HELLENIC
OPEN
UNIVERSITY

School of Social Sciences

Master in Business Administration (MBA)

Postgraduate Dissertation

# Machine Learning in Accounting and Finance Research: A Literature Review

Evangelos Liaras

Supervisor: Michail Nerantzidis

Patras, Greece, Mar 2022

# Machine Learning in Accounting and Finance Research: A Literature Review

Evangelos Liaras

Supervising Committee

Supervisor:
Michail Nerantzidis
University of Thessaly

Co-Supervisor:
Panagiotis Artikis
University of Piraeus

Patras, Greece, Mar 2022

*"Acknowledgments and / or Dedication"*

# Abstract

In recent years, accounting and finance scholars have shown a growing interest in using machine learning for research. By combining bibliographic coupling and literature review, this study analyzes (181) papers from (47) established quality journals in the field of accounting and finance published between 1998 and 2021; it addresses three interrelated research questions: (RQ1) How is research on the impact of machine learning on accounting and finance research being developed? (RQ2) What are the focus and critique of this corpus of literature? (RQ3) What are the future avenues of machine learning in accounting and finance research? We find increased interest in this field since 2015 with majority of studies focused either in U.S markets or global scale approach, while Asian markets increased their share by 600% during 2019-2021. Supervised models are by far the most applied (>60%) in contrary to unsupervised models mainly focused on topic extraction thru LDA algorithm; an indication that research in finding new patterns is limited. We firstly examine our corpus under (10) major research categories identifying proposed machine learning algorithms while neural networks being the most commonly used. We also created (7) clusters thru bibliographic coupling using VosViewer identifying key topics, current challenges and opportunities that machine learning appliance provides. Stock and commodities price forecasting, bankruptcy prediction and asset pricing can summarize current interest in which traditional models are either slow (e.g. Monte Carlo), or less accurate than machine learning which can cater for non-linear and high-dimensional data. Adaptation of machine learning is also hindered by lack of algorithm decision making transparency, regulations and wrong choice of tasks to be replaced. Accounting research is limited (15% of our corpus) on auditing, education and skills that practitioners should require to cope up with current and future demands. Finally, we outline some promising research avenues for future research.

**Keywords**

Machine Learning, Accounting, Finance, Literature Review, Deep learning, Artificial Intelligence

# Μηχανική Μάθηση στη Λογιστική και Χρηματοοικονομική Έρευνα: Βιβλιογραφική Ανασκόπηση
(in Greek)

Ευάγγελος Λιάρας
(in Greek)

## Περίληψη

## Abstract

Τα τελευταία χρόνια, οι ερευνητές λογιστικής και χρηματοοικονομικών, έχουν επιδείξει ιδιαίτερο ενδιαφέρον για τους αλγορίθμους μηχανικής μάθησης. Συνδυάζοντας βιβλιογραφική ανασκόπηση με βιβλιογραφική σύζευξη, αναλύουμε (181) άρθρα από (47) καταξιωμένα περιοδικά στο πεδίο της λογιστικής και των χρηματοοικονομικών που δημοσιεύτηκαν μεταξύ 1998 και 2021. Απαντάμε σε τρία επιστημονικά ερωτήματα: (ΕΕ 1) Ποια είναι η τρέχουσα κατάσταση της βιβλιογραφίας σχετικά με τη χρήση μηχανικής μάθησης στην λογιστική και χρηματοοικονομική; (ΕΕ 2) Που επικεντρώνεται και ποια τα συμπεράσματα της συγκεκριμένης βιβλιογραφίας; (ΕΕ 3) Ποιο είναι το μέλλον της έρευνας σχετικά με τη μηχανική μάθηση στην λογιστική και χρηματοοικονομικά; Αναγνωρίζουμε αυξημένο ενδιαφέρων στον συγκεκριμένο κλάδο από το 2015 με τις περισσότερες έρευνες να επικεντρώνονται είτε στην αγορά των Ηνωμένων Πολιτειών Αμερικής, είτε σε παγκόσμιο επίπεδο ενώ αύξηση ενδιαφέροντος κατά 600% παρουσίασαν οι Ασιατικές αγορές μεταξύ 2019 και 2021. Τα μοντέλα με επιβλεπόμενη μάθηση είναι με διαφορά τα πιο ευρέως χρησιμοποιούμενα (>60%) ενώ τα μη επιβλεπόμενα περιορίζονται κυρίως στην άντληση θεμάτων από κείμενο με την χρήση του «LDA» αλγορίθμου. Αυτό υποδηλώνει ότι η έρευνα σχετικά με την εύρεση νέων προτύπων είναι περιορισμένη. Αναγνωρίσαμε (10) κύριες κατηγορίες έρευνας για τις οποίες αναλύουμε τους προτεινόμενους αλγορίθμους με τα νευρωνικά δίκτυα να είναι

τα δημοφιλέστερα. Ακόμα, αναλύουμε (7) κατηγορίες που δημιουργήθηκαν από το «VosViewer» για τον σκοπό της βιβλιογραφικής σύζευξης αναγνωρίζοντας την κύρια θεματολογία, τα προβλήματα και τις ευκαιρίες που παρέχει η χρήση της μηχανικής μάθησης. Προβλέψεις τιμών σε μετοχές και παράγωγα, πρόβλεψη πτώχευσης και αποτίμηση περιουσιακών στοιχείων είναι η κύρια θεματολογία κατά την οποία τα παραδοσιακά μοντέλα είναι είτε αργά στην εφαρμογή τους είτε παρουσιάζουν μικρότερη ακρίβεια από τα μοντέλα μηχανικής μάθησης τα οποία διαχειρίζονται μη γραμμικά και πολλαπλών διατάσεων δεδομένα. Η αφομοίωση της μηχανικής μάθησης καθυστερεί από την διαφάνεια στον τρόπο λήψης αποφάσεων, τα κανονιστικά πλαίσια και από την λανθασμένη επιλογή εργασιών για αντικατάσταση. Η έρευνα στην λογιστική είναι περιορισμένη (μόλις 15% των άρθρων μας) σε λογιστικό έλεγχο, εκπαίδευση και δεξιότητες που θα πρέπει να αποκτηθούν για να ανταπεξέλθουν οι λογιστές στις σημερινές και μελλοντικές απαιτήσεις. Τέλος, αναφέρουμε μελλοντικά μονοπάτια έρευνας.

**Λέξεις – Κλειδιά**

Μηχανική Μάθηση, Λογιστική, Χρηματοοικονομικά, Βιβλιογραφική ανασκόπηση, νευρωνικά δίκτυα, Τεχνητή νοημοσύνη

# Table of Contents

# List of Figures

# List of Tables

# List of Abbreviations & Acronyms

| Abbreviation | Explanation |
|---|---|
| AI | Artificial Intelligence |
| AJG | Academic Journal Guide |
| CAPM | Capital Asset Pricing Model |
| CL | Computational Linguistics |
| CNN | Convolutional Neural Networks |
| GICS | Global Industry Classification Standard |
| LDA | Latent Dirichlet Allocation |
| LSTM | Long Short–term Memory |
| NLP | Natural Language processing |
| P2P | Peer to Peer |
| RQ | Research Question |
| SVM | Support Vector machine |
| SVR | Support Vector Regression |
| WoS | Web Of Science |

# 1. Introduction

The idea of "thinking" machines first proposed by Turing (1950) who introduced the "imitation game" by trying to assess the machine improvement on the field now known as Artificial Intelligence. Machine learning, subset of Artificial Intelligence, was first popularized by Samuel (1959) who defined it as *"field of study that gives computers the ability to learn without being programmed".* He managed to create, train and demonstrate the first self-learning program that was able to play Checkers as if it were a human (1959)**.**

Machine learning then, is far from new concept but was hindered by technological capabilities in order to be applied outside the scientific experimentation (Albanese, Crépey, Hoskinson, & Saadeddine, 2020). Turing (1950) forecasted that it would have taken close to fifty years as to be able for a machine to pass his imitation test. The challenges varied from processing power to storage and data retrieval costs which were prohibiting factors a decade ago in order to research and apply self-learning algorithms in a plethora of cases.

Previous studies have been done on the lifecycle of accounting Information Systems/Artificial Intelligence (AI) and researcher interest and adoption (Gray, Chiu, Liu, & Li, 2014)**.** The results indicated that in late 90s there was a decline of interest while on contrary Sutton et al. (2016)**,** who focused on a broader look of AI, found re-ignition of interest after 2003. Moreover, Sutton et al. (2016) mentioned the importance to investigate the machine learning application as a promised subset of AI. El-Haj et al. (2019)**,** studied the usage of Computational Linguistics (CL), a subset of AI, in financial disclosure and identified limiting factors, such as the small dataset, which applies to all machine learning algorithms. Weigand (2019), clustered papers as to explore machine learning usefulness in asset pricing. Even though it is unclear how the papers were collected and grouped, the (5) clusters were pricing equities, pricing derivatives, default prediction, appraising of Real Estate and Others. Another literature review to approximate asset pricing by focusing on stochastic discount factor belongs to Karolyi & Van Nieuwerburgh (2020) who stated the importance of researcher decisions in appliance of machine learning such as sample usage and hyper-parameters. Furthermore, they proposed to research community, to identify the differences between multiple dimensionality approaches as currently lead to different set of factors which fed to machine learning algorithms.

Our research contributes in several ways. Firstly, to the best of our knowledge this is the first literature review that investigates both finance and accounting exclusively under the prism of machine learning by combining both bibliographic coupling and literature review. That way we are able to get more insights on current literature status and propose future research. Secondly, our database is composed by only highly rated journals and papers to extract findings based on the top-notch academic research. We identify key journals, authors and countries on this field. We also extract research topics in which preferred machine learning algorithms are provided to help with future research. Furthermore, with bibliographic coupling, we analyze (7) clusters and propose new research paths.

We are trying to answer three interrelated research questions:

(RQ1) How is research on the impact of machine learning on accounting and finance research being developed?

(RQ2) What are the focus and critique of this corpus of literature?

(RQ3) What are the future avenues of machine learning in accounting and finance research?

In doing so, this dissertation is structured as follows. Section 2 presents the methodology conducted in this study. Section 3 incorporates in detail our findings in conjunction to our three research questions. In Section 4, we present our conclusions and limitations of present dissertation.

# 2. Methodology

Given that machine learning in accounting and finance is still an emerging field, we adopted bibliometric analysis in conjunction with literature review as the most appropriate method to approach our research (Rialti, Marzi, Ciappei, & Busso, 2019). As we will present in results section, this hybrid model allowed us to get insights about academic focus and find potential fields that are not yet explored in detail.

## 2.1 The Literature Search

There are several databases that could have helped us retrieving articles related to machine learning research in accounting and finance such as Web of Science (WoS), Scopus and Google Scholar. For our research, we collected exclusively our information from WoS database for two reasons. Firstly, using multiple sources could be unhelpful due to paper duplication (Harzing & Alakangas, 2016). Furthermore, even though WoS database includes fewer citations per article, its transparency on data inclusion and selection, promotes it as a more trusted source (Levine-Clark & Gil, 2021).

Our keyword selection was based on the idea to find the most used words (aka keywords) for accounting and finance as they are more likely to be found in articles and journals. Thus, we combined search engines tools and commercial ones to retrieve the general related terms and then filter and combine them.

The search engines are using data mining tools based on previous search logs to generate and cluster keywords that are suggested back to users as related searches (Thomaidou, Liakopoulos, & Vazirgiannis, 2014). Thus, search engines are able to provide search trends, number of searches for a specific term and cluster them based on relevance. Our preferred tool was GoogleAds Keyword Planner which even though is preliminary used by advertisers to find keywords before they publish their advertisements through web, is also found in Bibliometric analysis (Madani, 2015). The same procedure/tool has been extensively used by scholars in different research fields such as computer science (Bisikalo, et al., 2020) and Marketing (Dumitriu & Popescu, 2020). An extensive application of this procedure is illustrated in the study of Kaminski et al. (2020).

Our search for "accounting" and "finance" keywords has been executed on September 30rd and was limited between 2017-2021 for (10) countries as shown in Table 1.

| Country | Estimated Population using Google |
|---|---|

| | Services |
|---|---|
| **Australia** | 20.3M |
| **Brazil** | 165M |
| **Canada** | 34.3M |
| **France** | 70.7M |
| **Germany** | 73.2M |
| **India** | 428M |
| **Japan** | 76.1M |
| **Russia** | 68M |
| **United Kingdom** | 65.5M |
| **United States of America** | 312M |

**Table 1: Country Selection. Estimations provided by Google**

From the keywords provided, we decided to keep only the ones with more or equal to ten thousand monthly searches as provided by Keyword Planner. Moreover, we filtered out results related either to brands or country specific (e.g. tax authorities). Our keyword list was composed of (57) keywords related to accounting and (4) to finance even though the latter is having more searches overall. To further expand our searching query, we used Wordstream tool (WordStream Online Advertising Made Easy, 2021) to retrieve even more keywords in accounting and finance. This tool is employing the same principles that search engines are using in order to identify website topic; by releasing specialized programs named "Crawlers", is scanning the content of websites thus can identify the most relevant words for specific topic and cluster them. The above combination of techniques provided the most general keywords (Thomaidou, Liakopoulos, & Vazirgiannis, 2014) for our research. The total number of keywords (103) were further filtered and combined by using "regular expressions" when applicable. (Appendix A).

The filtering process was carried in three stages. In the first stage we filtered out repeated keywords by grouping and generalize them (e.g. ifrs 9, ifrs 11 both became ifrs). In the second stage we used "*" regular expression (means "Any") to further remove keywords with similar endings (e.g. audit and auditor became "audit*"). However, some keywords remained "as is" to avoid irrelevant article retrieval from our query. An example is the word "accountant" as "account*" would return also non accounting results. During the last stage, we removed the duplicate keywords and transformed them into a query (Appendix A). The finalized query was composed of (37) keywords related to accounting and finance and another (4) for machine learning.

We decided to limit our search by accepting only highly rated journals as provided by Chartered Association of Business Schools through Academic Journal Guide (AJG) (Chartered Assosiation of Business Schools, 2021). Thus, we included in our research journals rated as 2 or above as adapted by Harvey et al (2010).

| Criterion | Rule |
|---|---|
| **Keywords** | Having equal or more than 10M searches per month |
| **Database** | Web of Science |
| **Journal ratings** | 2 or above |
| **Documents** | Articles or Review Articles |

**Table 2: Search Criteria**

Our dataset was composed of (1098) articles out of which (239) complied with our rating criteria. Our final corpus is composed of (181) articles by removing the non-relevant articles for our search and (12) which we were unable to retrieve.

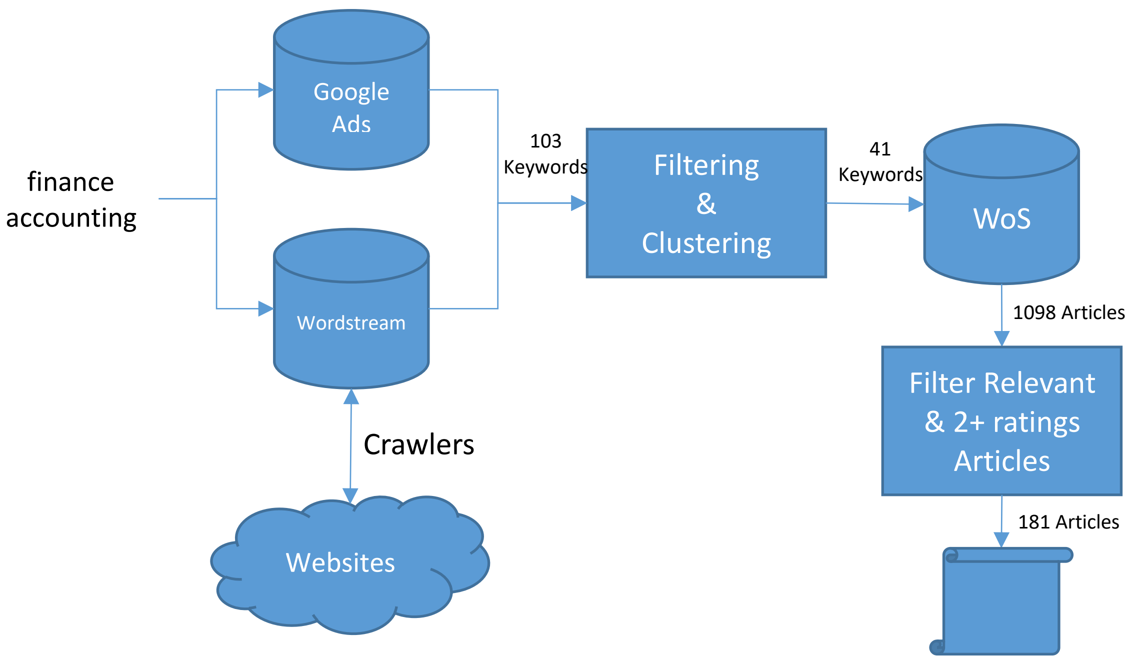


**Figure 1:Search & Filtering Process**

### *2.2 Bibliometric Analysis*

Co-citation, direct citation and bibliographic coupling can successfully cluster more than 92% scientific papers and articles (Boyack & Klavans, 2010). In direct citation, researcher investigates the relation between cited and citing documents. In Co-citation analysis, proposed by Small (1973), researcher links co-cited documents as referenced by previous papers. The higher the frequency the stronger the relation between the two cited papers.

For our research, we have chosen bibliometric coupling which overall generates the most accurate clustering (Boyack & Klavans, 2010). In bibliometric coupling, we are grouping articles based on the common citations used. The more citations found to be commonly used between articles, the stronger article relation is and higher the probability of addressing a shared topic. Our chosen tool to perform bibliographic coupling was VOSviewer (VOSviewer - Visualizing scientific landscapes, 2021) by Van Eck and Waltman (2010) as already used by scholars for literature review such as  Ciampi et al. (2021).

On top of bibliometric coupling, we analyzed all retrieved papers in order to get more insights for our research. We identified the purpose of each selected paper, industry that is referring\applied to and geographic area where machine learning algorithms were tested. As such, we are able to identify and categorize our selected papers into new dimensions and investigate opportunities for future research.

# 3. Results

### 3.1.1 (RQ1) How is research on the impact of machine learning on accounting and finance research being developed?

Given our selected dataset, we identified that the first highly rated paper related to machine learning and finance was written in 1998 in "Journal of Banking and Finance". As expected, only after 2014 the interest of accounting & finance researchers started to boom, which is still uptrend at the time of this writing, given the technological improvements that allowed usage of machine learning in conventional PCs. Our findings are presented in Figure 2 . We consolidated journals with less than (4) papers published under "Other" classification. Furthermore, in Table 3 we present the contribution of the (48) distinct journals, ordered top to bottom, which included on our dataset. "Quantitative Finance" is leading the field by publishing almost ¼ of papers in our corpus.

