Trends and Applications of Machine Learning in Quantitative Finance

Sophie Emerson, Ruairi Kennedy, Luke O'Shea, and John O'Brien

Abstract — **Recent advances in machine learning are finding commercial applications across many industries, not least the finance industry. This paper focuses on applications in one of the core functions of finance, the investment process. This includes return forecasting, risk modelling and portfolio construction. The study evaluates the current state of the art through an extensive review of recent literature. Themes and technologies are identified and classified, and the key use cases highlighted. Quantitative investing, traditionally a leading field in adopting new techniques is found to be the most common source of use cases in the emerging literature.**

Index Terms—**Machine Learning, Quantitative Finance, Portfolio Construction, Return Forecasting**

I. Introduction

Machine Learning (ML) is a subfield of Artificial Intelligence (AI) that uses statistical techniques that provide computer models with the ability to learn from a dataset, allowing the models to perform specific tasks without explicit programming [1]. ML is being applied to improve function across the finance industry in a wide range of areas including, for example, fraud detection, payment processing and regulation. This research evaluates current and potential applications of machine learning to the investment process. In particular, this includes the development of ML applications for return forecasting, portfolio construction and risk modelling.

The first widespread commercial use cases of artificial intelligence were "expert systems", originating in Stanford in the 1960s [2] and popularised in the 1980s and 1990s. Expert systems were designed to solve complex problems in a specific field, in a manner similar to a subject matter expert. Original expert systems were rule-based programs developed in languages such as LISP and Prolog. In recent years, there has been a significant drop in interest in classic expert systems, as they are superseded by systems incorporating artificial intelligence [3]. AI systems are systems that replicate human thought processes. [4]. Many of these systems are advertised today as cognitive computing systems.

Cognitive computing describes a computer system which mimics human cognitive process in some way, cognitive processes are those that allow individuals to remember, think, learn and adapt [5]. The term has gained recognition in the public domain in recent years, due in large to the introduction of Watson, IBM's cognitive computing system. These systems are constructed by combining computer science with statistical and ML techniques developed over the last century [1]. Watson, in its original form, was a question answering computing system, responding to questions posed in natural language. It was introduced on the television quiz show "Jeopardy!" – where it defeated two of the show's most celebrated contestants in the "IBM Challenge" [6]. Large-scale systems such as Watson combine many techniques [6] to provide "augmented human intelligence" services to users [7]. However, the use of individual techniques, for example deep learning neural networks or reinforcement learning, has found significant success across industry and applications [8-10].

Recently, there has been a proliferation of ML techniques and growing interest in their applications in finance, where they have been applied to sentiment analysis of news, trend analysis, portfolio optimization, risk modelling among many use cases supporting investment management. This paper explores the potential of ML to enhance the investment process. We begin with a broad survey of the area to determine the main programming languages, frameworks and use cases for ML from the perspective of the financial industry. We then focus on ML and its potential applications to quantitative investment. We look at research that has applied ML to the investment process, analysing the technologies used, the functions of the applications, and evidence of potential to improve investment outcomes. Our findings are relevant to both academics and practitioners with interest in investment management, and in particular quantitative investment, by providing a detailed discussion of the latest technologies, their potential uses, and probability of successful application.

The paper is organized as follows. In Section II, we provide an overview of the development of the area as a background for the discussion, this includes the emergence of ML, common algorithms and methodologies, and a review of the evolution and theory of quantitative investing We then describe the research methods in Section III. Section IV provides a detailed description of the current state of the art in the application of ML to investment. We conclude with a discussion of the evidence presented in Section V.

Manuscript received April 19, 2019. This work was supported by State Street Corporation and the authors wish to thank State Street for this support.

Sophie Emerson (sophieemerson@gmail.com), Ruairi Kennedy (r.ocinneide@umail.ucc.ie), and Luke O'Shea (davidluke.oshea@gmail.com) are researchers in the State Street Advanced Technology Centre, Cork University Business School, UCC, Ireland.

John O'Brien (j.obrien@ucc.ie) is a lecturer in the Department of Accounting & Finance, Cork University Business School, UCC, Ireland.

II. Background

A. Machine Learning

Although variations of ML have long been around, the discipline has developed rapidly in recent years. Many factors have combined to derive this development. Increased computer power has made real time processing feasible for many complex tasks, increased connectivity has driven innovation and automation in the delivery of traditional tasks and services, the potential to extract useful information from the vast amounts of data generated via the internet (Big Data) has led to novel analytical methods. Alongside this, the development of easy to use programming languages, such as Python and R, and ML focused frameworks such as TensorFlow, has contributed to the wide investigation of ML applications in industry. It has already found commercial application across multiple industries from automated trading systems in the finance industry to the health sector where ML algorithms assist decision making in fertility treatments [11]. The success of these applications is driving commercial research into further applications.

B. Common ML Approaches and Algorithms

Three main approaches to training ML algorithms are recognised; supervised learning, unsupervised learning and reinforcement learning. Supervised learning generates a function that maps inputs to outputs based on a set of training data. The algorithm infers a function linking each set of inputs with the expected, or labelled, output in the training set. Unsupervised learning finds hidden patterns in and draws inferences from unlabelled data. Unsupervised learning provides inputs to models, but does not specify an expected set of outcomes, the outcomes are unlabelled. Reinforcement learning enables algorithms to learn by trial and error, based on feedback from past experiences. Like unsupervised learning, it does not require labelled data. A hybrid system, semi-supervised learning, combines supervised and unsupervised learning, using both labelled and unlabelled data to train models. This is useful where there is limited data or the process of labelling data could introduce biases.

The main research areas in supervised learning are regression and classification (specifying the category or class to which something belongs), this approach is often used in developing predictive models. Regression techniques predict continuous responses using algorithms such as linear regression, decision trees and Artificial Neural Networks (ANNs). Classification techniques predict discrete responses using algorithms such as logistic regression, Support Vector Machines (SVMs) or K-Nearest Neighbors (KNN). The main research area in unsupervised learning is clustering. Clustering refers to grouping objects together, such that objects that are put in the same group are more similar to each other than objects in other groups.

