.297 | 2.1  | 0.015 | 0.1    | 0.015  | 0.1     | 0.713  | 1.9     |
| SHOP     | 0.701 | 2.1  | 0     | 0.1    | 0      | 0.1     | 0.982  | 0.100   |
| AA       | 0.8   | 0.8  | 0.163 | 0.6    | 0.163  | 0.6     | 0.417  | 13.5    |
| DIS      | 0.608 | 15.3 | 0.002 | 0.1    | 0.002  | 0.1     | 0.009  | 0.2     |
| VALE3    | 0.601 | 20.4 | 0.304 | 5.1    | 0.02   | 0.3     | NaN    | NaN     |
| ITUB4    | 0.567 | 46.8 | 0.438 | 54.9   | 0.4    | 21      | NaN    | NaN     |
| PETR4    | 0.572 | 40.7 | 0.253 | 3      | 0.299  | 4.8     | NaN    | NaN     |
| BBDC4    | 0.609 | 17.6 | 0.281 | 4      | 0.259  | 3.2     | NaN    | NaN     |
| B3SA3    | 0.5   | 1381.322 | 0.262 | 3.3    | 0.285  | 4.2     | NaN    | NaN     |
| PETR3    | 0.554 | 72.2 | 0.337 | 7.6    | 0.259  | 3.2     | NaN    | NaN     |
| ABEV3    | 0.569 | 43.9 | 0.546 | 98.4   | 0.848  | 1.3     | 0.002  | 0.2     |
| OIBR3    | 0.585 | 28.8 | 0.669 | 7      | 0.643  | 10      | NaN    | NaN     |
| ITSA4    | 0.48  | 546.4 | 0.37  | 12.2   | 0.007  | 0.2     | NaN    | NaN     |
| BBAS3    | 0.601 | 20.6 | 0.423 | 35.2   | 0.404  | 22.9    | NaN    | NaN     |

**TABLE XIV**

Four strategies results for thirty SX assets. A black cell indicates a situation where the strategy surpassed the zero Sharpe ratio threshold.
| Asset ID | Dark Venus MT5 | Neural Networks 2 Moving Averages | Dark Moon MT5 | Hyper Trade Global | CAP Random Trader EA MT5 |
|---------|----------------|----------------------------------|---------------|-------------------|--------------------------|
| AAPL    | 0.019          | 0.334                            | 0.035         | 131.2             | 0.382                    |
| AMD     | 0.163          | 40.6                             | 0.486         | 867.2             | 0.1                      |
| MSFT    | 0.471          | 215.8                            | 0.433         | 819.9             | 0.1                      |
| FB      | 0.557          | 55.6                             | 0.2           | 71.3              | 0.2                      |
| AAL     | 0.536          | 142.5                            | 0.478         | 1033.7            | 0.434                    |
| TSLA    | 0.803          | 9170.3                           | 0.237         | 143.8             | 0.998                    |
| TLYR    | 0.562          | 24.8                             | 0.51           | 974.4             | 0.918                    |
| V       | 0.522          | 360.3                            | 0.445         | 1242.2            | 1.0                      |
| JNJ     | 0.204          | 1.7                              | 0.559         | 1258.6            | 0.2                      |
| MU      | 0.469          | 21.8                             | 0.466         | 161.1             | 0.992                    |
| BA      | 0.368          | 5.2                              | 0.421         | 323.3             | 0.991                    |
| SQ      | 0.532          | 90.4                             | 0.555         | 481.5             | 0.856                    |
| NIO     | 0.597          | 9.8                              | 0.401         | 147.7             | 0.473                    |
| AMC     | 0.549          | 39.6                            | 0.532         | 83.9              | 0.491                    |
| JPM     | 0.529          | 221.1                            | 0.568         | 1036.2            | 0.997                    |
| ABEV    | 0.435          | 22.3                            | 0.554         | 85                | 0.015                    |
| SHOP    | 0.646          | 4.3                              | 0.605         | 1016.9            | 0.1                      |
| AA      | 0.528          | 121.7                            | 0.532         | 106.1             | 0.822                    |
| DIS     | 0.109          | 0.8                              | 0.556         | 670.6             | 0.998                    |
| VALE3   | 0.009          | 0.2                              | 0.574         | 3.4               | 0.388                    |
| ITUB4   | 0.975          | 0.400                            | 0.526         | 6.2               | 0.919                    |
| PETR4   | 0.046          | 4.3                             | 0.606         | 18.3              | 0.699                    |
| BBDC4   | 0.525          | 127.8                            | 0.3           | 132              | 0.583                    |
| PETR3   | 0.927          | 0.600                            | 0.522         | 12.7              | 0.5                      |
| B3AS1   | 0.461          | 141.4                            | 0.552         | 21.5              | 0.988                    |