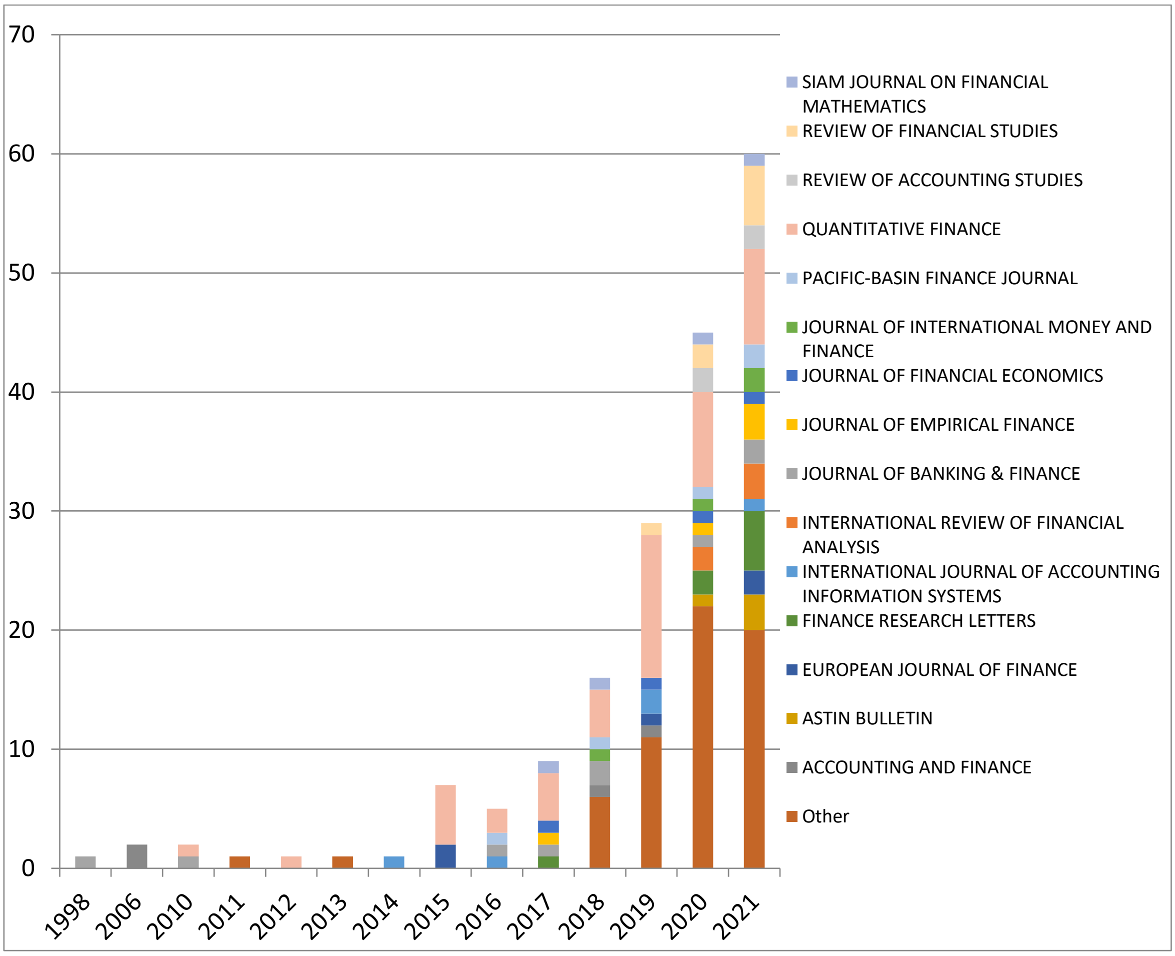


**Figure 2: Number of Papers per Journal and Published Year**

| Journal | Field | AJG Ranking | Number Of Papers | Percentage |
|---|---|---|---|---|
| **QUANTITATIVE FINANCE** | Finance | 3 | **45** | **24.86%** |
| **JOURNAL OF BANKING & FINANCE** | Finance | 3 | 9 | 4.97% |
| **FINANCE RESEARCH LETTERS** | Finance | 2 | **8** | **4.42%** |
| **REVIEW OF FINANCIAL STUDIES** | Finance | 4* | 8 | **4.42%** |
| **EUROPEAN JOURNAL OF FINANCE** | Finance | 3 | **5** | **2.76%** |
| **INTERNATIONAL JOURNAL OF ACCOUNTING INFORMATION SYSTEMS** | Account | 2 | 5 | **2.76%** |
| **INTERNATIONAL REVIEW OF FINANCIAL ANALYSIS** | Finance | 3 | **5** | **2.76%** |
| **JOURNAL OF EMPIRICAL FINANCE** | Finance | 3 | 5 | **2.76%** |
| **PACIFIC-BASIN FINANCE JOURNAL** | Finance | 2 | **5** | **2.76%** |
| **ACCOUNTING AND FINANCE** | Account | 2 | 4 | 2.21% |
| **ASTIN BULLETIN** | Finance | 2 | **4** | 2.21% |
| **JOURNAL OF FINANCIAL ECONOMICS** | Finance | 4* | 4 | 2.21% |
| **JOURNAL OF INTERNATIONAL MONEY AND FINANCE** | Finance | 3 | **4** | 2.21% |
| **REVIEW OF ACCOUNTING** | Account | 4 | 4 | 2.21% |

| | | | | |
|---|---|---|---|---|
| STUDIES | | | | |
| SIAM JOURNAL ON FINANCIAL MATHEMATICS | Finance | 2 | 4 | 2.21% |
| ACCOUNTING AUDITING & ACCOUNTABILITY JOURNAL | Account | 3 | 3 | 1.66% |
| FINANCIAL MANAGEMENT | Finance | 3 | 3 | 1.66% |
| INSURANCE MATHEMATICS & ECONOMICS | Finance | 3 | 3 | 1.66% |
| JOURNAL OF ACCOUNTING RESEARCH | Account | 4* | 3 | 1.66% |
| JOURNAL OF ASSET MANAGEMENT | Finance | 2 | 3 | 1.66% |
| JOURNAL OF BUSINESS FINANCE & ACCOUNTING | Account | 3 | 3 | 1.66% |
| JOURNAL OF CORPORATE FINANCE | Finance | 4 | 3 | 1.66% |
| JOURNAL OF PORTFOLIO MANAGEMENT | Finance | 3 | 3 | 1.66% |
| JOURNAL OF REAL ESTATE FINANCE AND ECONOMICS | Finance | 3 | 3 | 1.66% |
| RESEARCH IN INTERNATIONAL BUSINESS AND FINANCE | Finance | 2 | 3 | 1.66% |
| REVIEW OF QUANTITATIVE FINANCE AND ACCOUNTING | Finance | 3 | 3 | 1.66% |
| ACCOUNTING HORIZONS | Account | 3 | 2 | 1.10% |
| JOURNAL OF ACCOUNTING & ECONOMICS | Account | 4* | 2 | 1.10% |
| JOURNAL OF BEHAVIORAL FINANCE | Finance | 2 | 2 | 1.10% |
| JOURNAL OF FINANCIAL ECONOMETRICS | Finance | 3 | 2 | 1.10% |
| JOURNAL OF FINANCIAL STABILITY | Finance | 3 | 2 | 1.10% |
| JOURNAL OF INTERNATIONAL FINANCIAL MARKETS INSTITUTIONS & MONEY | Finance | 3 | 2 | 1.10% |
| JOURNAL OF RISK AND INSURANCE | Finance | 3 | 2 | 1.10% |
| MATHEMATICAL FINANCE | Finance | 3 | 2 | 1.10% |
| ABACUS-A JOURNAL OF ACCOUNTING FINANCE AND BUSINESS STUDIES | Account | 3 | 1 | 0.55% |
| ASIA-PACIFIC FINANCIAL MARKETS | Finance | 2 | 1 | 0.55% |
| AUDITING-A JOURNAL OF PRACTICE & THEORY | Account | 3 | 1 | 0.55% |
| FINANCIAL ANALYSTS JOURNAL | Finance | 3 | 1 | 0.55% |
| FINANCIAL MARKETS AND PORTFOLIO MANAGEMENT | Finance | 2 | 1 | 0.55% |
| FINANCIAL REVIEW | Finance | 3 | 1 | 0.55% |
| INTERNATIONAL JOURNAL OF FINANCE & ECONOMICS | Finance | 3 | 1 | 0.55% |
| JOURNAL OF ECONOMICS AND BUSINESS | Finance | 2 | 1 | 0.55% |
| JOURNAL OF EMERGING MARKET FINANCE | Finance | 2 | 1 | 0.55% |
| JOURNAL OF FINANCIAL AND QUANTITATIVE ANALYSIS | Finance | 4 | 1 | 0.55% |
| JOURNAL OF FINANCIAL SERVICES RESEARCH | Finance | 3 | 1 | 0.55% |

| | | | | |
|---|---|---|---|---|
| JOURNAL OF RISK | Finance | 2 | 1 | 0.55% |
| REVIEW OF FINANCE | Finance | 4 | 1 | 0.55% |

**Table 3: Main journals published machine learning in accounting and finance**

The top (5) journals are related to finance which is an indication that machine learning in accounting is not as explored in academic research. Indeed only (27) papers were published through accounting journals out of (181), a merely 15%. Interestingly the most cited paper belongs to Li (2010) and cited (350) times, was published by "Journal of accounting Research". The top (5) cited papers and published journals are presented on Top 5 Cited PapersTable 4.

| Paper | Author | Journal | Number of Cites |
|---|---|---|---|
| The Information Content of Forward-Looking Statements in Corporate Filings-A Naive Bayesian machine learning Approach | Li, F | JOURNAL OF ACCOUNTING RESEARCH | 350 |
| Consumer credit-risk models via machine-learning algorithms | Khandani, AE; Kim, AJ; Lo, AW | JOURNAL OF BANKING & FINANCE | 164 |
| Genetic algorithms applications in the analysis of insolvency risk | Varetto, F | JOURNAL OF BANKING & FINANCE | 129 |
| News implied volatility and disaster concerns | Manela, A; Moreira, A | JOURNAL OF FINANCIAL ECONOMICS | 112 |
| Measuring Qualitative Information in Capital Markets Research: Comparison of Alternative Methodologies to Measure Disclosure Tone | Henry, E; Leone, AJ | ACCOUNTING REVIEW | 83 |

**Table 4: Top 5 Cited Papers**

We have noticed that a big proportion of papers have been applied\studied specific geographic region. We find as important to replicate the same experiments to other markets, or globally, in order either to verify results or study any potential differences between the two groups. Therefore, we apply geographical data filter by reviewing the data used per paper. In some cases, we identified papers used data from multiple countries under the same or different continent. In the first case, we mark those papers as "global" while the latter based on the continent. Our results can be found on Figure 3. Literature which does not specifically mentions the composition of dataset or does not use at all, has been excluded from our results.

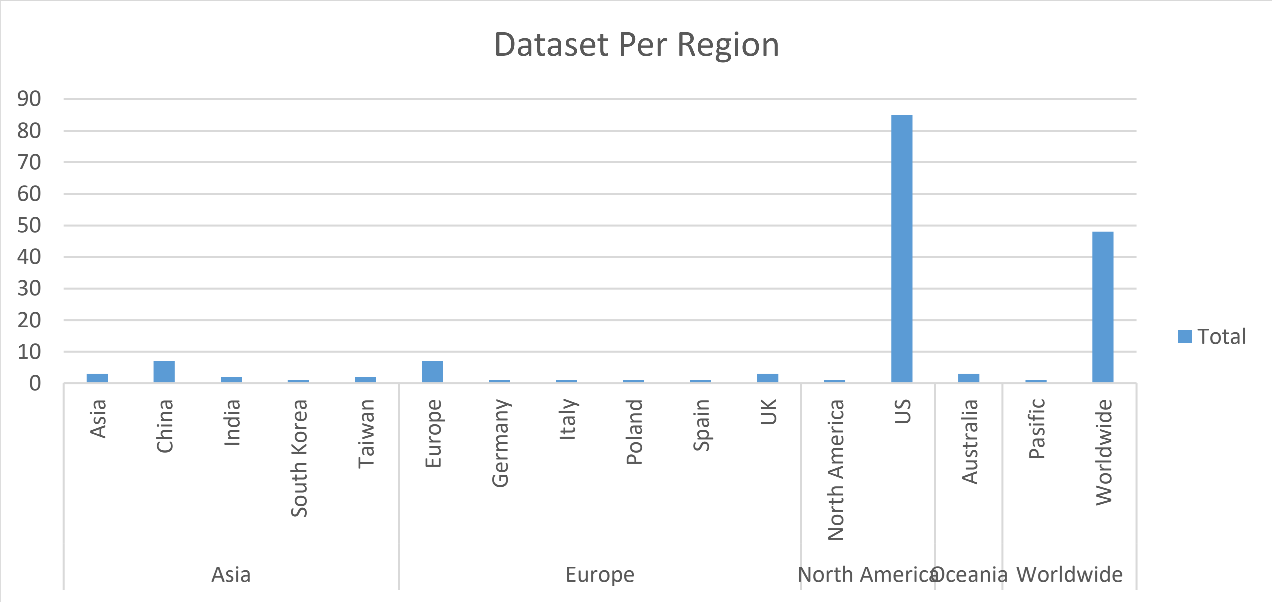


**Figure 3: Dataset per Region**

Literature is U.S data driven however (48) papers are having a more global scale approach. To identify any possible trend changes per year, we have constructed Figure 4 which depicts how the focus of scientists has changed over the years.

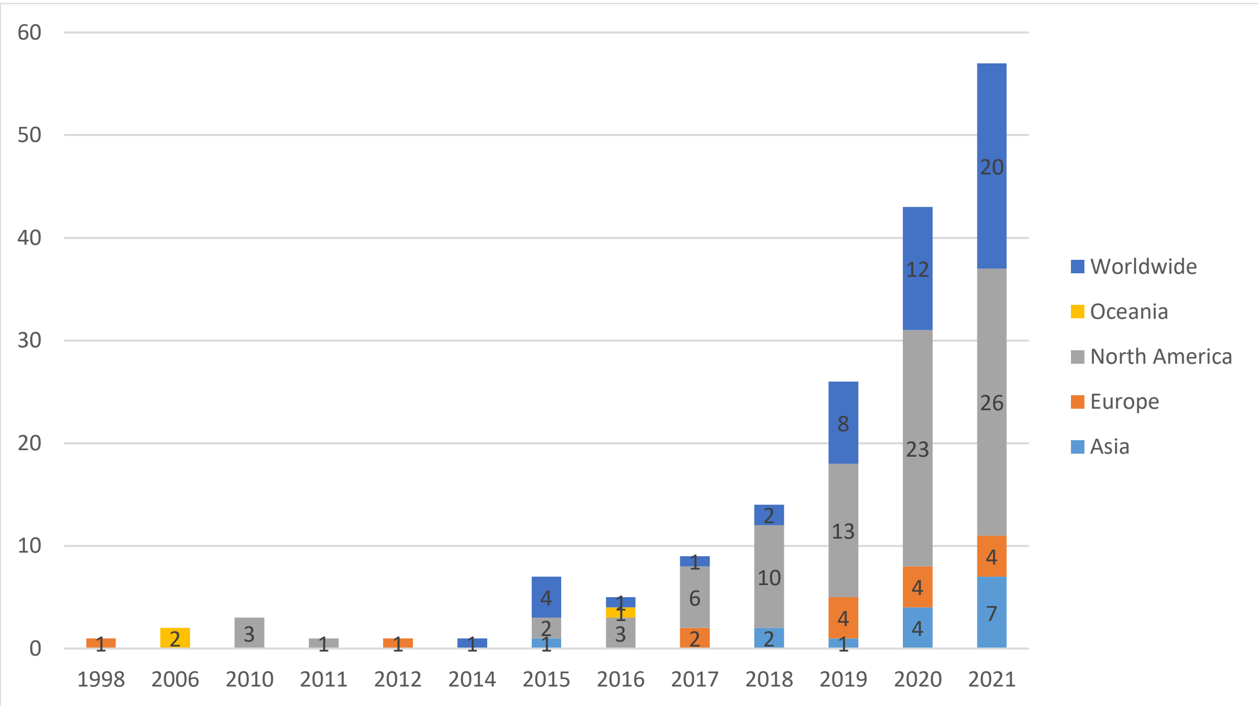


**Figure 4: Geographic Dataset Trends**

While number of papers produced over the last three years related exclusively to Europe is stable, literature produced overall is increasing. We also identified increased interest in Asia (600% increase during 2019-2021) and North America (100% increase) however the latter had only modest 13% increase for the past 2 years. Moreover, we haven't found any Oceania specific literature since 2016.In conjunction of the above with the global scale research increase as much as 67% over the last 3 years, might be an indication of moving towards of universal approach to machine learning in accounting and finance rather than geographic specific ones.

Lastly, we have also performed bibliographic analysis through VosViewer (VOSviewer - Visualizing scientific landscapes, 2021), to identify the key countries and average citations per paper for our corpus. The results can be found in **Error! Reference source not found.**Figure 5. It is evident that papers from Italy and United States are the most cited (with 21.86 and 20.85 average citations respectively) in contrary to China and Singapore having the lowest ones (2.9 and 4.25 respectively).