Artificial neural networks have become a key technology in the development of ML. They were first proposed over 75 years ago, inspired by the workings of the human brain [12]. They are a collection of algorithms that replicate the process of a biological brain at the neuron level [1].

There are a number of different classes of artificial neural networks, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and recursive neural networks, among others. CNNs are ideal for things such as image classification and video processing because they're able to identify patterns by focusing on fragments of images. RNNs are better for dealing with things like speech or text analysis because they use time-series information, such as monthly stock price figures to predict next month's figure. GANs have garnered much interest in recent years since they were first introduced in 2014 [13]. GANs are comprised of two neural networks that compete against each other. One neural network generates data similar to the training dataset, and the other tries to evaluate whether data is from the training dataset or generated by the generative network.

Aside from neural networks other well-known ML algorithms include SVMs, KNN and other. SVMs, used for classification and regression analysis, involve finding a hyperplane which minimizes the distance between a set of data points in an n-dimensional space. Bayesian networks are built from probability distributions and use probability laws for prediction and anomaly detection. KNN selects the most similar data points in the training data, this allows the algorithm to classify future data inputs in the same way. Some techniques are better suited to particular tasks than others. This research partly seeks to contribute to this area of knowledge. It is important to evaluate the effectiveness of certain algorithms, to assist in choosing appropriate algorithms for specific tasks in future applications and studies.

C. The Evolution of Quantitative Investing

Graham and Dodd's *Security Analysis*, published in 1934 following the Wall Street Crash of 1929 is the seminal work on fundamental investing and remains in publication today [14]. It is one of the first books to distinguish investing from speculation, advocating the use of a systematic framework for analysing securities for stock selection.

A systematic approach to portfolio construction and risk analysis was presented in *Portfolio Selection* [15], published in 1952. In this, Markowitz provides a mathematical definition of risk as the standard deviation of return. The approach focused on maximizing portfolio performance by optimizing the trade-off between risk and return. This was the foundation of modern portfolio theory, providing an analytical framework for the construction and analysis of investment portfolios [16], [17].

A quantitative approach to market analysis gained popularity as advances in computing technology made the collection and analysis of large amounts of market data possible. This allowed the development and verification of market models on a scale not previously possible, contributing to significant advances in the understanding of financial markets, including the Capital Asset Pricing Model (CAPM) [18]-[21] and Efficient Market Hypothesis (EMH) [22].

In 1973, Fama and MacBeth used the Center for Research in Security Prices (CRSP) financial dataset (one of the first of its kind) to perform an empirical analysis of the CAPM [23]. They showed that the CAPM provided a good quantitative approximation of the behaviour of security prices while setting a standard for empirical cross-sectional analysis of market data [23].

The empirical support for the EMH, enhanced by the success of market indices, such as the S&P 500, led to the dominant view, particularly in academia, that active investing was futile, as it was impossible to beat a passive investment. In comprehensive literature reviews, [16] and [17] provide evidence that research and empirical evidence that challenged the CAPM and EMH was strongly discouraged. At the same time many examples of research that argued that although difficult, it is possible for active management to beat passive management, by exploiting market inefficiencies not covered by the CAPM and EMH. Strategies based on risk factor models, first explored by Rosenberg [24] and Ross [25] in the 1970s, surged in popularity [26] after the publication of the Fama-French three-factor model [27].

From Markowitz portfolio optimization to CAPM, EMH and factor models more recently, quantitative investors have shown that they are willing to embrace new techniques and strategies. A key argument for applying ML techniques to financial problems is that ML methods capture non-linear relationships [28] in the data. Non-linear methods are required to model data where outputs are not directly proportional to the inputs [29] and many traditional analysis methods assume a linear relationship, or a non-linear model that can be simplified to a linear model. Typical examples of well-established non-linear ML methods include KNN, and ANN [20].

ML has been applied with positive results across many areas of quantitative investing, including portfolio optimization [30], [31], factor investing [32], bond risk predictability [29], derivative pricing, hedging and fitting [33], and back-testing [34]. The results section contains a comprehensive summary of papers where ML techniques are applied to areas of quantitative finance.

III. Methodology

Initially, a broad search was conducted to identify the major themes related to ML. This search yielded information on the popular use cases and technologies. This information informed a second, more focused investigation of relevant material. Here, the aim was to draw connections between popular use cases in finance and current ML techniques.

As quality and scope of published research can vary widely, measures were taken to reduce the possibility of including unreliable information in the final dataset. Before inclusion in the concept matrix, each paper was assessed on quality. This was achieved by using a variety of quality indicators including; the citation count, the quality of an institute's research activities associated with the paper, bias created from funding sources, and the impact factor of the journal.

An appropriate search strategy was devised and carried out based on the main topics that were identified during the first investigation of the literature. The arXiv and SSRN databases were searched to ensure that the most up-to-date research papers were included. However, as these are not peer-reviewed papers, extra care was taken to ensure that the papers were from reputable authors, focusing on the quality of authors' previous publications. The topic phrases used in search were "portfolio management", "stock market forecasting", and "risk management". All of these topic phrases were used in conjunction with the key phrase "machine learning" in an attempt to return only relevant research papers. The purpose of searching by topic was to identify which technologies are widely and effectively used within each area. As we are evaluating the current state of the art, we wanted to ensure that only recent papers were included. Thus, we only included papers that were submitted in 2015 or later. From the initial search we collected a total of 118 papers. After an initial review of abstracts, papers that were not relevant to machine learning in finance (specifically investing) were removed. Any papers that were duplicates under more than one search topic were kept under the topic that appeared most relevant. Papers were then assessed in relation to their quality using the quality indicators mentioned above. This reduced the number of papers to 55.