**TABLE XV**

Five strategies results for thirty SX assets. A black cell indicates a situation where the strategy surpassed the zero Sharpe ratio threshold.
E. Discussion of Results

In the fifty FX scenarios, we observed that only eleven strategies managed to produce Sharpe Ratio above the zero threshold, and only two strategies (MAMA,MAPS) were able to do so in the six month period. As expected, it is harder for a strategy to prove skill in shorter time periods. None of the strategies were able to pass the 0.1 threshold. The results were similar in the 270 SX scenarios. Only thirteen out of 270 scenarios showed strategies passing the zero Sharpe Ratio threshold. We tested these thirteen strategies against the 0.1 threshold and also against an easier target of 0.05. None of them were able to pass these thresholds, not even the easier 0.05 limit (see Table XVI).

One may argue that such limits are too high. However, it has been observed that professional human portfolio managers are able to pass a much harder threshold, 0.5 [30]. In fact, this study, discussed in Section III showed that twenty-nine (29) out of thirty-three (33) funds were able to pass the zero threshold. Furthermore, nine of them passed the much harder 0.5 threshold. These results seem to indicate that autonomous trading strategies still have a long road ahead to match a human expert’s performance. However, it is clear to us that using just one asset for a strategy contributed to the poor performance, because it does not allow risk mitigation through diversification, but as discussed in Section III-A it is common to see autonomous trading strategy focus only on return and disregarding risk mitigation through diversification. The simple use of optimization algorithms, such as quadratic programming [25], and trading with many assets would likely help to improve results. However, it is not clear to us that this would be enough to close the big gap observed among the tested autonomous strategies and experts’ performance.

Naturally, it is possible that autonomous trading strategies dealing with many assets would present much better performance. Nevertheless, the STSE method would still be able to evaluate such multi-asset strategies without any change, as discussed in Section IV-C.

The STSE method was tested on several strategies and assets (foreign currencies, cryptocurrency, and stocks). The strategies were not selected based on high expectation about their performance. They were selected based on the use of several different techniques and availability. We did not engage in a long process to refine the strategy’s parameters, which could have helped to improve their performance. Some strategies did surpass the zero threshold, which means that they have an expected Sharpe Ratio higher than zero with 95% level of confidence and that the observed track record is equal or above the minimum required track record (mTRL). The STSE method open source implementation is freely available in [19]. In Appendix A we detail all simulation setups, and Strategy’s parameters used in the tested scenarios. As discussed in Section V some of the tested strategies are not open source, but all allow binary download free of charge. Using the STSE open source implementation, the same strategies and setup specified in Appendix A it is possible to reproduce the results achieved and presented in this paper. In the next and final Section, we provide a brief summary of this study’s contribution.

VI. Conclusions and Future Work

In this paper, we presented a method to evaluate autonomous trading strategies regardless of the techniques used for their construction. This method, called STSE, was able to separate skillful trading strategies from those with no skill (expected Sharpe Ratio below zero). We also provided an open source implementation which we believe can be useful for autonomous trading strategy developers. The STSE implementation is freely available at GitHub [19].