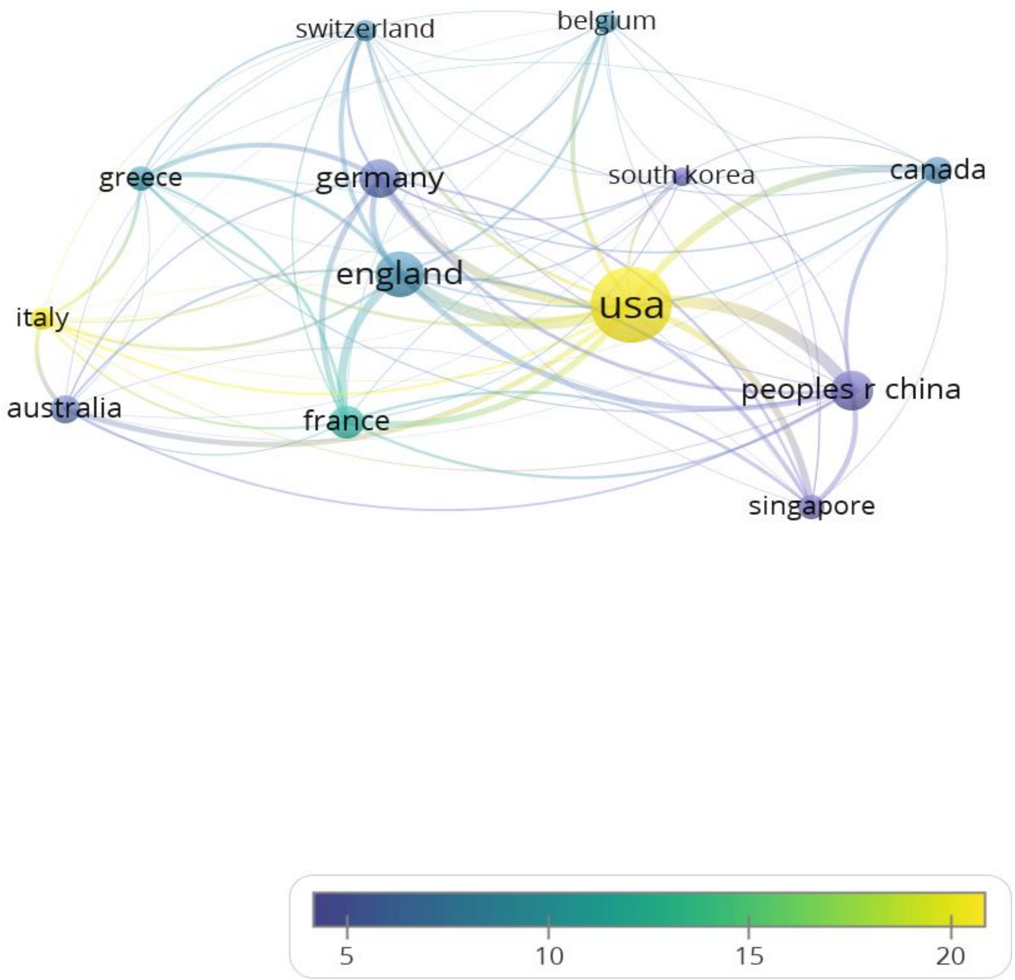


**Figure 5: Bibliographic Coupling Countries**

### 3.1.2 (RQ2) What are the focus and critique of this corpus of literature?

#### 3.1.2.1 Machine learning and Categories

Since Samuel (1959) who defined machine learning as *"field of study that gives computers the ability to learn without being programmed",* there is no universally accepted definition as it's often context specific (Gu, Kelly, & Xiu, 2020). Gu et al. (2020) provided a definition based on asset pricing *"(a) a diverse collection of high-dimensional models for statistical prediction, combined with (b) so-called "regularization" methods for model selection and mitigation of overfit, and (c) efficient algorithms for searching among a vast number of potential model specifications*". Apart Gu et al. (2020), we find only few papers (Wall, 2018; Sun, 2019; Rasekhschaffe & Jones, 2019) in our corpus which provided a clear definition of machine learning however their definitions differ. Therefore, this leaves an academic gap for at least two reasons. Firstly, it's not clear whether or not, all papers define the in the same way machine learning boundaries to distinct it from the more generalized AI or the purely statistical models. Thus, differences between samples of literature reviews are expected. Secondly, it's on the author's discretion to decide if a new algorithm or methodology is part of machine learning or not. As in our paper we used Samuel's (1959) definition by including statistical regression models, we have excluded "explicitly programmed" methods as considered part of AI. The first AI definition belongs to John McCarthy who defined it as "*the science and engineering of making intelligent machines."*.

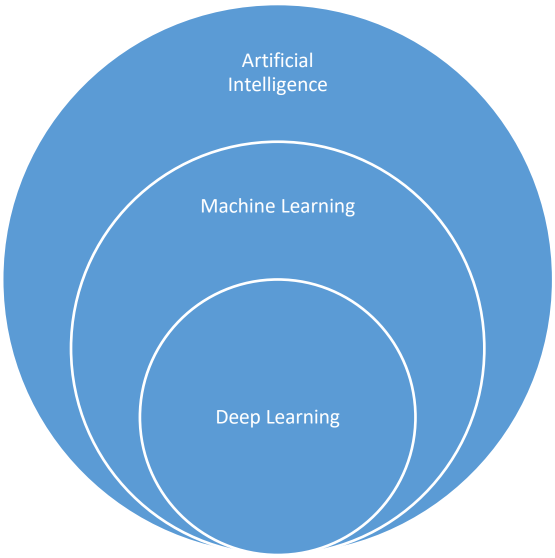


**Figure 6: Artificial Intelligence & Machine Learning**

In machine learning, in order for the model to "learn" and perform the specified task, needs to be trained and validated as to calculate achieved accuracy and proceed with any adjustments if needed. Therefore, the dataset is split into three sets; training, validation and testing set. With training the model is fit with data as to "learn" while with validation we provide the algorithm with dataset that has not been parsed before in order to calculate its accuracy before performing the actual task such as classification, regression with the validation set.

There are three main categories of models based on how are trained and the problem at hand. The first category called "Supervised" is composed by models that require labeled dataset in order to be trained. That way the model can identify if it has performed as expected and adjust accordingly as to improve. One example is the plant identification in which the machine learning model is provided with pictures of plants and relevant plant names in order to map them. With "Supervised" models, we aim to solve mainly classification and regression problems. The second category includes models called "Unsupervised" where training data are not labeled, therefore they are used mainly to identify patterns and hidden features. Clustering, association and dimensionality reduction are common problems addressed through those models. Last category is "Reinforcement" learning where algorithms are reacting to their environments and learning on their own by try end error. The scientist is providing rules, reward and penalty rules that are considered by the "agent" of the algorithm in order to distinct correct from wrong actions. This method is used mainly in gaming industry and robotics for navigation. We sum up the differences between the three categories in Table 5.

| Criteria | Supervised | Unsupervised | Reinforcement |
|---|---|---|---|
| Training Data | Labeled Data | Unlabeled Data | No-predefined data |
| Problem Types | Regression, Classification | Clustering, Dimensionality reduction, Association | Exploration, Finding best move |
| Aim | Calculation | Pattern Discovery | Learn series of Actions |
| Application Examples | Forecasting, Object Recognition | Recommendation systems, news categorization | Gaming, self-driven cars |

**Table 5 Machine learning Categorization**

## 3.1.2.2 Machine learning Models Used in Accounting and Finance

Having clarified the categorizations of machine learning models, we now identify the types of problems (e.g. classification, regression) literature is focused on and models that have been tested and used by researchers. Some models however, depending on what the problem is at hand, could be classified differently. Neural networks for example can be used for classification such us predict stock price movement (Upwards, Downwards, Sideways) or for clustering by finding fraud patterns. In our dataset we have identified (79) distinct models after removing 3rd party proprietary services such as "Amazon Rekognition" (Corazza, Truant, Scagnelli, & Mio, 2020) and IBM Watson (Hrazdil, Novak, Rogo, Wiedman, & Zhang, 2019). Also new models created are omitted, as not commonly used or named. We distribute our models by categorizing them in two dimensions; machine learning categorization (Supervised, Unsupervised, Reinforcement) and algorithm/problem type (e.g. regression, clustering, classification). Results can be found in Figure 7 by taking into consideration the times that algorithms have been used in our corpus. Supervised models are by far the most common which indicates tendency for solving regression or classification problems rather than finding new patterns to specific problems. As some papers are exploring more than one models, we have reviewed the proposed solution in each paper to extract the problem type, algorithms used and categorize appropriately. Our findings as shown on Figure 8 are confirming that Regression and Classification approximation to accounting and financial problems are the most applied.

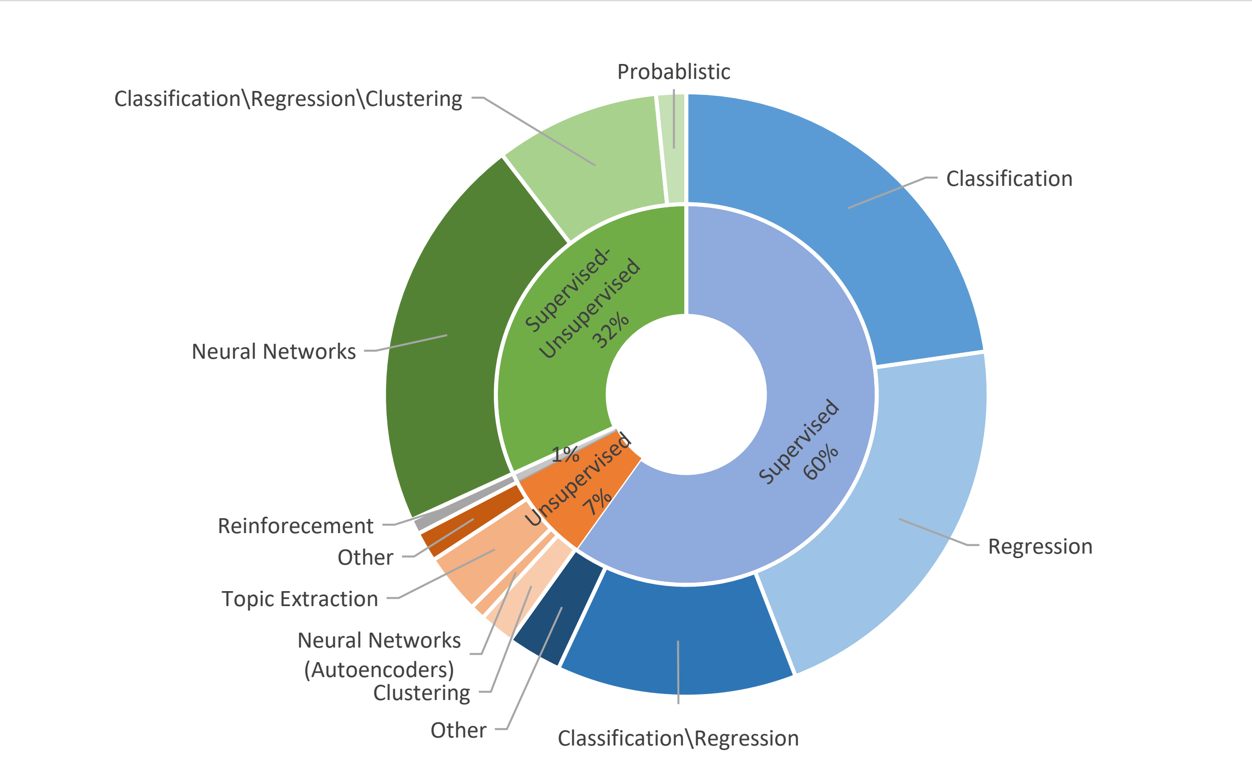


**Figure 7: Machine Learning Model Distribution**

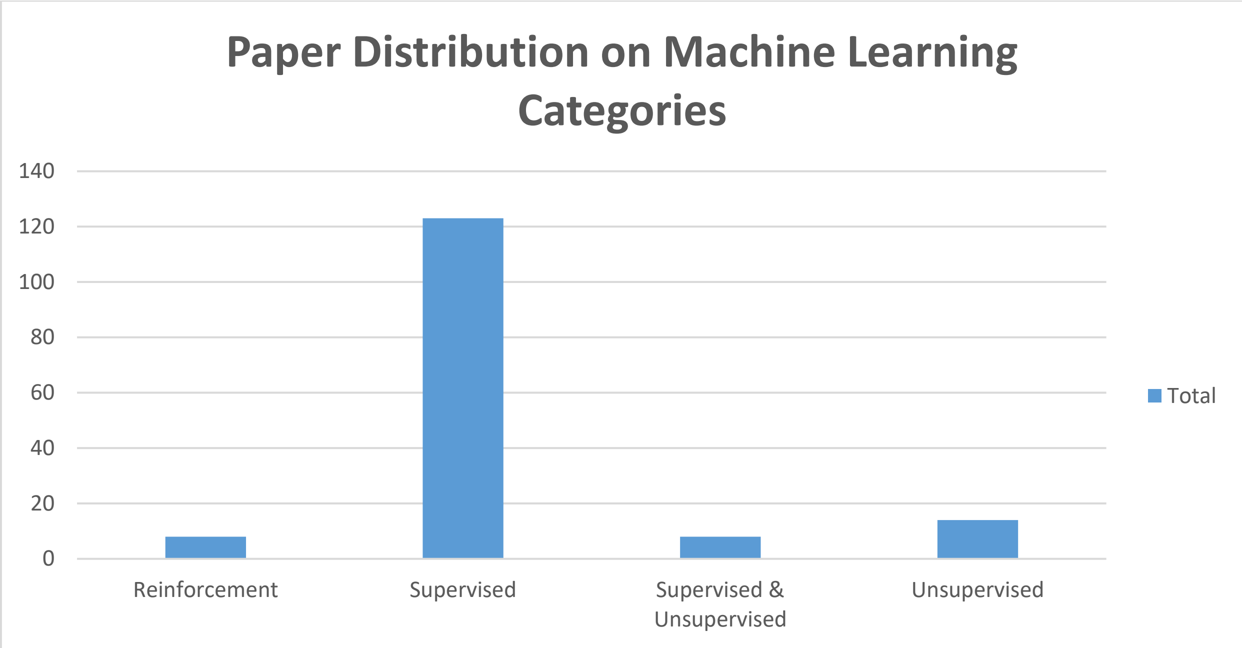


**Figure 8: Paper Distribution on Machine Learning Categories**

"Supervised & Unsupervised" category is composed by papers which either combine both methods, or compares them to explore differences for possible problem solution. The total number of machine learning algorithms found are summarized in Table 6.

| Algorithm | ML Category | Sub Category |
|---|---|---|
| AdaBoost | Supervised | Classification |
| ARIMA | Supervised | Regression |
| Autometrics | Supervised | Regression |
| Bagging | Supervised | Ensemble |
| Bayesian Network | Supervised | Probablistic |
| Binary switch portfolio | Supervised | Other |
| Boosted Linear Model | Supervised | Regression |
| BPNN | Supervised\Unsupervised | Neural Networks |
| C4.5 | Supervised | Classification |
| C5.0 | Supervised | Classification |
| CART | Supervised | Classification\Regression |
| Classification Trees | Supervised | Classification |
| CNN-LSTM-Attention | Supervised\Unsupervised | Neural Networks |
| Convolutional Neural Network | Supervised\Unsupervised | Neural Networks |
| Decision Tree | Supervised | Classification |
| Delta Boosting | Supervised | Regression |
| Discriminant Analysis Models | Supervised | Classification |

| | | |
|---|---|---|
| ElasticNet Regression | Supervised | Regression |
| ELM | Supervised\Unsupervised | Neural Networks |
| ENN | Supervised\Unsupervised | Neural Networks |
| Entropy & Linear Regression | Supervised | Regression |
| exponentially weighted average | Supervised | Other |
| external service | Unknown | Unknown |
| Extra Trees | Supervised | Classification |
| Extreme Gradient Boosting | Supervised | Classification |
| Factorization machines | Supervised | Classification |
| FCN | Supervised\Unsupervised | Neural Networks |
| FNN | Supervised\Unsupervised | Neural Networks |
| Gaussian Process Regression(GPR) with Tree | Supervised | Regression |
| Gaussian Process Regression—Exact Integration | Supervised | Regression |
| Gaussian Processes | Supervised | Classification\Regression |
| Generalized Additive Models | Supervised | Regression |
| Generative Bayesian Neural Networks | Supervised\Unsupervised | Neural Networks |
| GPIRL | Reinforcement | Reinforcement |
| Gradient Boosting | Supervised | Classification |
| GRNN | Supervised\Unsupervised | Neural Networks |
| Heterogeneous Autoregressive | Supervised | Regression |
| Markov Model | Unsupervised | Probabilistic |
| Hierarchical Clustering | Unsupervised | Clustering |
| JNN | Supervised\Unsupervised | Neural Networks |
| K-Means | Unsupervised | Clustering |
| K-nearest-neighbors | Supervised | Classification\Regression |
| K-Prototypes | Unsupervised | Clustering |
| LASSO | Supervised | Regression |
| Latent Dirichlet Allocation | Unsupervised | NLP |
| Le-Net | Supervised\Unsupervised | Neural Networks |
| Linear Discriminant Analysis | Supervised | Dimensionality Reduction |
| Linear Regression | Supervised | Regression |
| LNIRL | Reinforcement | Reinforcement |
| Logistic Regression | Supervised | Classification\Regression |
| LogitBoost | Supervised | Classification |
| LSTM | Supervised\Unsupervised | Neural Networks |
| LSTM-RNN | Supervised\Unsupervised | Neural Networks |
| Max Entropy | Supervised | Classification |
| Maximum Entropy Classifier | Supervised | Classification |
| Naive Bayes | Supervised | Classification |
| NARX | Supervised\Unsupervised | Neural Networks |
| Neural Networks | Supervised\Unsupervised | Neural Networks |
| Own Creation | Unsupervised | Multiple |
| Partial Least Squares | Supervised | Regression |
| PCR Regression | Supervised | Regression |
| PNN | Supervised\Unsupervised | Neural Networks |
| Principal Component Regression | Supervised | Regression |
| Principal convex hull analysis(PCHA) | Unsupervised | Archetypal Analysis |

| | | |
|---|---|---|
| Proprietary | Unknown | Unknown |
| QDA | Supervised | Classification |
| Random Forest | Supervised\Unsupervised | Classification\Regression\Clustering |
| RBFNN | Supervised\Unsupervised | Neural Networks |
| Recurrent Neural Network | Supervised\Unsupervised | Neural Networks |
| Regression Trees | Supervised | Regression |
| Reinforced Urn Process | Reinforcement | Reinforcement |
| Resnet-50 | Supervised\Unsupervised | Neural Networks |
| Ridge Regression | Supervised | Regression |
| Stacked Denoising Autoencoder | Unsupervised | Neural Networks |
| Stacking | Supervised | Classification |
| Genetic Programming | Supervised\Unsupervised | Probabilistic |
| Support Vector machine | Supervised | Classification |
| Support Vector Regression | Supervised | Regression |
| TRF Regression | Supervised | Regression |
| VAR | Supervised | Regression |
| RusBoost | Supervised | Classification |

**Table 6: machine learning Algorithms in Corpus**

### *3.1.2.2.1 Machine Learning Models*

Before diving into our findings for machine learning models used per topic, we briefly present the characteristics of the most commonly used ones in our corpus.