IV. Results

A. *Popular Machine Learning Use Cases and Algorithms*

A concept-centric matrix was utilised initially to identify which areas commonly use machine learning techniques. Recurring concepts and themes were noted and counted across a sample of 67 papers identified. An initial list of recurring themes was identified and analysed. Some themes, such as 'Geopolitics' were removed as they were deemed irrelevant due to the lack of research on the topic. A list of the most recurring themes with relevance to ML is presented in Table I.

TABLE I: Recurring themes from the literature review.

Theme	References
Return Forecasting	21
Portfolio Construction	12
Ethics	8
Fraud Detection	8
Decision Making	8
Language Processing	7
Sentiment Analysis	7

The most common use-cases identified were return forecasting and portfolio construction. Quantitative methods were introduced to finance through the equity market and it is unsurprising that it should lead the way in incorporating the latest advances in its processes. A large number of the papers above also discussed risk modelling. This led us to take return forecasting, portfolio construction, and risk modelling as our three core topics. The most popular ML techniques identified in the papers researched are presented in Table II, as well as a breakdown of the different acronyms used in the table.

TABLE II: POPULAR TECHNIQUES FEATURED IN MACHINE LEARNING AND FINANCE PAPERS

	MLP	SVM	LSTM	GRU	RNN	CNN	RF	GPR	LR
Return Forecasting	7	5	4	2	-	1	2	-	-
Portfolio Construction	7	2	3	1	1	1	4	2	1
Risk Modelling	6	2	2	1	1	1	4	3	4

MLP	Multilayer Perceptron
SVM	Support Vector Machine
LSTM	Long Short-Term Memory
GRU	Gated Recurrent Unit
RNN	Recurrent Neural Network (basic)
CNN	Convolutional Neural Network
RF	Random Forests/Decision Trees
GPR	Gaussian Process Regression
LR	Logistic Regression

Many techniques used in the papers only appear once, some twice. Since the purpose of this paper is to identify the most popular machine learning techniques used in finance, specifically in the topics above, only techniques which appeared in at least three papers were included in Table II. We also decided to include RNN, although it is only mentioned explicitly in two papers, it appears implicitly more frequently as both LSTM and GRU are subsets of the technology.

Artificial neural networks are used in all three areas of finance studied, with a standard feedforward network (MLP) being the most common. Useful results are found from networks that range from small to very large networks (deep neural networks). There is also evidence of preferences for some techniques in particular areas. For example, Gaussian process regression is used in both portfolio construction and risk modelling but has not been applied to return forecasting.

B. *Summary of Key Insights from Recent Papers*

The paper selection included ML papers published in recent years as well as papers yet to be published by established authors from reputable institutions. These papers have been submitted for publication and are awaiting acceptance. The most recent studies in this field were included to help evaluate the cutting edge and state of the art of the use of ML for financial applications.

I. *Portfolio Construction*

Portfolio construction is the process of combining return forecasts and risk models to create an optimum portfolio given an investor's constraints. A variety of ANN methodologies are applied to the portfolio optimisation problem, often outperforming traditional optimisation techniques. Deep learning reappeared a number of times during this search in the context of portfolio construction. Deep learning refers to models that consist of multiple layers or stages of nonlinear information processing (for example, a neural network with many hidden layers) [35]. Both hierarchical clustering and reinforcement learning were used to improve portfolio diversification. Multiple papers discuss the method of applying Markov models to predict the performance of stocks. Markov models are a type of ML method that model variables that change randomly through time. The complicated nature of the global market makes using this type of model a viable option.

- The authors present a deep learning framework for portfolio design, applying their framework to the stocks in the IBB index, demonstrating that their portfolio weighted using deep learning outperformed the index [31].
- The author outlines a reinforcement learning solution for a rational risk-averse investor seeking to maximize expected utility of final wealth, giving an example of a Q-learning agent exploiting an approximate arbitrage in a simulation [36].
- The authors of both papers make use of hierarchical clustering algorithms for constructing diversified portfolios. The portfolios are constructed using variations of risk parity [30] and equal risk contribution methods [37] which take the hierarchical correlation structure of the assets into account. The portfolios constructed are shown to have superior diversification and out-of-sample risk adjusted performance.
- The authors make use of convex analysis techniques to devise an optimal portfolio coupled with a Hidden Markov Model (HMM) used to estimate growth rates in the market model, which achieves improved results over a simple model using geometric Brownian motions [38].
- The authors provide an overview of the financial applications of Gaussian processes and Bayesian optimisation, providing examples for forecasting the yield curve with Gaussian processes, and using Bayesian optimisation to build an online trend-following portfolio optimisation strategy [39].
- The authors compare the use of Feature Salient Hidden Markov Models (FSHMM) and HMM for constructing factor investing portfolios. The FSHMM selects relevant factors for use from a pool of available factors, while the HMM uses the whole pool of factors. Both models outperformed benchmark portfolios, with the FSHMM portfolio showing better performance [40]
- The authors use factors as inputs to deep neural network, SVM and random forest models for predicting stock returns. While their research shows the effectiveness of a deep learning model, more

significantly they used Layer-wise Relevance Propagation (LRP) to determine individual factor contributions to the neural network's prediction [41].