We have used STSE to evaluate nine different strategies and thirty five assets including currency, cryptocurrency and stocks. The autonomous strategies include some based on technical analysis (Moving average, MACD and others), machine learning techniques (Random Forest, Neural networks) and some proprietary strategies developed by third-parties that we did not have much information about. We were still able to apply our STSE method to each one of them. These results indicate that the analyzed strategies are far from presenting performance levels comparable to human experts. We also believe that many claims about great performance achieved by some proprietary autonomous strategies may be due to overfitted strategies, since it is hard to avoid data leakage and other common problems. Furthermore, it is important to note that the saying “past returns do not guarantee future returns”, is also valid for autonomous trading. Finally, we observe that new approaches, such as deep reinforcement learning and the use of alternative data [28], may be really useful in building better autonomous strategies. However, significant evaluation processes will be needed to tell if you built a really successful strategy or you are just fooling yourself.

ACKNOWLEDGEMENTS

Paulo A.L. Castro is partially funded by CNPq (Brazil) Grant. No. 311838/2017-0.

REFERENCES

[1] Michael P. Wellman and Uday Rajan. Ethical issues for autonomous trading agents. Minds and Machines, 27(4):609–624, Dec 2017.
[2] Renato Oliveira and Adriano Pereira. Agente de negociação de ações utilizando aprendizado por reforço. In Proceeding of the Workshop of Artificial Intelligence Applied to Finance (WAIIF 2019), Sao Jose dos Campos, Brazil, 2019. WAIIF.
[3] Alexander Sherstov and Peter Stone. Three automated stock-trading agents: A comparative study. In Proceedings of the Agent Mediated Electronic Commerce (AMEC) Workshop - AAMAS 2004, New York, 2004.
[4] Paulo Andre Lima Castro and Jaime Sichman. Automated asset management based on partially cooperative agents for a world of risks. Applied Intelligence, 38:210–225, 2013.
[5] J. Paulin, A. Calinescu, and M. Wooldridge. Agent-based modeling for complex financial systems. IEEE Intelligent Systems, 33(2):74–82, Mar 2018.
[6] Leandro Anghinoni and Liang Zhao. Time series trend detection and forecasting using complex network topology analysis. In Proceeding of the Workshop of Artificial Intelligence Applied to Finance (WAIIF 2018), Sao Jose dos Campos, Brazil, 2018. WAIIF.
TABLE XVI

| TABLE XVI | The strategies that passed the zero threshold results for higher thresholds. A black cell indicates a situation where the strategy surpassed the 0.05 or 0.1 Sharpe ratio threshold. |
|----------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| **Asset ID** | **STRATEGY** | **PSR(0.05)** | **mTRL(0.05)** | **PSR(0.1)** | **mTRL(0.1)** |
| TSLA          | CAP Random Trader EA MT5 | 0.886 | 0.4 | 0.331 | 3.1 |
| V             | CAP Random Trader EA MT5 | 0.822 | 1.3 | 0.053 | 0.4 |
| BABA          | CAP Random Trader EA MT5 | 0.800 | 0.8 | 0.245 | 1.2 |
| JPM           | CAP Random Trader EA MT5 | 0.662 | 6.5 | 0.030 | 0.3 |
| DIS           | CAP Random Trader EA MT5 | 0.720 | 3.4 | 0.040 | 0.4 |
| ITSA4         | CAP Random Trader EA MT5 | 0.736 | 3.4 | 0.281 | 4.0 |
| TSLA          | MAMA                | 0.886 | 0.4 | 0.331 | 3.1 |
| AAL           | Neural Networks 2 Moving Averages | 0.439 | 48.6 | 0.019 | 0.3 |
| AMC           | Neural Networks 2 Moving Averages | 0.772 | 1.1 | 0.207 | 0.9 |
| ABEV          | Neural Networks 2 Moving Averages | 0.898 | 0.4 | 0.306 | 2.3 |
| AA            | Neural Networks 2 Moving Averages | 0.702 | 2.1 | 0.172 | 0.7 |
| ITUB4         | Dark Venus MT5      | 0.519 | 586.4 | 0.013 | 0.3 |
| ABEV3         | Dark Venus MT5      | 0.418 | 31.6 | 0.009 | 0.2 |
| SHOP          | MAPS2               | 0.722 | 1.7 | 0.176 | 0.7 |