#### 3.1.2.2.1.1 Neural Networks

Inspired by biological neural networks to learn and extract features, this family of algorithms is referred as artificial neural networks. McCulloch and Pitts (1943) pioneered in this field as proposed a mathematical model of the neuron. After mid-1980s, they have been popularized by Rumelhart et al. (1985) who implemented the back-propagation algorithm and since then they have been applied for speech and image recognition, forecasting and natural language processing to name a few. The main components of the neural networks are layers, nodes, channels and activation functions. Neural networks are composed at least by three layers; the input layer, hidden layer\s and output layer. Each variable fit to the model is transformed to a node in the input layer. One or more hidden layers can exist where nodes, channels and activation layers are processing the input variables and inform the output layer for the outcome. When more than two hidden layers exists, the neural network is referred as deep learning network, a subset field of machine learning. The whole system is connected via channels that links the nodes together and applying a weight number on top of the prior node outcome which is processed by the next node. Each node, sums the input of the previous connected nodes and applies a node specific bias before engaging a function called "activation". The activation function is a filter that will either activate, or not, the next node and can be decided by the scientist. The two most common functions are "ReLu" and "tanh". This process is called forward propagation as information passes from previous node to the next one. With the back propagation of Rumelhart et al. (1985), a supervised neural network, can calculate the distance between the result and desired outcome and alter the channel weights as to improve accuracy.

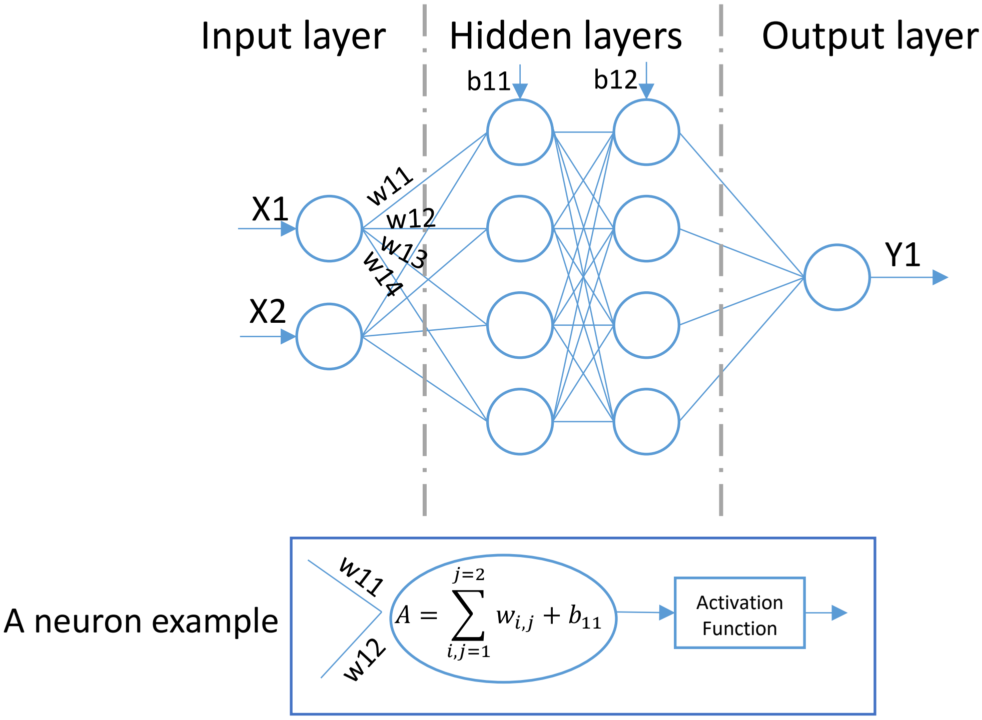


**Figure 9: Neural network (with (2) input variables & (2) hidden layers) & neuron example**

The described process covers the basics of a neural network however more advanced architectures such as convolutional neural network (CNN) and long-short term memory (LSTM) exists. CNN is applied for feature extraction especially when variables are not individually informative (Murphy, 2012). To transform the data in a format which can be processed by the network, it first uses convolution and pooling filters that extract features by simplifying data and removing the data "noise". Therefore, those networks are often used for image and phase recognition. LSTM on the other hand, was introduced by Hochreiter and Schmidhuber (1997) to deal with the vanishing gradient problem by proposing architecture called memory cells.

The biggest advantages of neural networks, is the ability to handle non-linear data, to generalize and be applied for plethora of applications. However, they are computation expensive and require more data as to be trained in comparison to regression or tree-based models.

#### 3.1.2.2.1.2 Tree based Algorithms & Bagging

Tree based algorithms are the main building blocks for other algorithms and can be used for decision making (decision trees), regression and classification. They are structured top to bottom and consisted of nodes branches and leaves. The root node\s, is composed of the input variables while each branch denotes a decision point. In the last layer of a tree based algorithm, we find leaves (or end-nodes) which contain the output of this model. An enhancement that increases the accuracy of decision trees is called random forest (Breiman, 2001). This algorithm creates multiple decision trees by selecting samples with replacement (bagging), and input variables in a random way from our training dataset. Thereafter, during validation or testing phase, all created trees estimate the new dataset and proposed the outcome value. In case of a classification problem, the most common prediction is considered as the solution, while in regression, solution is the average of the of the decision trees outcomes.

The major advantages of tree based algorithms are that we can trace the outcome, find the independent variables that explain the most our outcome, and can be used for both regression and classification. While decision trees tend to overfit by creating depth trees and are data "sensitive", random forest with the expense of solution clarity, tend to mitigate this problem. The number of trees and variables to be used for training a random forest however, needs to be experimented on as to minimize variance.

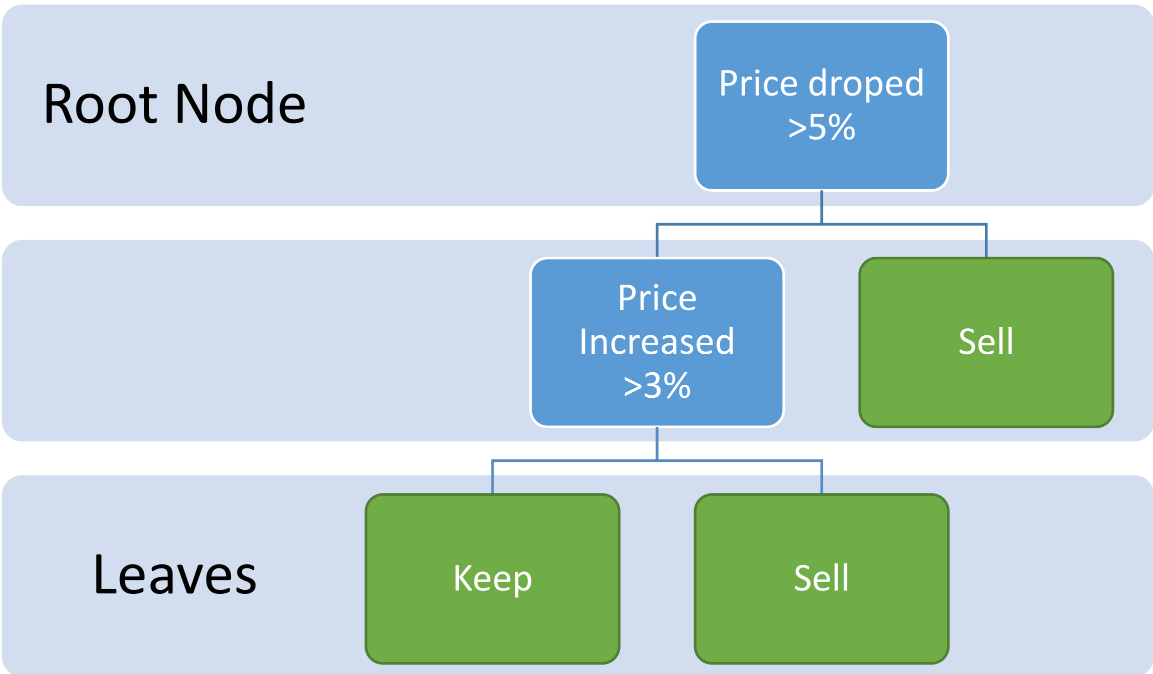


**Figure 10: Example of Decision Tree**

The Green boxes indicate leaves, the blue ones denote the decision nodes.

#### 3.1.2.2.1.3 Boosting

Boosting is an ensemble method which aims to convert a weak learner algorithm into a strong one. Similarly to Bagging, it uses random sampling with replacement however the learning process is sequential by evaluating each previous model and applying weights to ensure that the most accurate models are having the greater impact on outcome decision. Therefore, weak predictions are combined to a strong model using weights and majority vote. There are multiple variations in this category such as LogitBoost, Adaboost, Gradient Boosting and XGboost. Adaboost (Freund & Schapire, 1997) for example, discards weak models that have less than 50% accuracy.

Even though Boosting can increase the accuracy of a model, it is more prone to overfitting than Bagging due to weight mechanism.

#### 3.1.2.2.1.4 Regression Models

Simple regression models like linear and logistic regressions are commonly used in statistics before machine learning was formulized; nowadays are categorized under supervised learning family as during their training phase, the construction of best fit line, the dependent variable is known. Commonly used models are simple linear and logistic regressions, LASSO (Tibshirani, 1996), ElasticNet (Zou & Hastie, 2005), Support Vector Regression (SVR) and Ridge regression (Hoerl & Kennard, 1970). Tree based models can also perform regression activities such as random forest, that instead employing "voting" function that calculates the average of the results of the trees.

The least absolute shrinkage and selection operator model known as LASSO, and Ridge regression are both dealing with overfitting by applying penalty over the least squares line calculated during the training set. Therefore, they assume that the testing set will have more variance than the training one. On LASSO however, independent variables which are not significant can be eliminated in contrary to Ridge regression that is preferred when all independent variables are significant, as just performs coefficient reduction. To mitigate the problem of investigating all independent variables, a combination of the above methods was proposed named ElasticNet that applies both penalties (L1,L2).

Regression models are particularly important for predictions and can provide explanations on how each independent variable contributes to the calculation of the dependent one. However, they are susceptible to collinear problems, linearity cannot explain outliers and as independent variables increased their prediction reliability is decreased.

#### 3.1.2.2.1.5 Classification Models

While the result of Regression models are real numbers, in classification problems we aim to classify our population in pre-defined classes. Common classification models are Support Vector Machines (SVM), Decision Trees, Random Forest, K-nearest neighbors (KNN) and boosting algorithms such as RUSBoost. SVM model, aims to separate the data to classes through the creation of line called hyperplane. The optimal hyperplane, is the one which maximizes the distance between the closest points of the separated classes to the hyperlane. Thus, it aims to maximize the separation distance between the groups. When data are nonlinear, SVM is equipped with mathematical functions called kernel functions, which allows to transform data into higher dimensions as to be separated in a linear way. K-nearest neighbors model, is based on assumption that the closest the individuals are, the more related should be. By knowing the classes from the training set, any new individual is classified to a group in which has the closest distance. To do so, a parameter called "K" is set a priori to inform the algorithm on the number of classified individuals, which are closest to the new individual, will participate on the classification decision. The class with the most individuals will include the new datapoint. Therefore, in this method, experimentation is needed for the "K" parameter to find the optimal value for a given sample.

Classification models are quite fast and can be efficient for small datasets. However, they require a priori to train them with the final set of classes which limits our ability to uncover new ones.

#### 3.1.2.2.1.6 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) proposed by Blei et al. (2003) as a probabilistic model for topic extraction through three-level hierarchical Bayesian model. In machine learning, this algorithm is categorized as unsupervised model given that there is no need to train it by providing beforehand any word dictionary or topic categories. Thus, LDA is reducing data preparation, pre-work that is needed mainly in supervised models, and can be applied for any language. Another key advantage, is that it reduces the human bias by preventing explicit classification. To do so, it classifies large corpus by binding together terms that tend to be appearing together more often assuming that similar topics will use similar group of words. Lastly, LDA is capable of extracting more than one topic per document as demonstrated by Berninger et al. (2021) who identified topics on finance articles.

#### 3.1.2.2.1.7 Genetic Algorithms

Genetic algorithms have been introduced by Holland (1975) based on the concept of Darwin-type natural evolution. The basic idea is to find a set of optimal solutions by creating random population at first and combine characteristics, the gens, of individuals within the population to form a new generation. To identify the best performing individuals and discard the unacceptable solutions, criteria are set which composed the fitness function. There are three genetic algorithm operators that are performed after the evaluations of individuals to create a new generation. The first operation is called "Selection" where the algorithm selects the fittest individuals to pass their genes to the next generation. Even though it is not guaranteed that the best performing models will be chosen for this operation, they are having higher chance as scored higher during their evaluation. The next operation is named "Crossover" in which the selected individuals are exchanging genes to create new individuals. The last operation is executed with a given probability is called "Mutation" as randomly changes genes in the new individuals. Some variations are using "elitist selection" where instead of selecting individuals they simply pass the best solutions to the next generation without being mutated. The genetic algorithms are terminated either when fitness function evaluates an acceptable solution or after a provided number of generations.

One advantage of the genetic algorithms is the fact that as they are not deterministic, each execution may lead to different solution (if any). Moreover, they can be applied to problems where solution is not known beforehand or deterministic approach to find the best possible solution can be timing consuming. Their performance however is dependent on the complexity of fitness function to evaluate the individuals. Lastly, researcher needs to take into consideration on how the potential solutions will be assessed as this is a key component on the generation creation and algorithm performance overall.

#### *3.1.2.2.2 Corpus Classification per Topic*

We also classify our papers per topic to provide insights on the most used machine learning algorithms and potential model which substantially stands out. We have chosen (10) topics as shown on Figure 11:

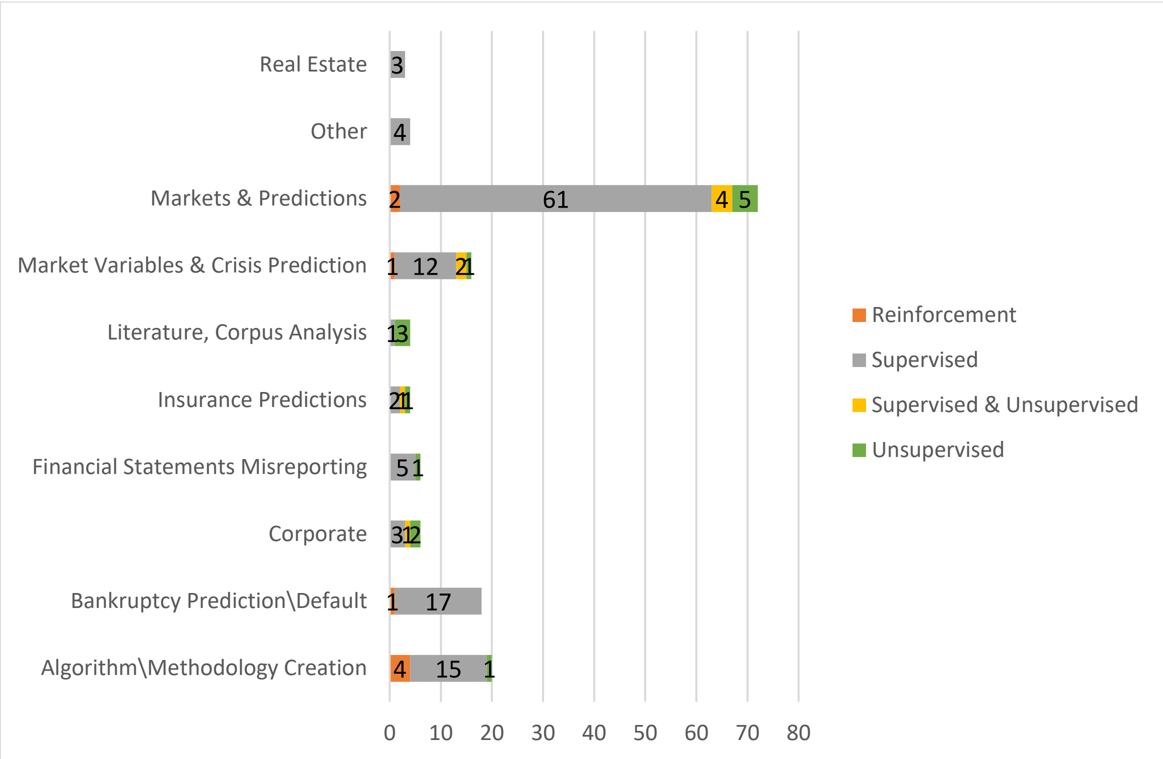


**Figure 11: Machine Learning Categories Per Topic**

On Real estate topic, we have classified (3) papers which approximate the hedonic pricing forecast estimation thru supervised algorithms but none of the papers are employing the same algorithm. Pace & Hayunga (2019) proposed a solution by comparing Gradient boosting with bagging, tree based algorithms where boosting improves accuracy and bagging minimizes variance. Bagging performed better in all (12) cases examined. In contrary Nowak et al. (2019) used LASSO, a regression based model, and Johnson et al. (2019) a combination of CNN and logistic regression to value a property based on photos and known variables. Thus, regression models are commonly used to predict property prices however our sample is small to conclude.

In Financial statement misreporting, the only unsupervised model found is LDA, commonly used to extract topics out of corpus. Brown et al. (2020) extracted topics from financial disclosures to identify misreporting patterns. Out of the papers which used supervised models, we find both classification and regression model comparison where RusBoost, a classification boosting algorithm that cater class imbalance (Seiffert, Khoshgoftaar, Van Hulse, & Napolitano, 2010), seems to outperform other models (Bao, Ke, Li, Yu, & Zhang, 2020) and (Bertomeu, Cheynel, Floyd, & Pan, 2020) except in the work of Perols (2011) who instead proposed another classification model named Support Vector Machine (SVM). By comparing the work of Perols (2011) and Bao et al. (2020), as both consider the serial fraud issue by employing SVM, we conclude that Perols (2011) results could have been different if used RusBoost especially as on his work we find great class imbalance by having only (51) cases of fraud out of almost 16K.

The "Corporate group" is consisted of general corporate related topics such as capital structure (Amini, Elmore, Öztekin, & Strauss, 2021), executive selection (Erel, Stern, Tan, & Weisbach, 2021), culture identification (Li, Mai, Shen, & Yan, 2020; Amin, Mohamed, & Elragal, 2021), impression management (Corazza, Truant, Scagnelli, & Mio, 2020) and venture capital communities existence (Bubna, Das, & Prabhala, 2019). Therefore, as there is plethora of topics and small sample, we cannot conclude of any prevailing method. We can identify though two unsupervised which both related to measure corporate culture through earning calls, by employing neural networks (Li, Mai, Shen, & Yan, 2020), or by tracking board tweets using LDA for topic extraction (Amin, Mohamed, & Elragal, 2021). The researchers in those cases could not rely on supervised models as they aimed to extract patterns instead of knowing beforehand the culture of the investigated companies. Bubna et al. (2019), combined Hierarchical clustering, unsupervised model, to realize (3) communities/groups and then feed results to CART and Random forest model, tree based classification models, to identify the key variable contributors for the previous clustering. In supervised models, we find both regression and classification models with neural networks, LASSO and Gradient boosting to be the most commonly used.