- The authors create a non-linear multi-factor model using LSTM to estimate the non-linear function. As in the previous paper the authors make use of LRP to identify which factors contribute to the model. The performance of the LSTM model is compared to the neural network model used in [32] and gives superior returns [42].
- The authors examine the use of three deep reinforcement learning algorithms, Deep Deterministic Policy Gradient (DDPG), Proximal Policy Optimization (PPO) and Policy Gradient (PG), in managing a portfolio of assets in the Chinese stock market. They determine that training conditions used in game playing and robot control are unsuitable for use with portfolio management, finding that DDPG and PPO gave unsatisfying performance in the training process. They propose the use of adversarial training methods and employ a revised PG algorithm which outperforms a Uniform Constant Rebalanced Portfolio (UCRP) benchmark [43].
- The authors employ models constructed using Gaussian processes and Monte Carlo Markov Chains which learn optimal strategies from historical data, based on user-specified performance metrics (e.g. excess return to the market index, Sharpe ratio, etc.). This approach addresses the inverse problem of Stochastic Portfolio Theory – devising suitable investment strategies that meet the desired investment objective, when initially given a user-defined portfolio selection. The models outperform the benchmark in-sample and out-of-sample for absolute terms (returns) and also after adjusting for risk (Sharpe ratio) [44].
- The author provides an ML framework for estimating optimal portfolio weights. They apply this framework using three ML methods – Ridge and Lasso regression, and two newly introduced methods; Principal Component regression, Spike and Slab regression. All methods outperform the mean-variance, minimum-variance, and equal weight portfolios. [45].
- The authors propose a way to find the risk budgeting portfolio by using optimisation algorithms to find a solution to the logarithmic barrier problem. They use algorithms such as cyclical coordinate descent, alternating direction method of multipliers (ADMM), proximal operators, and Dykstra's algorithm [46].
- The authors present a financial-model-free reinforcement learning framework as a solution to the portfolio management problem. The study tests the proposed framework with the following neural networks: CNN, a basic RNN, and LSTM [47].

II. Return Forecasting

Return forecasting, predicting the investment return from an asset or asset class, is central to investment management and features highly in the literature. Many types of ANN are tested on their ability to forecast returns. Deep neural networks, CNNs, LSTMs are all applied to the problem of return forecasting. In one theme, the new ML technology is applied to improve forecasts made from traditional inputs, such as fundamental accounting data or technical indicators. A second approach uses ML to extract new inputs from alternative data, such as sentiment from news data. Finally, authors predict movement at market level rather than at the level of individual securities, for example using ML to identify states.

- The authors use a CNN strategy to analyse and detect price movement patterns in high-frequency limit order book data. Multilayer neural network methods and SVMs were also considered. However, they conclude the CNNs provide better performance for this task [48].
- The authors implement several ML algorithms to predict future price movements using limit order book data. They employ two feature learning methods: Autoencoders, and Bag of Features. They compare three different classifiers: SVM, a Single Hidden Layer Feedforward Neural Network (SLFNN), and an MLP. They test the performance of the classifiers with an anchored walk forward analysis, to determine if the models can capture temporal information, as well as a hold-out per stock method, to determine if the models can learn features that can be applied to previously unseen stocks. The results from the MLP are better than the other classifiers. However, the use of the Autoencoder and Bag of Features in combination with the MLP lead to fewer correct predictions [49].
- The authors introduce a novel Temporal Logistic Neural Bag-of-Features approach, that can be used to tackle the challenges that come with data of a high dimensionality, in this case high-frequency limit order book data [50].
- The authors train a deep neural network on reported fundamental data from publicly traded companies (revenue, operating income, debt etc.). The model forecasts future fundamental data based on a trailing 5-years window. A value investing factor strategy based on forecasted fundamental data outperforms a traditional value factor investing approach with a compounded annual return of 17.1% vs 14.4% for a standard factor model [51].
- The authors create a simple buy-hold-sell strategy to predict direction of movement for 43 CME listed commodities and FX futures based on an ANN trained on a multitude of features for each instrument designed to capture co-movements and historical memory in the data. An average prediction accuracy of 42% is achieved across all instruments, with higher accuracies achieved for certain instruments [52].

- The authors use a random forest model to predict the direction of stock prices based on price information and a number of momentum indicators (Relative Strength Index, Moving Average Convergence Divergence, Stochastic Oscillator, Williams %R, On Balance Volume, and Price Rate of Change). The algorithm is shown to outperform existing algorithms found in the literature [53].
- The authors provide a sentiment analysis dictionary which they use to predict stock movements in the pharmaceutical market sector. With this model they achieve an accuracy of 70.59%. [54]
- The authors present a methodology to define, identify, classify and forecast market states. They use a Triangulated Maximally Filtered Graph network to filter information, and simple logistic regression for predicting market states. They compare five models, with a Gaussian Mixture Model as their baseline. All five models outperform the baseline in terms of risk/return significance [55].
- The authors compare five ANN models for forecasting stock prices: a standard neural network using back propagation, a Radial Basis Function (RBF), a General Regression Neural Network (GRNN), SVM Regression (SVMR), and Least Squares SVM Regression (LS-SVMR). However, they compare the models on just three stocks: Bank of China, Vanke A, and Kweichou Moutai. The standard neural network using back propagation outperforms all of the other models across all three stocks, in terms of both Mean Squared Error (MSE) and Mean Absolute Percentage Error (MAPE). [56]
- The authors use 25 risk factors as inputs to ML stock returns prediction models. Results show that deep neural networks generally outperform shallow neural networks, and the best networks also outperform representative machine learning models [57].
- The author employs ANNs to predict product demand for weather sensitive products in Walmart stores around the time of major weather events [58].
- The authors implement a Gaussian Naïve Bayes Classifier for prediction based on sentiment analysis of Twitter data. The data used was obtained from Twitter and pertained to the 2014 FIFA world cup. Their framework obtained an accuracy and Area Under the curve of the Receiver Operating Characteristic (AUROC) of around 80% and an 8% marginal profit when tested [59].

III. Risk

Three different themes are identified under the broad heading of risk. The first attempts to employ ML to improve traditional measures of risk used in the mean variance framework. The second theme looks for companies at risk of default or bankruptcy. Techniques such as natural language processing are used to identify words that indicate higher risk. The final theme uses ML to develop hedging strategies. Some authors look at identifying what selection of ML methods is best for risk modelling problems.