[7] Rafael Silva Wagner and André Alves Portela Dos Santos. Forecasting the direction of high-frequency returns: An ensemble-tree application. In Proceeding of the Workshop of Artificial Intelligence Applied to Finance (WAI AF 2018), Sao Jose dos Campos, Brazil, 2018. WAI AF.

[8] Harish Subramanian, Subramanian Ramamoorthy, Peter Stone, and Benjamin J. Kuipers. Designing safe, profitable automated stock trading agents using evolutionary algorithms. In GECCO ‘06: Proceedings of the 8th annual conference on Genetic and evolutionary computation, pages 1777–1784, New York, NY, USA, 2006. ACM Press.

[9] Flavio Abdenour Elias Cavalcante-Filho and Rodrigo De-Losso. Machine learning applied to accounting variables yields the risk-return metrics of private company portfolios. In Proceeding of the Workshop of Artificial Intelligence Applied to Finance (WAI AF 2019), Sao Jose dos Campos, Brazil, 2019. WAI AF.

[10] Felipe Dias Paiva and Carolina Magda da Silva Roma. Métodos de deep learning aprendidos a candlestick como estratégia de investimento. In Proceeding of the Workshop of Artificial Intelligence Applied to Finance (WAI AF 2019), Sao Jose dos Campos, Brazil, 2019. WAI AF.

[11] Y. Hu and S. Lin. Deep reinforcement learning for optimizing finance portfolio management. In 2019 Amity International Conference on Artificial Intelligence (AICAI), pages 14–20, Feb 2019.

[12] Jimin Lee, Hayeong Koh, and Hi Jun Choe. Learning to trade in financial time series using high-frequency through wavelet transformation and deep reinforcement learning. Applied Intelligence, Feb 2021.

[13] Salvatore Carta, Andrea Corriga, Anselmo Ferreira, Alessandro Sebastian Poddà, and Diego Reforgiato Recupero. A multi-layer and multi-ensemble stock trader using deep learning and deep reinforcement learning. Applied Intelligence, 51(2):889–905, Feb 2021.

[14] Andrew R. Young. Expert Advisor programming for MetaTrader 5. Edgehill Pushishing, Nashville, TN, USA, 2018.

[15] Abraham Othman and Tuomas Sandholm. Inventory-based versus prior-based options trading agents. Algorithmic Finance, 1(95–121, 2011.

[16] John C. Hull, Options, futures, and other derivatives. Pearson Prentice Hall, Upper Saddle River, NJ [u.a.], 6. ed., pearson internat. ed edition, 2006.

[17] Monya Baker. Is there a reproducibility crisis? Nature, 533:452–454, 2016.

[18] Philip B. Stark. Before reproducibility must come prereproducibility. Nature, 557:613, 2018.

[19] Murilo Sibrao Bernardini and Paulo Andre Lima de Castro. Sto: Significant trading -strategy evaluation. https://github.com/paulo-al Castro/sto, 2021.

[20] Stuart Russell and Peter Norvig. Artificial Intelligence A Modern Approach Third Edition. Prentice Hall, Englewood Cliffs - NJ, 2013.

[21] Elena Andreou and Eric Ghysels. Structural Breaks in Financial Time Series, pages 839–870. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

[22] Yoshua Bengio Ian Goodfellow and Aaron Courville. Deep Learning. The MIT Press, Boston, 2016.

[23] David H. Bailey, Jonathan Borwein, Marcos Lopez de Prado, and Qiji Jim Zhu. Pseudo-mathematics and financial charlatanism: The effects of backtest overfitting on out-of-sample performance. Notices of the American Mathematical Society, 61:458–471, May 2014.