Under insurance group, a mixture of supervised and unsupervised models exists with the first group of papers using different algorithms such as regression, classification and neural networks to predict mortality (Wang, Zhang, & Zhu, 2021) and lapse risk (Loisel, Piette, & Tsai, 2021). In the work of the latter, once more a classification model, Gradient boosting, outperformed the regression models SVM and CART. LDA extracted topics from annual reports to realize digitalization progress (Fritzsch, Scharner, & Weiß, 2021) and proved gains for insurance industry. Another unsupervised model K-Prototypes was used for clustering in conjunction with a supervised model, Gaussian process, to reduce the processing time for valuating of annuities in comparison to the sole usage of Monte Carlo (Gan, 2013).

In literature and corpus analysis, unsupervised models are prevailing with researchers aiming to extract topics and patterns of their selected literature. With Tagme tool, trends in accounting research were identified (Marrone, Linnenluecke, Richardson, & Smith, 2020) while LDA extracted topics from financial articles (Berninger, Kiesel, Schiereck, & Gaar, 2021). For measuring the impact of citations, as scientists have already a pre-defined set of variables to explore, they preferred Gradient boosting classification model to measure their influence (Jones & Alam, 2019).

Our corpus on "Market Variables & Crisis Predictions" is composed by papers related to Covid-19, crisis prediction, market sentiment and market variables. All papers in Covid-19 subcategory are classified as supervised except Xue et al. (2021) who combined neural networks to extract keywords related to "Big data" and regression model to evaluate relation between Covid-19 and corporate strategies. Therefore, we classified this paper as "Supervised & Unsupervised". Regression model named ElasticNet has been proposed to analyze almost (100) market variables to identify the most significant factors for market immunity over pandemic (Zaremba, Kizys, Tzouvanas, Aharon, & Demir, 2021). CART and random forest have been successful in measuring the importance of financial and economic variables which can explain the daily market volatility in conjunction to Covid-19 (Baek, Mohanty, & Glambosky, 2020).

On crisis prediction group, multiple supervised models are put into test however there is no clear winner. In fact, logistic regression seemed to outperform random forest, SVM, neural networks and k-nearest algorithms in the work of (Beutel, List, & von Schweinitz, 2019) which contradicts with the findings of (Holopainen & Sarlin, 2017; Samitas, Kampouris, & Kenourgios, 2020). As machine learning methods, and especially random forest, outperformed logistic regression during in-sample training, there is high change of overfitting the models. In case of overfitting, machine learning models are achieving almost perfect score in a small dataset but fail to recognize out of sample patterns. This situation can happen in cases such as when data used for training purposes do not include the full spectrum of real-life cases, small training dataset and high level of parameterization of the model to increase accuracy. Scientists can mitigate such problems by creating synthetic data or use training techniques such as re-sampling method named cross-validation. In the work of Holopainen & Sarlin (2017) we find k-nearest-neighbors, with close second SVM, as the best model out of (12) to predict financial crisis while in Samitas et al. (2020) who tested 5 algorithms including k-nearest, SVM was the winner. Therefore, both classifications are found as good candidates for the above problems. Lastly, for this group we find 2 unsupervised models, LDA to parse quarterly earning transcripts in banking sector to find topics and k-means to cluster words into uncertainty group (Soto, 2020) teamed with linear regression model as validation and exploration of variable importance.

On market sentiment subgroup, both papers are combining Natural Language processing (NLP), a subset of AI, with supervised models. Specifically, Slapnik & Lončarski (2021) parse Moody's & Fitch sovereign credit rating reports to extract sentiment and subjectivity and compare to Naïve Bayes classifier while Rybinski (2020) feed regression models ARIMA and VAR from NLP indexes to extract language differences in financial forecasting. In Slapnik & Lončarski (2021) approach, manual classification had to be done to train the machine learning algorithm while this is not needed for the approach of Rybinski (2020).

In last subgroup, named "Market Variables", we have categorized papers with general market scope such as Brexit (Polyzos, Samitas, & Katsaiti, 2020), risk identification and categorization (Bender & Panz, 2021), risk measurement (Avramov, Li, & Wang, 2021), financial integration (Akbari, Ng, & Solnik, 2021) and asset mispricing (Caglayan, Pham, Talavera, & Xiong, 2020). Regression and classification models are the most common ones however Bender & Panz (2021) employed two unsupervised models; LDA to extract topics and k-means for clustering market risks. Lastly, neural networks were combined with reinforcement techniques to choose the best buyback contract options (Guéant, Manziuk, & Pu, 2020).

To summarize "Market Variables & Crisis Predictions" group, we find plethora of supervised models that depending on the case, are combined with unsupervised models to achieve a desired outcome. Regression, tree based, SVM, k-nearest and neural networks models seem to be the most used while in unsupervised category LDA for topic extraction out of textual corpus is the preferred choice of researchers.

Bankruptcy\Default prediction group, is composed by papers examining the crucial alerting variables for upcoming insolvency both in corporate\industry and country level. In this group, we find in all cases out of which neural networks compared to other methods, outperformed other supervised algorithms except in the case of modeling loss given default for corporate dept in which simple statistical model Tobit exceled (Olson, Qi, Zhang, & Zhao, 2021). However, only one hidden layer configuration was examined while two or more could have produced different outcome. Neural networks also handle in efficient way qualitative input variables to predict corporate bankruptcy (Lahmiri & Bekiros, 2019). In the remaining papers composed by supervised models, classification models are preferred with SVM, gradient boosting and other tree-based algorithms achieving the best scores. SVM in commercial real estate default prediction was more accurate than random forest and boosting algorithms (Cowden, Fabozzi, & Nazemi, 2019). Even though gradient boosting and random forest are both tree based, the first one performed better in lending industry by predicting customer credit score (Tantri, 2020). Reinforcement model is also found for prediction of recovery rates and times after insolvency (Cheng & Cirillo, 2018). The approximation of such problem by supervised model would require the knowledge of variables which impact the recovery times however this might not be possible for every single case.

The biggest group, "Market and Prediction" includes multiple topics from asset allocation, crypto currency, exchange rates and forecast of bonds, stocks, options and commodities. We have noticed that neural networks not only are commonly identified as having the best prediction results, but also are preferable in papers applied reinforcement techniques. In Figure 12 we depict the number of times each algorithm "Family" is employed.

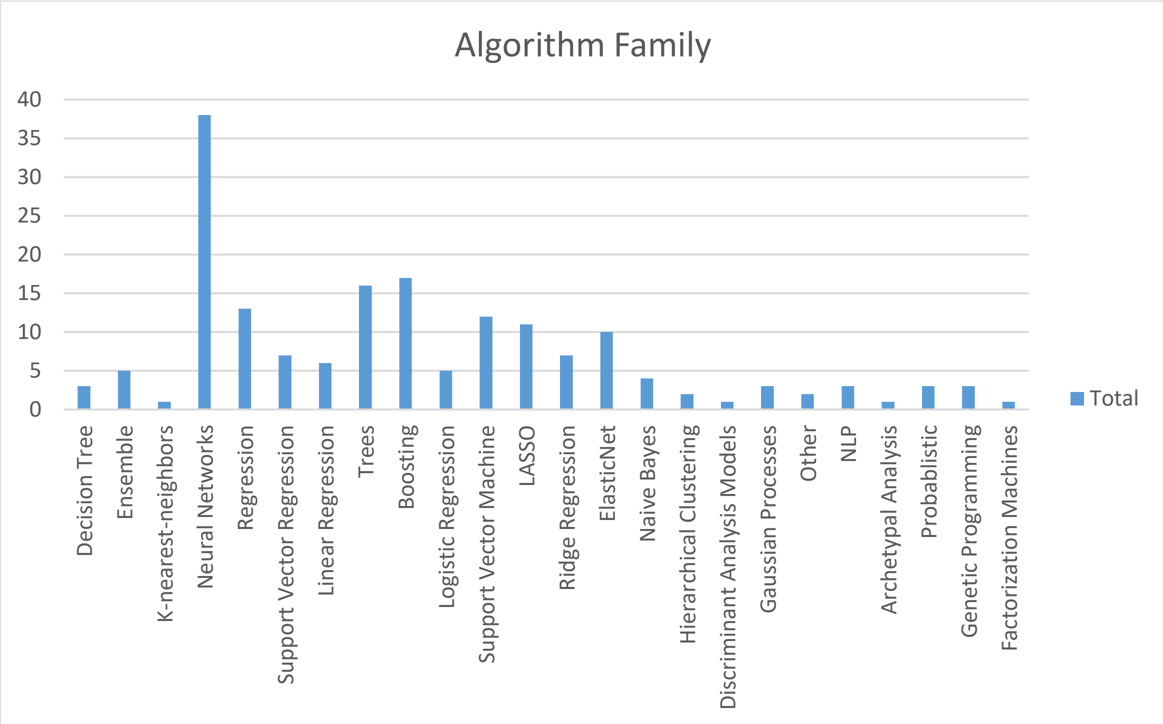


**Figure 12: "Market & Prediction" Algorithm Families**

In asset allocation, researchers investigate if machine learning can construct portfolios by either minimize risk (Pun & Wang, 2020), switch between strategies (Li, Chen, Feng, & Ying, 2016) and compare factors with asset allocations (Konstantinov, Chorus, & Rebmann, 2020). Only supervised algorithms have been found in this cluster with the majority being in regression category. This implies that researchers are familiar or testing specific variables which are important for optimizing the asset allocation. Except ElasticNet, LASSO, Support Vector Regression and Ridge regression, neural networks have been used for high-dimensional problem by calculating S&P500 daily returns for (24) stocks to compose the optimal portfolio (Tsang & Wong, 2020).

For options, in the majority of papers, neural networks were trained to price either American or European options, compare them with traditional methods Monte Carlo, Heston Model and Black-Scholes (Black & Scholes, 1973), and predict equity risk premium. Even though neural networks tend to outperform other algorithms in this group, overfitting can have huge impact as in the case of prediction of risk premium using implied volatility (Félix, Kräussl, & Stork, 2020). Both reinforcement papers, (Carbonneau & Godin, 2020) and (Zhang & Huang, 2021) are dealing with hedging by training neural networks and compare them with Monte Carlo simulations. The first paper is applied in American options while the second one in European. In the work of Goudenège, Molent and Zanette (2020), combination of tree and regression models are exercised to replace the traditional Monte Carlo simulation. Monte Carlo, due to its resource consuming and slow execution time (Risk & Ludkovski, 2018), cannot be easily applied in quick and adaptable environment while at the same time once machine learning has been trained, results can be produced almost simultaneously. De Spiegeleer et al. (2018) managed to replace the Monte Carlo with machine learning algorithms at small expense of accuracy. Black-Scholes model (Black & Scholes, 1973), has also being compared with machine learning algorithms to validate their fit for Option pricing (Zhang & Huang, 2021).

In exchange rates forecasting, corpus is limited to (3) papers however regression is almost exclusively used. In the work of Amat et al. (2018) coefficients are composed by "traditional" fundamentals however other coefficients could be as equally important to forecast exchange rates. In that essence, unsupervised models might be used in the future either to parse textual analysis for specific currencies on top of supervised models to uncover new useful patterns.

In Crypto currency market, the (4) papers in our corpus are investigating different aspects of the market thus supervised, unsupervised and combination of both techniques have been found. For Bitcoin price prediction, Liu et al. (2021) compared support vector regression, back propagation neural network and stacked denoising autoencoder to forecast price for the next day. Stacked denoising autoenconder, a deep learning method, outperformed the purely supervised models by trained at first in unsupervised way and then fined tuned in a supervised one. Except for Bitcoin, Anghel (2021) forecasted trend for multiple crypto currencies (up, down, sideways) comparing regression and classification models with trading patterns using only (5) variables. In the case of asset allocation purely composed by crypto, unsupervised model "Hierarchical Clustering" has been compared with traditional risk-based asset allocation strategies (Burggraf, 2021). The model was more accurate in capturing the volatility as clustering is able to identify new patterns. Tree based algorithm named "Genetic Programming", can be under both supervised and unsupervised category based on the selected fitness function, even though it finds the optimal solution based on probabilistic transition rules such as mutation and crossover. This algorithm is known for exploring problems in which we don't know the result beforehand, therefore to identify market efficiency due to high frequency trading, Manahov & Urquhart (2021) have chosen this algorithm instead of other unsupervised one.

The application of this algorithm is also demonstrated on supervised cases as well. It has been tested in case of corn/ethanol commodities trading to forecast price based on (13) market indexes and spreads (Dunis, Laws, Middleton, &

Karathanasopoulos, 2013). Researchers compared the algorithm with two neural networks under five statistical methods which indicated slightly better results in three of them. The results of Sermpinis, Laws & Dunis (2013), on Oil Brent and Gold value at risk (1) day prediction contradicts the Genetic algorithm superiority over neural networks. The difference between the two papers can be explained in the fact that the performance difference between the two algorithm categories were not decisive; Genetic algorithm as probabilistic it might take multiple times to be executed to construct the optimal tree-based solution, while at the same time neural network configuration was not standardized. The remaining (3) papers on commodities group, also focused on forecasting in a supervised way either in a daily timeframe (Creamer & Lee, 2019; Bonato, Gkillas, Gupta, & Pierdzioch, 2021) or intraday (Gkillas, Gupta, & Pierdzioch, 2020). SVM, random forest and Heterogeneous Autoregressive are the three different algorithms we find in this category.

The three papers in our corpus related to bonds, aiming to forecast or value them by comparing multiple regression and classification models. Random forest and neural networks managed to stand out with the second one to suffer in the case of catastrophe bonds due to the small dataset (Götze, Gürtler, & Witowski, 2020).

The group with the most papers, is related to stock price predictions achieved by measuring financial indicators, investor sentiment and news classification. Under the unsupervised category, market anomaly clustering via "Hierarchical Clustering" has been performed in order to identify the best variables to forecast stocks (Geertsema & Lu, 2020). Hidden Markov model, can also help in prediction process via pattern recognition of stock prices by investigating the short-term stock order imbalance (Wu & Siwasarit, 2019). To identify assets which are performing closely to a relevant index, unsupervised neural networks named autoencoders can be trained (Kim & Kim, 2019).

Our analysis on supervised models starts with the comparison of algorithms in various situations. We found (14) papers employing more than one method to compare results out of which (9) have researched the predictive power of the neural networks. In this category, researchers have realized that linearity is restrictive and cannot explain the prices well enough especially during financial crisis (Rasekhschaffe & Jones, 2019). Therefore, classification methods boosting, random forests and neural networks perform substantially better than regression models. Specifically, Rasekhschaffe & Jones (2019) have tested neural networks, gradient boosting, SVM and AdaBoost in different timeframes and proved that a composite of the algorithms is a better predictor than standalone predictions. Parcing limit order book to predict stock jumps (Mäkinen, Kanniainen, Gabbouj, & Iosifidis, 2019) or stock prices (Booth, Gerding, & McGroarty, 2014) within the next minutes is also part of our corpus. To predict price jumps within the next minute, Mäkinen et al. (2019) combined CNN, known for feature extraction and pattern recognition, with long short–term memory (LSTM), a neural network model that incorporates time feature by taking into account previous signals. The combination of those neural networks is powerful and outperformes the standalone usage of methods as CNN can extract patterns over a big dataset in a quick manner, in this case over (23k), and feeding LSTM which takes into account previous stock price by appling less weight as we go backwords in time. In contrary, to mitigate the impact of time in the stock prediction, Booth et al. (2014) retrain their neural network every (15) minutes. LSTM has also outperformed other neural netwoks in the volatility prediction of S&P500 index however LASSO regression model has been chosen to identify the most relevant variables to feed the models (Bucci, 2020). Pyo & Lee (2018) predict stock volatility based on last (10) months performance to construct a low risk Black-Litterman portfolio by comparing support vector regression model(SVR), gaussian model and neural network. Especially during crisis, neural network was more accurate on constructing a low risk portfolio with close second the SVR. In prediction of SSE50 index and CSI300, neural network was not always performing better than regression models even when different setups of hidden layer were used (Li & Mei, 2020) . We find this expected as the training dataset was small enough, less than (2k), and overfitting can be present in such cases. The last paper in which neural networks are compared, is investigating (6) financial statement variables to detect future cash flows for the next one or two years (Oz, Yelkenci, & Meral, 2021). All machine learning algorithms achived more than 97% prediction rate for the next year pedictions, droping to 82% for the second year, with neural networks, SVN and decision trees having the best performance (Oz, Yelkenci, & Meral, 2021). In summary, neural networks are having outstanding performance over other methods, if a big dataset is available, however when applicable, combined together with other algorithms can increase the predictability.

The next sub-group is composed by papers which are comparing different algorithms but neural networks are not among them. The majority of corpus has achived better results with classification models except in the case of Wolff & Neugebauer (2019) who compared tree based models, including boosting ones with regression models in prediction of S&P500 index. In this case however, only (168) observations were used to train the models therefore it is likely to have different results if the training set was extended (Wolff & Neugebauer, 2019). Index trading using both long and short positions, has been performed by comparing the boosting algorithms AdaBoost, LogitBoost with Logistic regression and Bagging (Creamer G. , 2012). Adaboost and LogitBoost, a modification of Adaboost proposed by Friedman et al. (2000) followed by Collins et al. (2002), both perform additive logistic regression while Bagging aims to reduce the variance. LogitBoost with Bagging as close runner up achived the best results therefore Creamer (2012) proved that boosting could have advantage over variance reduction and simple regression models. Sentiment identification over stock price has also been researched by taking into consideration investor comments (Knoll, Stübinger, & Grottke, 2018; Liu, Wang, Zhang, & Zheng, 2021). In both papers, we find SVM classification model being used however in Knoll et al. (2018) Factorization machines, a supervised classification model, managed to combine twitter data with S&P500 stock prices and achive better result than SVM strategy. Liu et al. (2021), recognized the importance of avoiding oversampling by creating sythetic data to balance the classified groups.