- The authors use k-means clustering to construct risk models by clustering stock returns normalized by standard deviation squared and adjusted by mean absolute deviation using a method proposed in [60]. They demonstrate that this ML approach outperforms statistical risk models [61] in quantitative trading applications [62].
- The authors present a framework for hedging a portfolio of derivatives in the presence of market frictions such as transaction costs, market impact, liquidity constraints or risk limits [63].
- The authors show how Gaussian Process Regression can assist in pricing and hedging a Guaranteed Minimum Withdrawal Benefit (GMWB) Variable Annuity with stochastic volatility and stochastic interest rate [64].
- The authors show that machine learning can be as effective as other existing algorithms at solving difficult hedging problems in moderate dimension. They use techniques such as a modified LSTM neural network to calculate their hedging strategies [65].
- The authors aim to explore the optimal model for business risk prediction. They attempt to do this using XGBoost, and by simultaneously examining feature selection methods and hyper-parameter optimization in the modelling procedure [66].
- The authors try to predict daily stock volatility using news and price data. Their model, which utilizes a Bidirectional Long Short-Term Memory (BiLSTM) neural network and stacked LSTM's, outperforms the well-known Generalized Autoregressive Conditional Heteroskedasticity (GARCH) model in all sectors analysed (financial, health care, etc.) [67].
- The authors exploit a heterogeneous information network of 35,657 global firms to improve the predictive performance for firms likely to be added to a blacklist. Blacklists are used to keep track of entities that have unacceptable problems, such as financial or environmental issues. Blacklists help keep portfolios profitable and "green". Their model consists of a simple MLP with thirty hidden units [68].
- The authors estimate corporate credibility of Chinese companies using a CNN and natural language processing. They use Latent Dirichlet Allocation to summarise the text of news articles and use a CNN to extract the most important words from each topic. The CNN learns how news articles may reflect the credibility of a company though the wording of articles and word occurrence. They verify their model works by building a negative rating system and showing a correlation between their model's results and the negative rating [69].
- The authors compare different strategies for solving a variation of the multi-armed bandit problem. In their version of the problem, the learner can pull several arms simultaneously, or none at all. This could easily be applied to assist in investment decisions. Out of the strategies compared, Bayes-UCB-4P and TS-4P perform the best [70].

- The authors compare several ML algorithms: Logistic Regression, K-Dimensional Tree (K-D Tree), SVM, Decision Trees, AdaBoost, ANN, and Gaussian Processes (GP) for forecasting business failures (corporate bankruptcy). Models are compared on datasets of manufacturing companies in Korea and Poland. All of the models are compared on their performance when combined with different dimensionality reduction techniques. The techniques used are: Principal Component Analysis (PCA), Linear Discriminate Analysis (LDA), Isometric Feature Mapping (ISOMAP), and Kernel PCA. On the Korean dataset, all models perform similarly. K-D Tree, SVM, and GP perform best over all of the dimensionality reduction methods used. On the Polish dataset, the linear regression model performs the best. Although having a lower accuracy than some of the other models, it is the best performing method when compared over other results such as precision, recall, F1 score, and AUC (Area Under Curve) [71].

V. Discussion

A. Strategy Development & Analysis

The results of the literature search demonstrate that there is a wide range of ML techniques being successfully applied to many areas in the development of quantitative investing strategies, outperforming traditional benchmarks, previously used techniques and algorithms in many cases. Algorithms that assume a linear relationship between data can result in reduced accuracy. [28] highlights this issue in terms of many of the econometric models employed by finance academics and investment managers. The author argues for the use of more advanced mathematical models and ML techniques such as unsupervised learning that are capable of modelling complex non-linear relationships in financial systems.

Taking factor investing as an example of this, [72] and [73] make use of statistical algorithms to show that many factors discovered over the last number of years (particularly those found using empirical evidence) can be considered inaccurate or invalid. In the aptly named paper, *Taming the Factor Zoo*, a double selection LASSO ML method was used to analyse the contribution and usefulness of individual factors amongst the large number available today [74]. LASSO (Least Absolute Shrinkage and Selection Operator) is a regression analysis method capable of reducing the dimensionality of a large sample while selecting variables significant to the final result [75]. In [57] the author uses twenty-five factors as model inputs, comparing the use of shallow and deep neural networks, as well as SVMs and random forests for predicting stock returns, finding the deep neural networks (more layers) superior to the other methods. Using a similar approach [41] uses factors as inputs to deep neural network, SVM and random forest models for predicting stock returns. While their research again showed the effectiveness of a deep learning model, more significantly they used layer-wise relevance propagation to determine individual factors contributions to the neural network's prediction. In these cases, not only has ML been used to develop investment strategies, but also to detect which input features were significant and which were not.

B. The use of Alternative Data

The use of ML for the analysis and application of alternative data for example, sentiment analysis, supply chain data etc. has opened up opportunities for new investment strategies. As seen in Table I, sentiment analysis was identified as a popular use case for ML. [17] provides a thorough overview of the growth of big data and sentiment analysis research over the last 30 years, highlighting the use of techniques such as NLP, SVMs and ANNs for the analysis of news, conference calls, reports, and social media activity. They concluded that to date, sentiment information has provided short-term, easy to exploit insights but long-term persistent insights are hard to achieve (falling in line with EMH). [16] acknowledges the effectiveness of big data for the modern fundamental investor, as it can provide insights and improve decision making by widening their research capabilities. This sentiment is echoed in [28] where the author makes reference to the recently emerged term “quantamental" – describing a fundamentally leaning investor who manages their portfolio based on data-driven insights provided by ML algorithms. Examples of ML and alternative data being applied together in the results section mainly fall under return forecasting or risk modelling, where decisions may be made based on good or bad news [54], weather [58], or social media sentiment [59].