[24] Mark A. Hall Eibe Frank and Ian H. Witten. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, New York, 2016.

[25] Frank K. Reilly and Keith C. Brown. Investment Analysis & Portfolio Management. South-Western, Cengage Learning, New York, 2012.

[26] J. Jussa S. Wang A. Wang Luo Y., M. Alvarez and G. Rohal. Seven sins of quantitative investing, September 8 2014.

[27] Allan Timmermann Ryan Sullivan and Halbert White. Data-snooping, technical trading rule performance, and the bootstrap. The Journal of Finance, 54(5):1647–1691, Oct 1999.

[28] Marcos Lopez de Prado. Advances in Financial Machine Learning. Wiley, New York, 2018.

[29] Harry M. Markowitz. Portfolio selection. Journal of Finance, 7(1):77–91, 1952.

[30] Michael Greenberg. etoro launches its own forex social community – openbook, december 2010.

[31] R. Jesse McWaters. The future of financial services: How disruptive innovations are reshaping the way financial services are structured, provisioned and consumed, June 2015.

[32] Tom Rollinger and Scott Hoffman. Sortino ratio: A better measure of risk. Futures magazine, 1(1):40–42, 2 2013.

[33] Simone Farinelli, Manuel Ferreira, Damiano Rossello, Markus Thoeny, and Luisa Tibiletti. Beyond sharpe ratio: Optimal asset allocation using different performance ratios. Journal of Banking and Finance, 32(10):2057–2063, 10 2008.

[34] Fl. A.Sortino and H. J.Forsey. On the use and misuse of downside risk. Journal of Portfolio Management, 22(1):35–42, 1996.

[35] Aswath Damodaran. Applied Corporate Finance. Wiley, New York, NY, 2010.

[36] David H. Bailey and Lopez de Prado. The Sharpe ratio efficient frontier. Journal of Risk, 15:3–44, Feb 2012.

[37] Benjamin Graham. The Intelligent Investor. Harper Business, Revised Edition, New York, 2006.

[38] Interactive brokers. https://www.interactivebrokers.com/, 2021.

[39] Alpaca securities. https://alpaca.markets/, 2021.
[40] CORY MITCHELL. An introduction to contract for differences (cfds), 2021.
[41] Corporate Finance Institute. Contract for difference (cfd), 2021.
[42] Jefferey Katz and Donna McCormick. The Encyclopedia of trading strategies. McGraw-Hill, New York, 2000.
[43] Jr J. Welles Wilder. New Concepts in Technical Trading Systems. Trend Research, Greensboro, NC, 1978.
[44] Paulo Andre Lima de Castro. mt5b3: A framework for building autonomoustraders, January 2021.
**APPENDIX**

There are some parameters that need to be defined in order to execute simulations in the Metatrader platforms. In Table XVII you will find a list of them, with explanations and the values used in the simulations described in this paper. In Table XVIII we provide a list of the specific parameters for each Trading agent (or Expert Advisor). The Random Forest trader (RFOR) was used with the basic setup described in [44].

| Name               | Description                                                                 | Value                                                                 |
|--------------------|-----------------------------------------------------------------------------|----------------------------------------------------------------------|
| Expert             | Identification of the Trading Strategy implementation (called Expert Advisor - EA - in Metatrader) | Several EAs were used, see Section V-A for complete list |
| EA parameters      | Specific parameters for each EA                                             | see table XVIII for complete list                                     |
| Symbol             | Identification of the target asset                                         | Several assets were used, see Table I for complete list              |
| Period             | The timeframe for testing/optimization                                      | D1                                                                  |
| Deposit            | Initial Balance of EA                                                       | 100000                                                              |
| Currency           | Currency used                                                               | BRL (Brazilian currency, Real)                                      |
| Leverage           | Leverage for testing                                                       | 1:100                                                               |
| Model              | Tick mode                                                                   | 2 (Open price only)                                                 |
| ExecutionMode      | Execution of trade orders with a random delay                              | 0 (no delay)                                                       |
| Optimization       | Genetic optimization of EA parameters                                      | We used both possibilities with and without optimization           |
| OptimizationCriterion | Criteria used for optimization                                            | 0 (it means maximize balance value)                                 |
| FromDate           | Start date of the simulation                                               | Several values used see in Section V                                |
| ToDate             | End date of the simulation                                                 | Several values used see in Section V                                |
| ForwardMode        | Custom mode of forward testing                                             | 1 (Half of the testing period was used for parameter fitting)       |