The last sub-group contains papers which either employed one machine learning method or combination to achieve the desired result. Eight papers in this category are related to news, volatility and textual analysis to predict future prospect of companies (Li F. , 2010), stock implied volatility (Manela & Moreira, 2017) intraday predictions (Renault, 2017; Calomiris & Mamaysky, 2019). Only the latter paper combined regression with unsupervised model as the majority of papers relied on manual sentiment classification to train their classification and regression models. In those cases, an NLP model introduction could have reduced the manual labor and still provide accurate results for the research. Gómez Martínez et al. (2018), combined NLP with supervised model to achieve positive returns by constructing a bot which could trade in both long and short positions based on media news. Moreover, Creamer (2015) compared quarterly portfolio re-allocation based on financial statements, by training classification models, with portfolio impacted by news and sentiment using classification and LDA unsupervised model. Both strategies outperformed market however the transaction costs on "following the news" strategy had impact on performance. Search engines can also indicate the investor interest on specific stocks and aid to predict prices combining market variables of (5) previous days with number of searches and feed an LSTM neural network (Zhang, Chu, & Shen, 2021). There is no preferred regression or classification model in this sub-category however we find papers classifying first the textual analysis result and then pass the sentiment as a variable to regression model (Renault, 2017; Agrawal, Azar, Lo, & Singh, 2018). Four papers are related to high frequency trading and/or examining limit order book to predict stock price movements in US markets (Kercheval & Zhang, 2015; Sirignano & Cont, 2019), news volatility that explains 5-min returns (Engle, Hansen, Karagozoglu, & Lunde, 2020) and impact of cancellation orders in markets (McInish, Nikolsko-Rzhevska, Nikolsko-Rzhevskyy, & Panovska, 2019). On those papers, classification models have been used however given the small sample there is no conclusion of preferred model.

The last group contains papers which modify existing machine learning algorithm or provide new methodology to be applied in finance and accounting. Out of (25) papers, (11) are related to neural networks and (4) to reinforcement learning. Starting with reinforcement learning, apart from Yang et al. (2015) who depended on Markov decision process to categorize trading accounts based on trading behaviors, Buehler et al. (2019) and Sirignano and Spiliopoulos (2017) employ neural networks for hedging and American options pricing respectively. Wang and Zhou (2020), develop a new framework for portfolio selection, based on time and mean-variance, which was then compared with neural network. Cheng et al. (2020), combined neural networks with Markov chain model to approximate the optimal dividend and reinsurance strategies applied to insurance industry. In the supervised models category, researchers are aimed to enhance existing regression algorithms such as SVR (Law & Shawe-Taylor; Fischer, Pohl, & Ratz, 2020) and LASSO (Devriendt, Antonio, Reynkens, & Verbelen, 2021).

Replacement of Monte Carlo either via Gaussian Processes for portfolio risk management (Risk & Ludkovski, 2018) or neural networks for derivative pricing (Liang, Xu, & Li, 2021) has also been explored. We find that new algorithms have been proposed for optimal portfolio creation (Pun, 2021) but tested against only a handful of existing machine learning algorithms or not at all.

In Table 7, we summarize our findings to help readers understand the commonly used machine learning algorithms in specific topics. However, as this is an emerging field, we expect new algorithms, improvements and combinations of them to be researched and surpass existing ones. Except the plethora of cases where neural networks have been applied, setting up between one and three hidden layers, we find interesting the fact that LDA is the most commonly used unsupervised model.

| Topic | Key Topic\s | Common Family\s | Proposed Algorithm\s |
|---|---|---|---|
| Algorithm\Methodology Creation | Replace existing algorithms | Neural Networks | Neural Networks |
| Bankruptcy Prediction\Default | Insolvency Prediction | Neural Networks | Neural Networks |
| Corporate | Culture Identification<br>Executive Selection<br>Capital Structure | Neural Networks<br>Tree Based Algorithms | Neural Networks<br>Random Forest<br>Boosting |
| Financial Statements Misreporting | Financial Statements Misreporting Detection | Boosting | RusBoost |
| Insurance Predictions | Lapse & Mortality Predictions | Neural Networks<br>Boosting | CNN<br>Gradient Boosting |
| Literature, Corpus Analysis | Topic Extraction | Unsupervised | LDA |
| Market Variables & Crisis Predictions | Covid-19 | Neural Networks<br>Regression<br>Classification | CART, ElasticNet<br>Random Forest,<br>Neural Networks |
| Market Variables & Crisis Predictions | Crisis Prediction | Unsupervised<br>Regression<br>Classification | LDA, KNN, SVM |
| Market Variables & Crisis Predictions | Market Variables | Neural Networks<br>Unsupervised<br>Regression<br>Classification | Neural Networks<br>LDA<br>LASSO<br>Random Forest |
| Markets & Predictions | Asset Allocation | Regression<br>Neural Networks | Regression Models, Neural Networks |

| | | | |
|---|---|---|---|
| Markets & Predictions | Options | Neural Networks | Neural Networks |
| Markets & Predictions | Exchange Rates Forecasting | Regression | Ridge Regression, SVR |
| Markets & Predictions | Crypto currency | All | All |
| Markets & Predictions | Commodities Forecasting | Regression Neural Networks | Neural Networks Heterogeneous Autoregressive |
| Markets & Predictions | Bonds | Tree Based Algorithms | Random Forests |
| Markets & Predictions | Stocks Forecasting | Neural Networks Classification Boosting | Neural Networks LogitBoost, SVM |
| Real estate | Hedonic Pricing Forecast | Regression | LASSO, CNN, Bagging |

**Table 7: Summary of machine learning algorithms & corpus groups**

### 3.1.2.3 Sector\Industry Filter

Another important filter to categorize our literature, is regarding the industry in which the selected papers are applied. Specifically, by reviewing each paper, we categorized them based on the data used to fit machine learning models. Thus, we are able to identify on which industries researchers are focused on and which ones are neglected under the accounting and finance prism. Our classification is aligned with Global Industry Classification Standard (GICS) (S&P Global Market Intelligence delivers essential information so you can make decisions with conviction, 2021) which defines (11) sectors and (24) industry groups. As a big portion of papers have applied machine learning to estimate stocks and market indexes, we have excluded them from our results given the fact that they are composed by multiple industries. Papers focused on industry agnostic accounting topics, are excluded as well. Results can be found on Figure 13, while major topics per industry group are presented on Table 8.

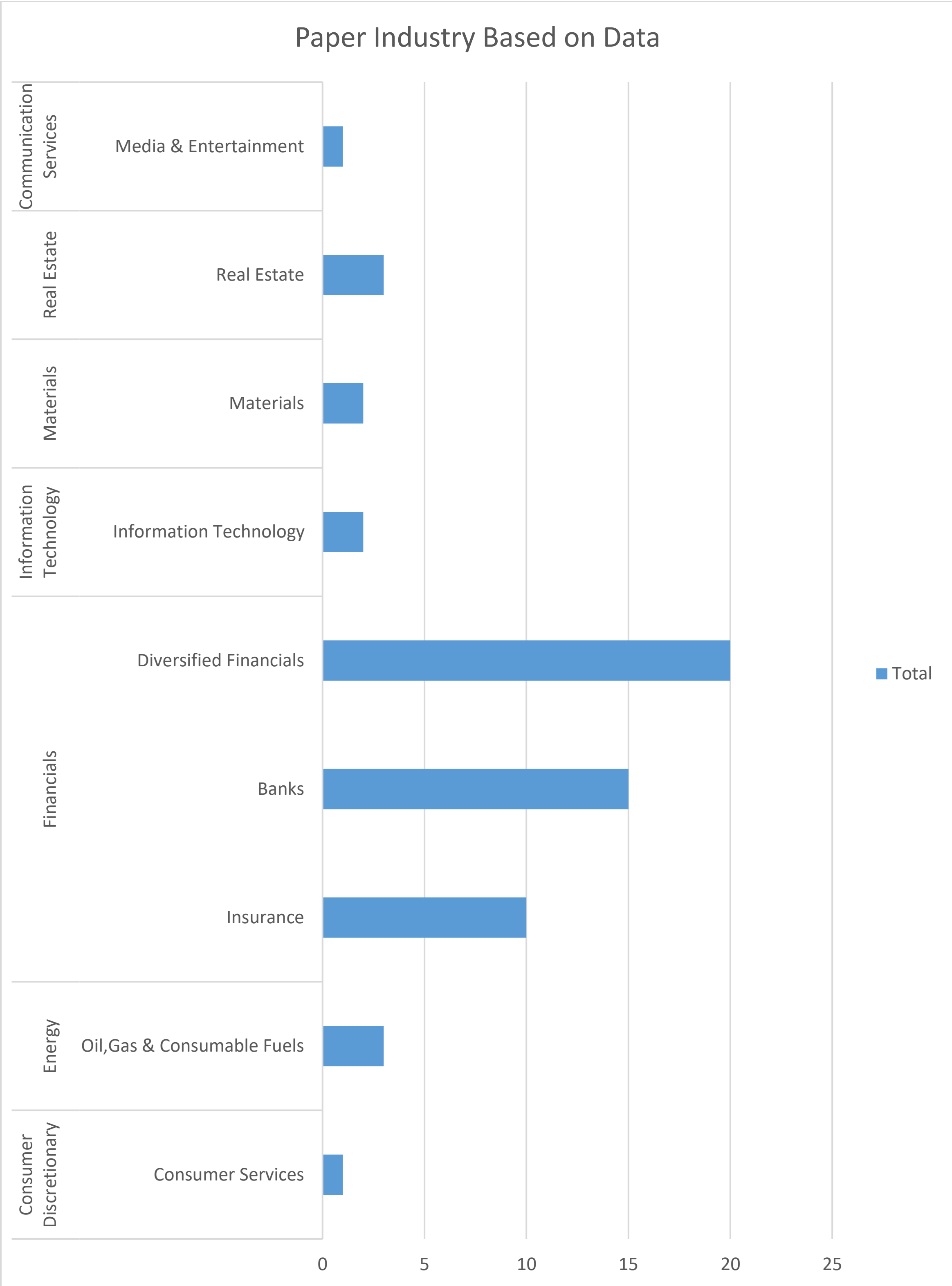


**Figure 13: Sectors & Industries**

In Real Estate, we find all papers focused on prediction of Hedonic prices based on pictures, house attributes and/or advertisements posted by real estate agents. The researchers, either trying to identify the most important variables for house pricing or explore new ways to value real estate which were unable to be evaluated by traditional statistical models. The Curb evaluation by using pictures taken from Google maps (Johnson, Tidwell, & Villupuram, 2019) and textual analysis (Nowak, Price, & Smith, 2019) are both examples of research that are enabled by the usage of machine learning.

On Communication Services sector, out of the two industry groups, we found papers only on "Media & Entertainment". On that group, social media are used, such as Twitter, to predict stock prices (Renault, 2017) through calculation of the sentiment of investors measured by their comments on specific stocks. The key takeaway is the speed and volume that is offered for intraday trading decisions which humans cannot cope up with.

For information technology, we do not to distinct the industry group as per GICS given that papers are using data from NASDAQ. Both papers, are exploring the usage of limit orders to predict price movements given that NASDAQ is electronic exchange with high-frequency limit and cancellation transactions (Sirignano & Cont, 2019). Intraday trading is again the main focus, achieved by scanning continuously the order book and asses the price movement.

In the Energy sector, all papers are predicting Oil, or other related commodities, prices either intraday (Gkillas, Gupta, & Pierdzioch, 2020) or in daily timeframe (Sermpinis, Laws, & Dunis, 2013; Creamer & Lee, 2019). We haven't found any paper related to Energy companies per se.

In contrary we find in Material sector (2) papers with different scope of research. Wu et al. (2006) proposed financial ratios and economic factors to predict bankruptcy for Australian materials industry while Bonato et al. (2021) used investor happiness index as approximation to forecast gold pricing.

In Consumer Discretionary, Corazza et al. (2020) investigated the sustainability report of Costa Crociere before, during and after Costa Concordia disaster to realize impression management exercised by using both textual and image recognition patterns.

As expected, the majority of literature is classified to "Financials" sector which is the only one having representatives from all three industry groups (Banks, Insurance, Diversified Financials). In banks group, we have identified (14) papers out of which (7) are related to bankruptcy prediction and credit risk. Le & Viviani (2018) and Viswanathan et al. (2020) are predicting bank insolvency using financial ratios while Varetto (1998) is predicting in 3 year timeframe corporations insolvency. In credit risk, we find Butaru et al. (2016) and Khandani et al. (2010) predicting credit card delinquency using anonymous data however using different machine learning algorithms. Appliance of machine learning has also been investigated in order to increase loan acceptance and keeping default rate under specific threshold in comparison to loan assistance (Tantri, 2020). The last paper in this sub-group, is dealing with the recovery process and specifically aims to predict time and recovery level by splitting recovery process into phases in conjunction to down payments (Cheng & Cirillo, 2018). Fintech is also examined as potential threat for traditional banks given the flexible usage of technology and soft data (Jagtiani & Lemieux, 2019) for their operations while current banks' response is limited (Boot, Hoffmann, Laeven, & Ratnovski, 2021). Chen et al. (2019) on the other hand, points out the importance of measuring the value of technological innovations in Fintech and how they affect banking industry. Two papers are predicting banking crisis using economic variables (Beutel, List, & von Schweinitz, 2019) or textual analysis of conferences (Soto, 2020). The importance of machine learning in banking daily operations has been studied by (Kumar, Muckley, Pham, & Ryan, 2019) who managed to protect elderly clients from external financial fraud by identifying them accurately. Lastly, Guéant et al. (2020) identified the optimal strategy for buying back shares based on VWAP-minus programs.

In the second industry group "Insurance" we find literature employed machine learning or proposed new algorithms mainly for predictions such as mortality (Wang, Zhang, & Zhu, 2021), lapse behaviors (Loisel, Piette, & Tsai, 2021), claims (Devriendt, Antonio, Reynkens, & Verbelen, 2021; Ding, Lev, Peng, Sun, & Vasarhelyi, 2020) and misrepresentation (Li, Song, & Su, 2021). Frintzsch et al. (2021) investigated the impact of digitalization that resulted in higher profits for insurance companies by parsing annual reports of (86) U.S insurance companies. Cheng et al. (2020), applied hybrid machine learning methods to approximate the optimal dividend and reinsurance strategies for insurance companies. In "Insurance" industry, the creation of claim simulator is not uncommon topic, however as large data sets are not always available, Avanzi et al. (2021) proposed a simulator using a mixture of real life and non-real-life data. Moreover, as it is common for insurance financial data to be imbalanced by having one class dominating, Lee (2020) proposed a new approximation procedure to address this problem. Monte Carlo in valuation of variable annuity, is quite slow process therefore, Gan (2013) managed to speed up the valuation process by (14) times using machine learning classification methods.

The last industry group, "Diversified Financials", is composed by literature used market data to predict bankruptcy and credit ratings, forecast stocks, bonds, crypto currency and exchange rates, and investigate market inefficiencies. Market inefficiencies in peer-to-peer lending environment has been investigated concluding that overpriced assets were perceived by investors as safer than underpriced ones (Caglayan, Pham, Talavera, & Xiong, 2020). Textual analysis applied to identify qualitative determinants used by Moody's and Fitch for transitioning sovereign credit rating (Slapnik & Lončarski, 2021) as well as to predict economic variables by pacing both English and Polish texts and investigate language differences in prediction results (Rybinski, 2020). Investor comments seems to be important prediction tool regarding peer-to-peer platform collapses within (30) days (Wang C. , Zhang, Zhang, & Gong, 2021). Machine learning is also investigated as assistant on lending decision making however human discretion is still important factor as increases sales (Costello, Down, & Mehta, 2020). On bankruptcy prediction, machine learning algorithms have outperformed statistical ones for default prediction of mortgage-backed securities loans (Cowden, Fabozzi, & Nazemi, 2019). Also such models can be used as a subjectivity tool for sovereign credit rating using macroeconomic variables and predict default within 1-2 years (De Moor, Luitel, Sercu, & Vanpee, 2018) as well as identify country credit risks using CDS and macroeconomic fundamentals out of which Dept/GDP and unemployment ratios seems to have the biggest impact (Arakelian, Dellaportas, Savona, & Vezzoli, 2019). Treasury bonds returns by using macroeconomic and yield-based variables, can be predicted by neural networks (Bianchi, Büchner, & Tamoni, 2020). Literature for exchange rates is also present by predicting GBP/USD rate post-Brexit (Plakandaras, Gupta, & Wohar, 2017) and verification of non-linear relationship between exchange rates flexibility, exchange rates regimes and monetary independence (Ahmed, 2021). One day forecast for Bitcoin price has been achieved using multiple finance and economic variables (Liu, Li, Li, Zhu, & Yao, 2021). Manahov & Urquhart (2021) on the other hand, proved by using crypto market, the higher trading frequency exists in a market, the more efficient it is. Asset allocation named "Hierarchical Risk Parity", has also been applied to crypto currency market as an alternative to traditional risk-based strategies with promising results (Burggraf, 2021). According to Anghel (2021), machine learning might not outperform statistical trading methodologies in crypto market in terms of trending prediction, however they still had significant trading performance. Lastly, Viebig (2019) predicted with the help of machine learning that low market returns over long horizon when market is currently over-valuated.

| Industry Group | Topic | Major Journal Contribution |
|---|---|---|
| Real Estate | Hedonic Prices Prediction | Journal of Real Estate finance and Economics |
| Media & Entertainment | Stock Price Prediction | Journal of Banking and finance |
| Information Technology | Stock Price Prediction (Intraday) | Quantitative finance |
| Oil, Gas & Consumable Fuels | Oil Market Price Prediction | Quantitative finance<br>European Journal of finance<br>Journal of International Money and finance |
| Materials | Bankruptcy | accounting and finance |
| Materials | Gold Price Prediction | finance Research Letters |

| | | |
|---|---|---|
| Consumer Services | Impression Management | accounting Auditing & Accountability Journal |
| Banks | Bankruptcy Prediction | Journal of Banking & finance |
| Banks | Banking Crisis Prediction | Journal of Financial Stability<br>Journal of Financial Services Research |
| Banks | Fintech | Review of Financial Studies<br>Journal of Financial Stability |
| Banks | Fraud Detection | European Journal of finance |
| Banks | Stock Buybacks | Quantitative finance |
| Insurance | Mortality & Risk Predictions | Astin Bulletin |
| Insurance | Misrepresentation | Journal of Risk and Insurance |
| Insurance | Industry Transformation | Journal of Risk and Insurance<br>Insurance Mathematics & Economics |
| Insurance | Insurance Claims Predictions | Insurance Mathematics & Economics Review of accounting Studies |
| Diversified Financials | Lending | Journal of accounting & Economics |
| Diversified Financials | Bankruptcy Prediction | Quantitative finance<br>Journal of Portfolio Management<br>Research in International Business and finance |
| Diversified Financials | Stocks & Bonds Predictions | Review of Quantitative finance & accounting |
| Diversified Financials | Crypto currencies | finance Research Letters |
| Diversified Financials | Exchange Rates Prediction | finance Research Letters<br>Journal of International Money and finance |
| Diversified Financials | Market Inefficiencies & Sovereign Credit Rating | finance Research Letters<br>Journal of Corporate finance<br>Journal of International Financial Markets Institutions |