C. Choosing Machine Learning Algorithms

It is important to understand the relevant factors that contribute to the choice of ML algorithms, given the wide range available. These factors include accuracy, training time, linearity, number of parameters, the number of features and the structure of the data [76]. Some systems do not need a high level of accuracy. Estimates may be sufficient, for example, when calculating different route times for a journey. Model training times can also vary hugely between algorithms, making some algorithms more appealing than others when under time constraints. Many algorithms assume a linear relationship between input and output (linear regression, logistic regression, SVMs). This can result in reduced accuracy when dealing with non-linear problems. The number of parameters an algorithm has can indicate its flexibility, but also indicates that more time and effort may be required to find optimal values for training the model. The number of features can also be overwhelming for some algorithms. This is particularly a problem with textual data, where the number of words in the dictionary vastly outweighs the number of words in say, a paragraph being used for sentiment analysis. It's important to consider the structure of the data and the specific problem, as some algorithms are better suited for certain problems and data structures [77].

D. Backtesting & Strategy Verification

While ML techniques can provide superior performance, financial data is notorious for having a low signal-to-noise ratio, which can lead to the detection of false patterns and results. Backtesting protocols have been proposed to tackle this [78]. ML solutions have also been applied to this problem. In [34] the authors present an unsupervised learning strategy which makes use of a modified k-means clustering algorithm to extract the number of uncorrelated trials from a series of backtests, which can be used in estimating the probability of false positives and estimating the expected value of the maximum Sharpe ratio. While in [79] the authors use a machine learning strategy for backtesting and the evaluation of automated trading strategies which is trained on a number of performance and risk metrics, demonstrating that this strategy outperforms standard metrics such as Sharpe ratio out-of-sample.

The development of new backtesting strategies and protocols is welcome and necessary, especially taking into account recent “black box” criticisms by leading deep learning researchers regarding a lack of testing and reproducibility in the field of ML. In their acceptance speech after winning the “test-of-time” award at NIPS, the leading AI conference, the authors of [80] compared much of recent ML research to “alchemy”, highlighting a situation where algorithms were being created and trained using trial and error methods, with the researchers unable to explain the fundamental operation. They later published a paper highlighting instances of this [81].

VI. Conclusion

As the previous section discusses, ML offers an opportunity for more complex financial analysis than was previously possible. The literature shows that quantitative investors have embraced new tools and techniques as they have emerged [16], [17].

There is a growing body of literature applying ML techniques to investment problems. Varieties of ML methods have been applied to areas of quantitative finance– the most popular methods are MLPs, followed by SVMs, and LSTM. ML has been applied to problems in areas such as return forecasting, portfolio construction, and risk modelling.
These ML methods utilize traditional financial data, as well as making use of new types of alternative data. Big data is providing new datasets that need to be analysed and ML techniques are capable of modelling complex (non-linear) relationships and analysing new data.

[28] notes the recent trend of traditional hedge funds hiring an increasing proportion of STEM graduates for portfolio construction positions, as they possess the required mathematical skillset for performing complex analysis and computer modelling. An understanding of machine learning, as well as the languages (Python, R, etc.) and frameworks (e.g. TensorFlow) needed to construct complex models could certainly be considered advantageous for any quantitative investor looking for an edge.

References

[1] P. Domingos, "A few useful things to know about machine learning," *Communications of the ACM,* vol. 55, no. 10, pp. 78-87, 2012.

[2] R. K. Lindsay, B. G. Buchanan, E. A. Feigenbaum, and J. Lederberg, "DENDRAL: a case study of the first expert system for scientific hypothesis formation," *Artificial intelligence,* vol. 61, no. 2, pp. 209-261, 1993.

[3] W. P. Wagner, "Trends in expert system development: A longitudinal content analysis of over thirty years of expert system case studies," *Expert systems with applications,* vol. 76, pp. 85-96, 2017.

[4] S. Russell and P. Norvig, *Artificial Intelligence A Modern Approach Third Edition*. Prentice Hall, 2009.

[5] D. Modha and C. Witchalls, "A computer that thinks," vol. 224, ed: New Scientist Ltd., 2014, pp. 28-29.

[6] D. Ferrucci *et al.*, "Building Watson: An overview of the DeepQA project," *AI magazine,* vol. 31, no. 3, pp. 59-79, 2010.

[7] H. Reynolds, "AI? Or cognitive computing?," *KM World,* Article vol. 26, no. 9, pp. 4-5, 2017.

[8] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, and F. E. Alsaadi, "A survey of deep neural network architectures and their applications," *Neurocomputing,* vol. 234, pp. 11-26, 2017/04/19/ 2017.

[9] R. Rana and F. S. Oliveira, "Dynamic pricing policies for interdependent perishable products or services using reinforcement learning," *Expert Systems with Applications,* vol. 42, no. 1, pp. 426-436, 2015.

[10] G. Choy *et al.*, "Current applications and future impact of machine learning in radiology," *Radiology,* vol. 288, no. 2, pp. 318-328, 2018.

[11] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[12] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The bulletin of mathematical biophysics,* vol. 5, no. 4, pp. 115-133, 1943.

[13] I. Goodfellow *et al.*, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672-2680.

[14] B. Graham and D. Dodd, *Security Analysis: Foreword by Warren Buffett*. McGraw-Hill Professional, 2008.

[15] H. Markowitz, "Portfolio selection," *The journal of finance,* vol. 7, no. 1, pp. 77-91, 1952.

[16] R. N. Kahn, *The Future of Investment Management*. CFA Institute Research Foundation, 2018.

[17] Y. L. Becker and M. R. Reinganum, *The Current State of Quantitative Equity Investing*. CFA Institute Research Foundation, 2018.

[18] W. F. Sharpe, "Capital asset prices: A theory of market equilibrium under conditions of risk," *The journal of finance,* vol. 19, no. 3, pp. 425-442, 1964.

[19] J. Mossin, "Equilibrium in a capital asset market," *Econometrica: Journal of the econometric society,* pp. 768-783, 1966.