**TABLE XVII**

**METATRADER TRADING AGENT (EXPERT ADVISOR) PARAMETERS**
| Expert Advisor | Parameter name                          | Value       |
|----------------|----------------------------------------|-------------|
| MACD           | Inp_Expert_Title                       | ExpertMACD  |
| MACD           | Inp_Signal_MACD_PeriodFast             | 12          |
| MACD           | Inp_Signal_MACD_PeriodSlow             | 24          |
| MACD           | Inp_Signal_MACD_PeriodSignal           | 9           |
| MACD           | Inp_Signal_MACD_TakeProfit             | 50          |
| MACD           | Inp_Signal_MACD_StopLoss               | 20          |
| MAMA           | Inp_Expert_Title                       | ExpertMAMA  |
| MAMA           | Inp_Signal_MA_Period                   | 12; 12; 1;120; N |
| MAMA           | Inp_Signal_MA_Shift                   | 6; 6; 1; 60; N |
| MAMA           | Inp_Signal_MA_Method                   | 0; 0; 0; 3; N |
| MAMA           | Inp_Signal_MA_Applied                 | 1; 1; 0; 7; N |
| MAMA           | Inp_Trailing_MA_Period                 | 12; 12; 1;120; N |
| MAMA           | Inp_Trailing_MA_Shift                 | 0.01;1.0;N |
| MAMA           | Inp_Trailing_MA_Method                 | 0.00;0.3;N |
| MAMA           | Inp_Trailing_MA_Applied               | 1.1;0.7;N |
| MAPS           | Inp_Expert_Title                       | ExpertMAPSAR |
| MAPS           | Inp_Signal_MA_Period                   | 12;12;1;120; Y |
| MAPS           | Inp_Signal_MA_Shift                   | 6;6;1;60; Y |
| MAPS           | Inp_Signal_MA_Method                   | 0.00;3;Y |
| MAPS           | Inp_Signal_MA_Applied                 | 0.00;3;Y |
| MAPS           | Inp_Trailing_ParabolicSAR_Slope       | 0.02;0.02;0.02;2.0;Y |
| MAPS           | Inp_Trailing_ParabolicSAR_SlopeMax    | 0.2;0.2;0.02;2.0;Y |
| MAPS2          | Inp_Expert_Title                       | ExpertMAPSARSizedOptimized |
| MAPS2          | Inp_Signal_MA_Period                   | 15;12;1;120; N |
| MAPS2          | Inp_Signal_MA_Shift                   | 4;6;1;60;N |
| MAPS2          | Inp_Signal_MA_Method                   | 0;0;3;N |
| MAPS2          | Inp_Signal_MA_Applied                 | 1;1;0;7;N |
| MAPS2          | Inp_Trailing_ParabolicSAR_SlopeMax    | 0.2;0.2;0.02;0.0000;N |
| MAPS2          | Inp_Trailing_ParabolicSAR_SlopeMax    | 0.2;0.2;0.02;0.0000;N |
| MAPS2          | Inp_Money_SizedOptimized_DebtFactor    | 3.310;4.0000;5.000000;N |
| MAPS2          | Inp_Money_SizedOptimized_Percent      | 10;10;1.00000;100.00000;N |

**TABLE XVIII**  
LIST OF EXPERT ADVISORS AND THEIR SPECIFIC PARAMETERS