**Table 8: Topics per Industry Group**

Out of (11) total sectors, scientists used data from (7) as to apply machine learning in accounting and finance. As expected, due to our scope of our research, literature is denser in "Financials" sector than others. From financial and accounting perspective, we find research opportunities on other sectors and industry groups as defined from Global Industry Classification Standard (GICS) (S&P Global Market Intelligence delivers essential information so you can make decisions with conviction, 2021). To illustrate, please refer to Figure 14. Color coding is as follows:

- Red: Literature not found on our dataset
- Yellow: Literature found but with less than 5 papers
- Green: Literature found with more or equal 5 papers

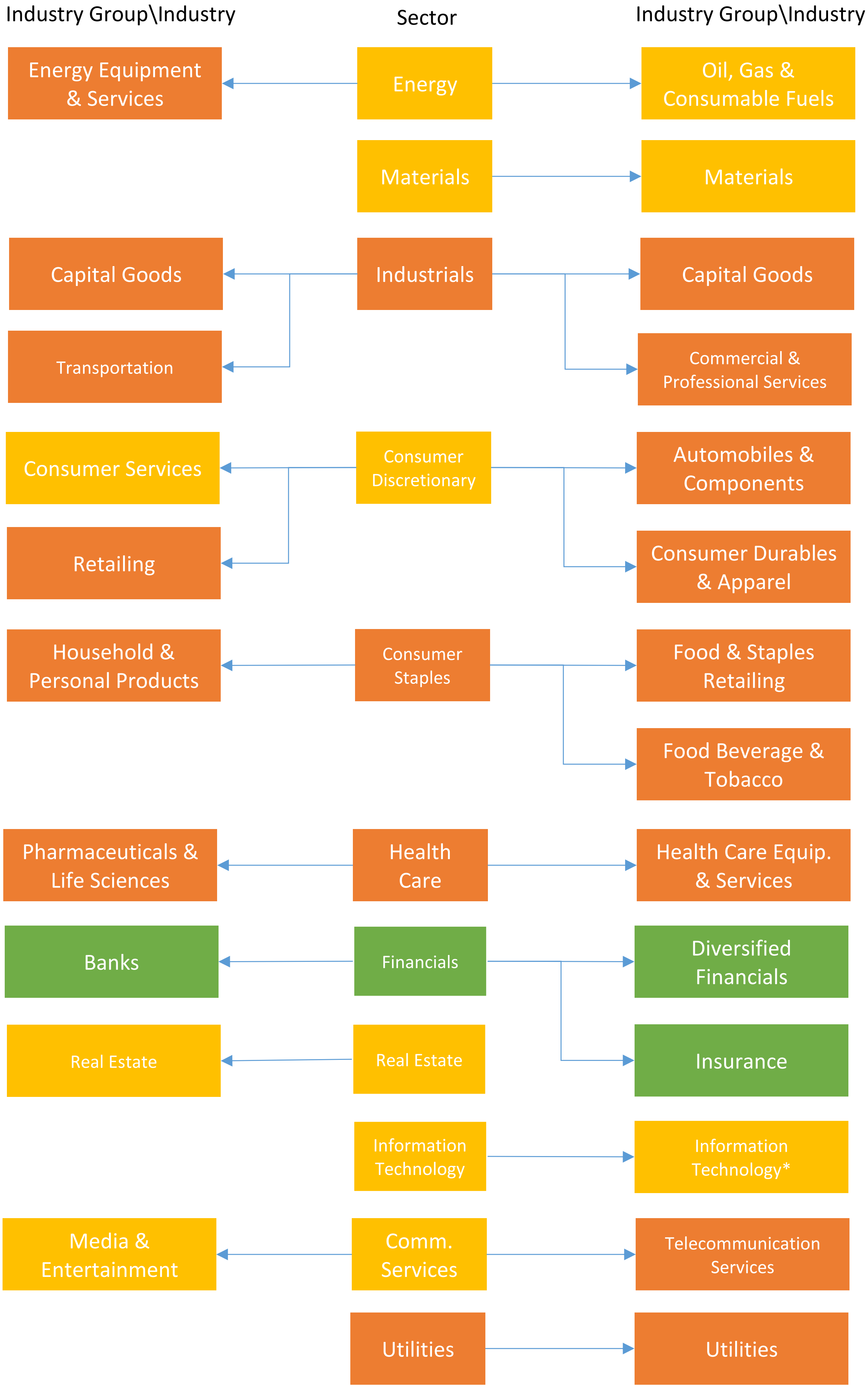
Industry Group\Industry
Sector
Industry Group\Industry
Energy Equipment & Services
Energy
Oil, Gas & Consumable Fuels
Materials
Materials
Capital Goods
Industrials
Capital Goods
Transportation
Commercial & Professional Services
Consumer Services
Consumer Discretionary
Automobiles & Components
Retailing
Consumer Durables & Apparel
Household & Personal Products
Consumer Staples
Food & Staples Retailing
Food Beverage & Tobacco
Pharmaceuticals & Life Sciences
Health Care
Health Care Equip. & Services
Banks
Financials
Diversified Financials
Real Estate
Real Estate
Insurance
Information Technology
Information Technology*
Media & Entertainment
Comm. Services
Telecommunication Services
Utilities
Utilities

**Figure 14: Sectors & Industry Groups**
***Information Technology was not sub-grouped during this research**

### 3.1.2.4 Bibliographic Coupling Results

Except the topics we have identified manually, bibliographic coupling analysis indicated (7) clusters by linking the citing documents based on the number of common sited papers. VOSviewer (VOSviewer - Visualizing scientific landscapes, 2021) constructs distance-based maps therefore the smaller the distance between two items, the stronger their relation (Van Eck & Waltman, 2010). In Figure 15, we present the map constructed by VOSviewer where each node represents a paper and its color indicate the cluster assigned. The node size indicates the times a paper is sited. The produced map enhances our decision of reviewing all papers to identify the characteristics of each clustering; common citations between groups and overlapping districts indicate somewhat strong correlation between some of the papers in different clusters. Full corpus is available in Appendix B.

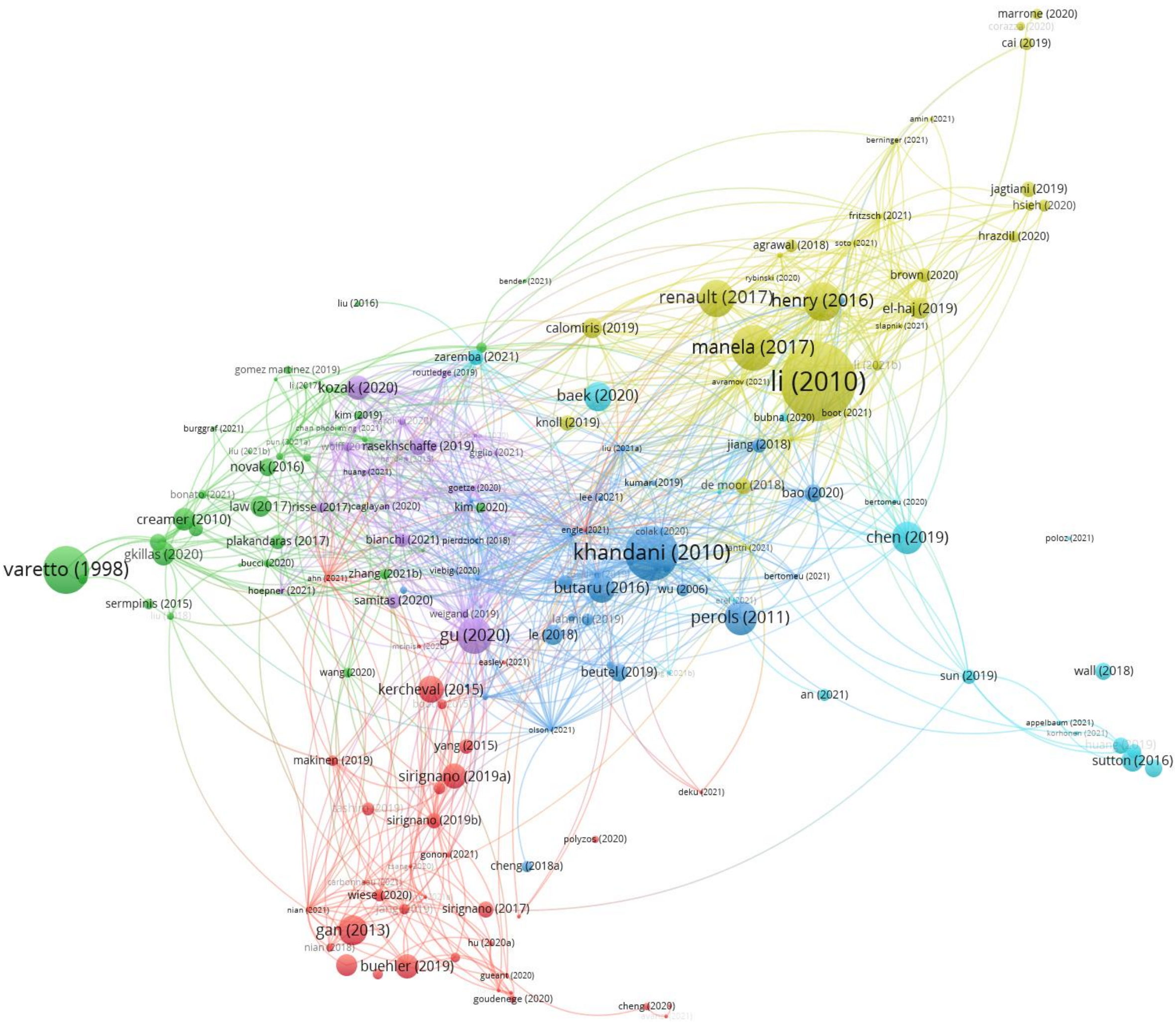


**Figure 15: VOSviewer bibliographic coupling map**

### *3.1.2.4.1 Yellow Cluster – Textual Analysis and Qualitative data*

This cluster is composed of (31) papers which employ textual analysis techniques, image feature extraction, and generally investigates qualitative data usefulness for plethora of cases. Specifically, Hsieh et al. (2020) and Duan et al. (2020) investigate the impact of facial trustworthiness with the first paper confirming that facial characteristics of CTO and CFO impact audit fees and the latter correlation of amount raised from crowdfunding and facial characteristics. Sentiment extraction from social media for stock prediction (Renault, 2017; Knoll, Stübinger, & Grottke, 2018) or corporate social responsibility evaluation from board tweets (Amin, Mohamed, & Elragal, 2021) have been achieved. Literature and corpus analysis using unsupervised models (Cai, Linnenluecke, Marrone, & Singh, 2019; Marrone, Linnenluecke, Richardson, & Smith, 2020; Berninger, Kiesel, Schiereck, & Gaar, 2021) mainly for topic extraction. Financial reports are another source for misreport identification (Brown, Crowley, & Elliott, 2020), crisis prediction (Soto, 2020), prospects of the company based on forward-looking statements (Li F. , 2010), corporate policies identification (Avramov, Li, & Wang, 2021) and analysis of risk management (Friberg & Seiler, 2021). Conference calls transcripts measure credit risk, when combined with financial statements, (Donovan, Jennings, Koharki, & Lee, 2021) or aid culture identification using vector approach such as Word2Vec (Li, Mai, Shen, & Yan, 2020). Word2Vec is under "word embedding" family techniques that aim to capture the semantic and syntactic similarity using vector representations. Supervised models successfully parsed headlines and articles, to measure implied volatility (Manela & Moreira, 2017) and for stock markets returns and risk prediction (Calomiris & Mamaysky, 2019).

The first great challenge researchers are facing upon textual analysis, is the cleanup phase due to unstructured nature of text. Abbreviations, misspelling, punctuations, stop words and combination of uppercase and lowercase letters may impact the accuracy of machine learning model. Abbreviations, misspelling and stop words can be dealt with the creation of word lists that perform replacement actions while stop words are removed from the dataset as irrelevant. In the case of stop words and punctuations, are mitigated by common programming libraries, however abbreviations and misspelling are normally corpus related therefore lexicons need to be found or created. Nowak et al. (2019) concluded small performance from correcting misspellings and abbreviations however the effort spend was not justified. Another crucial issue is the usage of tenses, singular\plural and generally different form of the same word. Two possible directions that researchers can use is the Stemming, chops off the end of the words, or Lemmatization that instead use language specific vocabulary to replace the words.

For sentiment extraction, dictionaries and lexicons are available for classification purposes that remove subjectivity and relative effort for researchers (Loughran & McDonald, 2016). This methodology is an example of "bag of words" which contrary to word embedding, assumes word sequence is not important but number of occurrences. However, word lists are not always available and problem of homographs is present (Loughran & McDonald, 2016). One alternative common solution found in our corpus, is the manual classification of the training and verification sets to fit a supervised machine learning algorithm like naïve bayes (Li F. , 2010; Slapnik & Lončarski, 2021). Surprisingly, researchers seemed to rely on manual classification as (18) out of (31) papers chosen the supervised approach. Unsupervised model LDA for topic extraction is another option which in fact is the most commonly used model in this cluster.

#### *3.1.2.4.2 Cyan Cluster – Automation Adaptation, Job Redesign and Covid-19*

In this cluster of (17) papers, the key topic is the adaptation of technology and the implications on accounting and finance profession. In banking industry, machine learning and AI have already been applied to detect credit card fraud and for chatbot development to support customers (Wall, 2018). Moreover, Blackrock announced the replacement of (40) employees by machines for stock picking with others to follow (Wall, 2018). In accounting, widespread use in audit, tax and consulting is becoming a target goal (Sutton, Holt, & Arnold, 2016) while KPMG and Deloitte confirmed to be customers of AI and machine learning services (Huang & Vasarhelyi, 2019). Neural networks, are able to analyze and process unstructured and semi-structured data from multiple sources (e.g. text data, images, video & audio) therefore by extracting insights they can be useful tool for judgment support in multiple audit phases (Sun, 2019). Even though there is somewhat adaptation, the challenges of selecting the correct processes to automate, reliability issues, replacement of current expert workforce, regulatory restrictions and job redesign emerge.

During 90s, expert systems to automate processes of auditing have been employed, soon abandoned as usage was not mandatory and at some point their maintenance cost couldn't be justified (Gray, Chiu, Liu, & Li, 2014). Also misjudging some task by marking them as able to be automated where in fact this is not the case, is another reason for hindering adaptation (Korhonen, Selos, Laine, & Suomala, 2020). Accounting processes need to be understood and defined before automated (Möller, Schäffer, & Verbeeten, 2020). For example, Korhonen et al. (2020) provided a case study in which they have demonstrated that calculation task performed better by humans than machines even though thought as "routine". Machine learning should not mimic human actions but rather learn how the decision making is done to perform more complex tasks (Huang & Vasarhelyi, 2019). Reliability concerns are also part of the discussion for three reasons. Firstly, machine learning cannot identify edge cases therefore wrong decisions with high impact can be made (Wall, 2018). Secondly, decisions made by algorithms like neural networks, is lucking of transparency and interpretation therefore are avoided due to possible regulation restrictions (Wall, 2018; Kokina & Davenport, 2017). At the same time complexity of audit decision making Issa et al. (2016) hinders the transition.

The appliance of machine learning can impact the accounting and financial labor force not just by replacing them. For example, when a system is more knowledgeable than its user, the decision making may be impacted (Arnold & Sutton, 1998). Moreover, experts might be de-skilled as being reliant of the new system (Arnold & Sutton, 1998). New specialists who can work on data preparation will be hired while accountants need to become more technology inclined to keep up with the new skill demands. Appelbaum et al. (2020) found that data analytics is the most desirable skill for accounting students. Therefore, should be trained in statistics, big data analytics, deep learning, databases and other emerging technologies as to be able to use those tools for auditing (Appelbaum, Showalter, Sun, & Vasarhelyi, 2020).

Regarding Covid-19, we find (3) papers that aim to explain the impact of the pandemic and identify the key features that minimize risk. Xue et al. (2021), as in the previous cluster, performed textual analysis with Word2Vec on annual reports to extract words related to big data. By correlating company stock price with Covid-19 cases, they have identified that companies used big data suffered the less from market fluctuations. On Industry level (Baek, Mohanty, & Glambosky, 2020) have correlated the Covid-19 cases with US stock values by clustering per industry to identify systematic risk and correlation of negative and positive news on stock prices. Zaremba et al. (2021) by investigating (67) markets under multiple factors identified that low unemployment rates, conservative investment policies and undervalued companies tend to be more protected.

#### *3.1.2.4.3 Red Cluster – Options and Limit Order Trading*

We have identified two key topics, first one related to the option pricing by replacing traditional models, and the second one related to high frequency trading and limit order trading. In the first topic, we find European, American and Bermudan option pricing and hedging being investigated due to current traditional method limitations. For European options, Black-Scholes model (Black & Scholes, 1973) and Heston model (Heston, 1993) are among the most commonly used models to identify the underline price. Both models however are parametric based in which estimation assumptions can impact the accuracy. Black-Scholes model for example, assumes constant volatility therefore inaccurate prices have been noticed (Funahashi, 2020; Nian, Coleman, & Li, 2021). On the other hand, slow Monte Carlo methods, another common choice for option pricing, increase accuracy in expense of speed (Horvath, Muguruza, & Tomas, 2020) which is undesired in a high volatility market. In contrary, machine learning models, neural networks in our corpus, are trained directly from market data therefore avoid misspecification that parametric models suffer (Nian, Coleman, & Li, 2021). American options are even more complicated as can be exercised at any time during the contract validity. Regression based Monte Carlo approach is becoming popular option however the objectivity of variable choice on regression-based methods could be compromised especially with high dimensional options (Hu & Zastawniak, 2020). In this sub-group, we find papers that either combining Monte Carlo with machine learning or replacing it. For example, in the work of Goudenège et al. (2020), combination of trees and Monte Carlo for pricing American options found to decrease the computation time without sacrifice accuracy.

In limit order trading, evaluation of the immediate stock pricing movement is based on the asks and bids order flows in a high-frequency trading environment. One approximation in this field, are the statistical models in which assumptions and intensive computations are required therefore are hardly a good fit in existing environments (Bouchaud, Mézard, & Potters, 2002). Moreover, modeling is challenging due to high dimensionality of a limit order book as consists of multiple price levels (Sirignano J. A., 2018). Deep learning is an alternative solution that can generalize and find relations between order flow and market price in a non-parametric way as such to be applied for different stocks (Sirignano & Cont, 2019). Furthermore, neural networks are able to identify whether or not price change is due to successful or cancellation order by pacing the order series (Tashiro, Matsushima, Izumi, & Sakaji, 2019). However, for machine models to be trained, multiple depth\level of the limit orders is needed including both size (number of stocks), price and precise time of placement. Data in high-frequency trading market are abundant therefore training a model is time consuming. For that reason, execution is done on graphic cards instead of CPUs as can parallelize the training process (Sirignano & Cont, 2019).