[20] J. Lintner, "The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets," in *Stochastic Optimization Models in Finance*: Elsevier, 1975, pp. 131-155.

[21] C. W. French, "The Treynor capital asset pricing model," *Journal of Investment Management,* vol. 1, no. 2, pp. 60-72, 2003.

[22] B. G. Malkiel and E. F. Fama, "Efficient capital markets: A review of theory and empirical work," *The journal of Finance,* vol. 25, no. 2, pp. 383-417, 1970.

[23] E. F. Fama and J. D. MacBeth, "Risk, return, and equilibrium: Empirical tests," *Journal of political economy,* vol. 81, no. 3, pp. 607-636, 1973.

[24] B. Rosenberg, "Extra-market components of covariance in security returns," *Journal of Financial and Quantitative Analysis,* vol. 9, no. 2, pp. 263-274, 1974.

[25] S. A. Ross, "The arbitrage theory of capital asset pricing," in *HANDBOOK OF THE FUNDAMENTALS OF FINANCIAL DECISION MAKING: Part I*: World Scientific, 2013, pp. 11-30.

[26] J. H. Cochrane, "Presidential Address: Discount Rates," *The Journal of Finance,* vol. 66, no. 4, pp. 1047-1108, 2011.

[27] E. F. Fama and K. R. French, "Common risk factors in the returns on stocks and bonds," *Journal of financial economics,* vol. 33, no. 1, pp. 3-56, 1993.

[28] M. Lopez de Prado, "Mathematics and economics: a reality check," 2016.

[29] D. Bianchi, M. Büchner, and A. Tamoni, "Bond risk premia with machine learning," *Available at SSRN 3232721*, 2018.

[30] M. Lopez de Prado, "Building diversified portfolios that outperform out-of-sample," *Journal of Portfolio Management*, 2016.

[31] J. Heaton, N. Polson, and J. H. Witte, "Deep learning for finance: deep portfolios," *Applied Stochastic Models in Business and Industry*, vol. 33, no. 1, pp. 3-12, 2017.

[32] K. Nakagawa, T. Uchida, and T. Aoshima, "Deep Factor Model," *arXiv preprint arXiv:1810.01278*, 2018.

[33] J. De Spiegeleer, D. B. Madan, S. Reyners, and W. Schoutens, "Machine learning for quantitative finance: fast derivative pricing, hedging and fitting," *Quantitative Finance*, pp. 1-9, 2018.

[34] M. Lopez de Prado and M. J. Lewis, "Detection of False Investment Strategies Using Unsupervised Learning Methods," *Available at SSRN: https://ssrn.com/abstract=3167017 or http://dx.doi.org/10.2139/ssrn.3167017*, 2018.

[35] L. Deng and D. Yu, "Deep learning: methods and applications," *Foundations and Trends® in Signal Processing*, vol. 7, no. 3–4, pp. 197-387, 2014.

[36] G. Ritter, "Machine learning for trading," 2017.

[37] T. Raffinot, "Hierarchical clustering based asset allocation," *Available at SSRN 2840729*, 2017.

[38] A. Al-Aradi and S. Jaimungal, "Active and Passive Portfolio Management with Latent Factors," *arXiv preprint arXiv:1903.06928*, 2019.

[39] J. Gonzalvez, E. Lezmi, T. Roncalli, and J. Xu, "Financial Applications of Gaussian Processes and Bayesian Optimization," *arXiv preprint arXiv:1903.04841*, 2019.

[40] E. Fons, P. Dawson, J. Yau, X.-j. Zeng, and J. Keane, "A novel dynamic asset allocation system using Feature Saliency Hidden Markov models for smart beta investing," *arXiv preprint arXiv:1902.10849*, 2019.

[41] K. Nakagawa, T. Uchida, and T. Aoshima, "Deep factor model," in *ECML PKDD 2018 Workshops*, 2018, pp. 37-50: Springer.

[42] K. Nakagawa, T. Ito, M. Abe, and K. Izumi, "Deep Recurrent Factor Model: Interpretable Non-Linear and Time-Varying Multi-Factor Model," in *AAAI-19 Workshop on Network Interpretability for Deep Learning*, Honolulu, Hawaii, USA, 2019: arXiv preprint arXiv:1901.11493.

[43] Z. Liang, K. Jiang, H. Chen, J. Zhu, and Y. Li, "Deep Reinforcement Learning in Portfolio Management," *arXiv preprint arXiv:1808.09940*, 2018.

[44] Y.-L. K. Samo and A. Vervuurt, "Stochastic Portfolio Theory: a machine learning perspective," presented at the Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence, Jersey City, New Jersey, USA, 2016.

[45] D. Kinn, "Reducing Estimation Risk in Mean-Variance Portfolios with Machine Learning," *arXiv preprint arXiv:1804.01764*, 2018.

[46] J.-C. Richard and T. Roncalli, "Constrained Risk Budgeting Portfolios: Theory, Algorithms, Applications & Puzzles," *arXiv preprint arXiv:1902.05710*, 2019.

[47] Z. Jiang, D. Xu, and J. Liang, "A deep reinforcement learning framework for the financial portfolio management problem," *arXiv preprint arXiv:1706.10059*, 2017.

[48] A. Tsantekidis, N. Passalis, A. Tefas, J. Kanniainen, M. Gabbouj, and A. Iosifidis, "Forecasting Stock Prices from the Limit Order Book Using Convolutional Neural Networks," in *2017 IEEE 19th Conference on Business Informatics (CBI)*, 2017, vol. 01, pp. 7-12.

[49] P. Nousi *et al.*, "Machine learning for forecasting mid price movement using limit order book data," *arXiv preprint arXiv:1809.07861*, 2018.

[50] N. Passalis, A. Tefas, J. Kanniainen, M. Gabbouj, and A. Iosifidis, "Temporal Logistic Neural Bag-of-Features for Financial Time series Forecasting leveraging Limit Order Book Data," *arXiv preprint arXiv:1901.08280*, 2019.

[51] J. Alberg and Z. C. Lipton, "Improving Factor-Based Quantitative Investing by Forecasting Company Fundamentals," *arXiv preprint arXiv:1711.04837*, 2017.

[52] M. Dixon, D. Klabjan, and J. H. Bang, "Classification-based financial markets prediction using deep neural networks," *Algorithmic Finance*, no. Preprint, pp. 1-11, 2017.

[53] L. Khaidem, S. Saha, and S. R. Dey, "Predicting the direction of stock market prices using random forest," *arXiv preprint arXiv:1605.00003*, 2016.

[54] D. Shah, H. Isah, and F. Zulkernine, "Predicting the Effects of News Sentiments on the Stock Market," in *2018 IEEE International Conference on Big Data (Big Data)*, 2018, pp. 4705-4708: IEEE.

[55] P. F. Procacci and T. Aste, "Forecasting market states," *Available at SSRN: https://ssrn.com/abstract=3215945 or http://dx.doi.org/10.2139/ssrn.3215945*, 2018.

[56] Y.-G. Song, Y.-L. Zhou, and R.-J. Han, "Neural networks for stock price prediction," *arXiv preprint arXiv:1805.11317*, 2018.

[57] M. Abe and H. Nakayama, "Deep Learning for Forecasting Stock Returns in the Cross-Section," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2018, pp. 273-284: Springer.

[58] E. Taghizadeh, "Utilizing artificial neural networks to predict demand for weather-sensitive products at retail stores," *arXiv preprint arXiv:1711.08325*, 2017.

[59] L. Le, E. Ferrara, and A. Flammini, "On predictability of rare events leveraging social media: a machine learning perspective," in *Proceedings of the 2015 ACM on Conference on Online Social Networks*, 2015, pp. 3-13: ACM.

[60] Z. Kakushadze and W. Yu, "Statistical industry classification," *Journal of Risk & Control*, vol. 3, no. 1, pp. 17-65, 2016.

[61] Z. Kakushadze and W. Yu, "Statistical risk models," *The Journal of Investment Strategies*, vol. 6, no. 2, pp. 1-40, 2017.

[62] Z. Kakushadze and W. Yu, "Machine Learning Risk Models," *Journal of Risk & Control*, vol. 6, no. 1, pp. 37-64, 2019.

[63] H. Buehler, L. Gonon, J. Teichmann, and B. Wood, "Deep hedging," *Quantitative Finance*, pp. 1-21, 2019.

[64] L. Goudenège, A. Molent, and A. Zanette, "Gaussian Process Regression for Pricing Variable Annuities with Stochastic Volatility and Interest Rate," *arXiv preprint arXiv:1903.00369*, 2019.

[65] S. Fecamp, J. Mikael, and X. Warin, "Risk management with machine-learning-based algorithms," *arXiv preprint arXiv:1902.05287*, 2019.

[66] Y. Wang and X. S. Ni, "A XGBoost risk model via feature selection and Bayesian hyper-parameter optimization," *International Journal of Database Management Systems (IJDMS) preprint arXiv:1901.08433*, 2019.

[67] M. Sardelich and S. Manandhar, "Multimodal deep learning for short-term stock volatility prediction," *arXiv preprint arXiv:1812.10479*, 2018.

[68] R. Hisano, D. Sornette, and T. Mizuno, "Predicting Adverse Media Risk using a Heterogeneous Information Network," *arXiv preprint arXiv:1811.12166*, 2018.

[69] M. Zhang, Z. Luo, and H. Lu, "Latent Dirichlet Allocation with Residual Convolutional Neural Network Applied in Evaluating Credibility of Chinese Listed Companies," *arXiv preprint arXiv:1811.11017*, 2018.

[70] M. Achab, S. Clémençon, and A. Garivier, "Profitable Bandits," presented at the Proceedings of The 10th Asian Conference on Machine Learning, Proceedings of Machine Learning Research, 2018. Available: http://proceedings.mlr.press

[71] J. C. Chow, "Analysis of Financial Credit Risk Using Machine Learning," *arXiv preprint arXiv:1802.05326*, 2018.

[72] C. Harvey and Y. Liu, "Lucky factors," 2017.

[73] C. R. Harvey, Y. Liu, and H. Zhu, "… and the Cross-Section of Expected Returns," *The Review of Financial Studies*, vol. 29, no. 1, pp. 5-68, 2016.

[74] G. Feng, S. Giglio, and D. Xiu, "Taming the factor zoo," 2017.

[75] A. Belloni, V. Chernozhukov, and C. Hansen, "Inference on treatment effects after selection among high-dimensional controls," *The Review of Economic Studies*, vol. 81, no. 2, pp. 608-650, 2014.

[76] R. Barga, V. Fontama, and W. H. Tok, "Introducing Microsoft Azure Machine Learning," in *Predictive Analytics with Microsoft Azure Machine Learning*: Springer, 2015, pp. 21-43.

[77] P. Harrington, "Machine learning in action," *Shelter Island, NY: Manning Publications Co*, 2012.

[78] R. Arnott, C. R. Harvey, and H. Markowitz, "A Backtesting Protocol in the Era of Machine Learning," *The Journal of Financial Data Science*, vol. 1, no. 1, pp. 64-74, 2019.

[79] T. Wiecki, A. Campbell, J. Lent, and J. Stauth, "All That Glitters Is Not Gold: Comparing Backtest and Out-of-Sample Performance on a Large Cohort of Trading Algorithms," *The Journal of Investing*, vol. 25, no. 3, pp. 69-80, 2016.

[80] B. Recht and A. Rahimi, "Reflections on random kitchen sinks, 2017," ed, 2017.

[81] D. Sculley, J. Snoek, A. Wiltschko, and A. Rahimi, "Winner's Curse? On Pace, Progress, and Empirical Rigor," 2018.