### *3.1.2.4.4 Green Cluster – Volatility. Portfolio Selection & Trading*

Key topic for papers on this cluster, is volatility as such to be incorporated for portfolio selection and trading to maximize profits for investors. It is not limited though to stock markets, but also to crypto-currencies, commodities and forex. The distinction from previous cluster, is that research is focused on price movements, volatility and investor mood to forecast upcoming trend.

Volatility is important for risk management, pricing and allocation of assets therefore has gained a lot of attention in financial and economic literature. The goal is to minimize the tolerated risk and maximize gains. Traditional models GARCH and Stochastic Volatility, are not suitable anymore in current high frequency data environment (Liu, Pantelous, & von Mettenheim, 2018). Realized volatility can approximate stock volatility however requires nonlinear approach in which neural networks outperform GARCH model (Hu & Tsoukalas, 1999). Another popular model, Markowitz mean-variance portfolio (Markowitz, 1952), ignores transactions costs and is more prone to estimation error than minimum variance models (Clarke, de Silva, & Thorley, 2011). Moreover, rebalancing models and naïve 1/N rule, equally invests on N assets, usually performs better than Markowitz model (Mulvey, Lu, & Sweemer, 2001; DeMiguel, Garlappi, & Uppal, 2007). Thus, creating portfolios of as independent assets as possible and rebalancing only when a certain risk threshold is exceeded, is a promising alternative (Liu, Mulvey, & Zhao, 2015; Li, Chen, Feng, & Ying, 2016). By recognizing low-risk anomaly, where low risk portfolio outperform high risk, Pyo & Lee (2018) experimented with multiple machine learning algorithms and GARCH model to predict volatility and then combined them with Black-Litterman model (Black & Litterman, 1992) for portfolio construction.

In the trading group, crypto-currencies is a good playground to investigate market efficiency on high frequency trading by training machine learning algorithms to simulate investor actions (Manahov & Urquhart, 2021), or predict price in one day timeframe by proving that macroeconomic variables are significant and effective (Liu, Li, Li, Zhu, & Yao, 2021). Prediction on commodities pricing, is another challenging topic due to higher volatility levels in comparison to other markets (Dunis, Laws, Middleton, & Karathanasopoulos, 2013). Once more, GARCH model is found to be one of the most commonly used model in our corpus in which researchers aiming to replace (Sermpinis, Laws, & Dunis, 2013; Gkillas, Gupta, & Pierdzioch, 2020). Investor sentiment also affect prices (Da, Engelberg, & Gao, 2011), therefore its impact is investigated in this group. However, as investor sentiment is not directly measurable, we find one approach of using data from internet searches for predicting gold price (Bonato, Gkillas, Gupta, & Pierdzioch, 2021) or stock price (Zhang, Chu, & Shen, 2021).

Our Corpus indicates that, machine learning is a good fit on portfolio selection and price prediction as used for its non-parametric nature, for neural networks, ability to handle multiple dimensions and process of non-linear data therefore offering advantages over traditional models such as GARCH.

### *3.1.2.4.5 Blue Cluster – Bankruptcy, Delinquency and Fraud Detection*

A key topic in our corpus is related to bankruptcy prediction, recovery after delinquency and misstatement detection as to protect investors, financial instruments and employees. Starting with bankruptcy prediction, scientists are focusing on identifying the correct attributes that can forecast imminent delinquency however we find different methods and variables being selected. For example, Lahmiri & Bekiros (2019) are the only ones in our corpus experimenting with qualitative instead of quantitative data by indicating that statistical model assumptions, such as multivariate normality, are violated. From information asymmetry perspective, Cheng et al. (2018) identified that in U.S market, institutional shareholders change their trading patterns by selling abnormally several quarters before bankruptcy while probably due to SEC enforcement insiders seize trading. In banking sector, it is common for rating agencies to misclassify banks therefore making a necessity for improved predictions (Viswanathan, Srinivasan, & Hariharan, 2020). Le & Viviani (2018) found small accuracy improvements using machine learning over traditional statistical models by measuring (31) different ratios from banking financial statements. In contrary, Beutel et al. (2019) by focusing on just (10) variables, suggested that for banking bankruptcy prediction, statistical models are performing better in out of sample dataset. In this group, we also find distress research in Chinese market which regardless of economic boost, default rates are increasing (Jiang & Jones, 2018). In this paper they have used (90) financial and non-financial variables with annual market returns, GDP growth, GDP per capita, unemployment rates, ROA, ROE to be among the most important ones. Therefore, in this sub-group there is no consensus of the most important variables thus more research is expected.

The recovery phase after delinquency is complex subject as recovery time and amount is not known until the end of recovery process. Cheng and Cirillo (2018), split the recovery into phases and timeframe to deal with the above problem and forecast possible recovery trajectory.

For lending institutions, it is vital to forecast and manage consumer credit risk for their survival as savings could be millions of dollars in annual basis (Butaru, et al., 2016). Currently they rely on human discretion and "hard" information however model and algorithms should be involved on decision making (Khandani, Kim, & Lo, 2010). The models currently in use, are based on past behavior while rapid changes makes a necessity to apply more data driven models. Khandani et al. (2010) combined traditional ratios with banking transactions, were categorized based on expenses, to predict credit score within next three to twelve months resulting from 6% to 23% cost savings. Similar exercise however incorporating data from multiple banks, has been executed to factor as well as the differences in risk management approach by Butaru et al. (2016). They found that cost savings depends on how aggressively the credit lines will be cut with decision trees to save up to 55% over perfect risk management.

Early warning models for financial crisis prediction can be achieved through machine learning. By measuring quarterly (14) variables for EU countries, Holopainen & Sarlin (2017) demonstrated the exceptional predictability performance of KNN, neural networks and ensemble methods over simple statistical models.

Under fraud detection, the focus is to detect misreporting on financial statements as affects investors, employees and produce uncertainty on financial markets. The challenges that researchers had to overcome are the ratio of the non-fraud firms to fraud ones, to find the right attributes which were noisy and the attempt to mask the statements to be as similar as to the ones found in non-fraud firms (Perols, 2011). We found three approaches with Bao et al. (2020) focusing on raw financial data, based on the assumption that fraud is designed to be harder for dedection therefore theories which financial ratios are based on may be misleading, Perols (2011) using combinations of ratios, raw data and dummy variables and Bertomeu et al. (2020) focusing on restatement items including non-accounting variables such as governance, capital markets and auditing. All three papers have raised different sets of variables as most significant for fraud detection. Bertomeu et al. (2020), mentioned that cumulatively, the accounting ones were more important than other groups of variables. There are cases though, were accounting estimations need to be done therefore accuracy of forecasting calculations is of major importance to minimize risk. Ding et al. (2020), experiment with machine learning algorithms in insurance sector to forecast future claims and depict them to financial statements.

#### *3.1.2.4.6 Purple Cluster – Asset Pricing*

Asset pricing is the Holy Grail for investors and financial institutes as accurate estimation of the fair price would allow to find investment opportunities while minimizing the risk. Many models have been proposed since capital asset pricing (CAPM) as research has shown empirical failures (Karolyi & Van Nieuwerburgh, 2020) and growing number of anomalies incorporated on newer models (Geertsema & Lu, 2020). Three-factor model (Fama & French, 1993) replaced by four-factor model (Hou, Xue, & Zhang, 2015) and five-factor model (Fama & French, 2015) to name a few. However, still those models cannot encapsulate the full spectrum of cases as the problem is high-dimensional therefore a larger number of characteristics is needed (Kozak, Nagel, & Santosh, 2020). In search for finding predictors for both cross-section and time series problems, hundreds of estimators have been proposed often highly correlated and investigated under linear prism (Gu, Kelly, & Xiu, 2020; Weigand, 2019). This situation is far from ideal as requires a-priori knowledge of multiple predictors and transform them to be incorporated in regression models.

One solution proposed, is from plethora of predictors to combine them and generate new ones that can reduce dimensionality and correlation problems (Fang, et al., 2020). Statistical tools were used for factor selection but not for construction of new ones. Fang et al. (2020), proposed a machine learning approach that combined multiple neural networks with "prior knowledge" feature to create and select the best features for the prediction model. Dynamic model averaging, has been proposed as to rely on multiple forecasting models instead of one. For financial time-series problems, Dynamic Occam's Window (Onorante & Raftery, 2016) has been introduced which selects a portion of models depending on the period of time. Thus, predictors are not stable depending on economic phase. Geertsema & Lu (2020), through machine learning provided nine factors that scored higher in Sharp ratio than traditional models proposed. They first clustered anomalies and tested (41) factors to identify the ones that can explain all of them.

Forecasting excess bond returns was traditionally based on yield curve, or multiple macroeconomic indicators (Bianchi, Büchner, & Tamoni, 2020). Both cases suffered from the same issues already mentioned above. Due to high-dimensionality and non-linear advantages of machine learning, tree-based algorithms and neural networks improved forecast accuracy while at the same time has been proven that thru neural networks both yield and macroeconomic variables outperform yield-only variables (Bianchi, Büchner, & Tamoni, 2020).

One factor that affects market prices is the non-universal perception of the investors about underline value of assets even if they possess the same set of information. Caglayan et al. (2020), thru machine learning methods estimated the selling success on secondary market for faired priced peer-to-peer loans was much lower than the overpriced ones. The factors that leads to mispricing seems to be transaction costs, delayed search, inattention and capital constraints (Caglayan, Pham, Talavera, & Xiong, 2020).

In asset pricing prediction advanced machine learning algorithms do not always perform better than improved regression models such as ElasticNet and LASSO (Wolff & Neugebauer, 2019), however this is subject to the data at hand.

##### *3.1.2.4.7 Orange Cluster – Image classification and Real Estate*

This clustering is composed by one paper (Johnson, Tidwell, & Villupuram, 2019) which is not sharing common citations with any other in our selected corpus even though we have found papers related to real estate and value estimation. This paper however, is the only one classifying images used for house price prediction alongside with other parameters.

### 3.1.3 (RQ3) What are the future avenues of machine learning in accounting and finance research?

As above research has shown, machine learning has multiple advantages over traditional methods however there is still a debate on how it can be adapted with current strict regulations on accounting and finance and the impact of existing workforce in those fields. We believe that further research will be conducted as to improve the tracking and logging of decision making of the algorithms as well as actions needed to support new or updated regulatory requirements. Otherwise traditional model will not be replaced even though we find promising results in our corpus. Discrimination on decision making, or lack of it, also needs to be part of further examination as will enable the adaptation of machine learning further. We argue that practitioners in accounting and finance should be more technology inclined and be able to work alongside with advanced automation technologies as to improve decision making.

It is evident from our corpus that researchers are choosing their model either by experimenting with multiple ones or based on past experiences. One goal of this paper is to indicate the top performer algorithms per case to guide future researchers however as more advanced algorithms are expected to be invented and enlargement of appliance scope for machine learning, we expect a creation of "best practices" in this field. We strongly propose to include the best algorithms and preferable parameters, confirmation of number of least data needed to produce accurate results, verification measurements and data preparation and selection. Moreover, as environment is far from static, we need to answer for how long a model is still relevant before re-training is required.

Dimension reduction will continue to be a key topic in the near future as we have revealed previous attempts on asset pricing. Inclusion of more market anomalies and convergence of current variables seem to be important for researchers to improve accuracy of their models, reduce execution and preparation times and understand the market factors. We believe that consensus on major factors will be fruitful before experimenting with additional new ones.

In finance, to our surprise we hardly find papers related to Forex market where liquidity is almost unlimited and high frequency trading could be easily applied. For those reasons, we propose research on machine learning for trading in this market by either forecasting currency exchange rate or investigate arbitrage opportunities. Moreover, early warning systems in a volatile economy will be priority in the upcoming years; a combination of bankruptcy prediction with identification of misstatements and price prediction through volatility will be of major importance to protect investors and restore trust to financial instruments. In accounting except in the work of Ding et al. (2020), we fail to find other paper related to accounting forecasts predictions to construct a more accurate financial statement. However, those estimates could be crucial for managerial decision making and for avoiding mistakes that could potentially impact firm's image.

Even though, most papers are conducted to U.S market, we have found increased interest in Asian markets. Therefore, we believe this trend will continue in the near future alongside with increase in a more global approach.

Lastly, we expect new algorithms to be proposed fit for specific tasks instead of broader use that we find today. Also, combinations of existing algorithms are to be tried out to improve their forecast.

# 4. Conclusions

Machine learning has multiple advantages over traditional methods currently in place for accounting and finance such as to extract features, pattern recognition, processing of high-dimensional data and nonlinearity. Our paper shed some light on current research\appliance of machine learning and suggested new paths for further investigation. Through both Literature review and Bibliographic coupling we explore three research questions.

In our first research question (RQ1), we identify a boom in this research field since 2014 which continues up to date. Finance papers makes up of 85% of our corpus while 25% of papers published in Quantitative finance journal. Out of (13) countries which are the main contributors, Italy and USA are the most cited.

In our second research question (RQ2), by categorizing our corpus manually, we identified (10) major research topics in which asset predictions was the most common topic. We have provided the most commonly used machine learning algorithms per topic to provide guidance for future research as multiple models are currently explored to find the best one for specific case. Data structure and approximation of the problem however seems to be the most important factor. Neural networks are a good fit for plethora of cases as long as enough training data are at hand; otherwise out of sample predictions usually achieve lower accuracy than traditional models. Supervised models were the most commonly used while unsupervised were employed mainly for topic extraction through LDA algorithm. This is an indication that the academic focus is not on pattern recognition. We have also identified combination of models being tried out or creation of new ones to improve either accuracy or to allow investigation of other problems in which were not possible previously. Also, we identified increased interest in Asian markets even though still U.S concentrates most of our corpus. Our analysis continued by constructing (7) clusters, some entirely different from our previous analysis, through Bibliographic coupling which enabled us to provide a more holistic analysis of this research topic. We present for each cluster the current challenges that practitioners in finance and accounting are facing as well as the proposed solutions that machine learning evolution is bringing. However, have found a clear gap in academic definition of machine learning as such it can blur the distinction between AI and statistical methods.

Lastly, in our last research question (RQ3) we propose further advancements on current research such as creation of best practices, emphasis on accounting research and investigation of transition in the new automation area for finance and accounting practitioners. As machine learning is suffering from regulations and decision tracking, focus on those aspects will increase the acceptance and transition into this new technology.

Of course, there are some limitations in this paper. Firstly, our corpus is composed exclusively by WoS therefore it is likely other important papers not to be included in our study. Also, in lack of unified definition of machine learning, we have in concluded papers in our corpus based on "general consensus" on which algorithms are included in machine learning or not.

# References

Agrawal, S., Azar, P. D., Lo, A. W., & Singh, T. (2018). Momentum, mean-reversion, and social media: Evidence from Stocktwits and Twitter. *The Journal of Portfolio Management, 44*(7), 85-95. doi:10.3905/jpm.2018.44.7.085

Ahmed, R. (2021). Monetary policy spillovers under intermediate exchange rate regimes. *Journal of International Money and finance, 112*. doi:10.1016/j.jimonfin.2020.102342

Ahn, Y., & Tsai, S.-C. (2021). What factors are associated with stock price jumps in high frequency? *Pacific-Basin finance Journal, 68*. doi:10.1016/j.pacfin.2021.101602

Akbari, A., Ng, L., & Solnik, B. (2021). Drivers of Economic and Financial Integration: A machine learning approach. *Journal of Empirical finance, 61*, 82-102. doi:10.1016/j.jempfin.2020.12.005

Albanese, C., Crépey, S., Hoskinson, R., & Saadeddine, B. (2020). XVA analysis from the balance sheet. *Quantitative finance, 21*(1), 99-123. doi:10.1080/14697688.2020.1817533

Amat, C., Michalski, T., & Stoltz, G. (2018). Fundamentals and exchange rate forecastability with simple machine learning methods. *Journal of International Money and finance, 88*, 1-24. doi:10.1016/j.jimonfin.2018.06.003

Amin, M. H., Mohamed, E. K., & Elragal, A. (2021). CSR disclosure on Twitter: Evidence from the UK. *International Journal of accounting Information Systems, 40*. doi:10.1016/j.accinf.2021.100500

Amini, S., Elmore, R., Öztekin, Ö., & Strauss, J. (2021). Can machines Learn Capital Structure Dynamics? *Journal of Corporate finance, 70*. doi:10.1016/j.jcorpfin.2021.102073

An, J., & Rau, R. (2019). finance, technology and disruption. *The European Journal of finance, 27*(4-5), 334-345. doi:10.1080/1351847x.2019.1703024

Anghel, D.-G. (2021). A reality check on trading rule performance in the cryptocurrency market: machine learning vs. technical analysis. *finance Research Letters, 39*. doi:10.1016/j.frl.2020.101655

Appelbaum, D., Showalter, D. S., Sun, T., & Vasarhelyi, M. A. (2020). A framework for Auditor Data Literacy: A normative position. *accounting Horizons, 35*(2), 5-25. doi:10.2308/horizons-19-127

Arakelian, V., Dellaportas, P., Savona, R., & Vezzoli, M. (2019). Sovereign risk zones in Europe during and after the debt crisis. *Quantitative finance, 19*(6), 961-980. doi:10.1080/14697688.2018.1562197

Arnold, V., & Sutton, S. G. (1998). The theory of technology dominance: Understanding the impact of intelligent decision aids on decision maker's judgments. *dvances in accounting behavioral research, 1*(3), 175-194.

Avanzi, B., Taylor, G., Wang, M., & Wong, B. (2021). Synthetic: An individual insurance claim simulator with feature control. *Insurance: Mathematics and Economics, 100*, 296-308. doi:10.1016/j.insmatheco.2021.06.004

Avramov, D., Li, M., & Wang, H. (2021). Predicting corporate policies using downside risk: A machine learning approach. *Journal of Empirical finance, 63*, 1-26. doi:10.1016/j.jempfin.2021.04.009

Baek, S., Mohanty, S. K., & Glambosky, M. (2020). Covid-19 and stock market volatility: An industry level analysis. *finance Research Letters, 37*. doi:10.1016/j.frl.2020.101748

Bao, Y., Ke, B., Li, b., Yu, Y. J., & Zhang, J. (2020). Detecting accounting fraud in publicly traded U.S. firms using a machine learning approach. *Journal of accounting Research, 58*(1), 199-235. doi:10.1111/1475-679x.12292

Bender, M., & Panz, S. (2021). A general framework for the identification and categorization of risks: An application to the context of Financial Markets. *The Journal of Risk*. doi:10.21314/jor.2021.004

Berninger, M., Kiesel, F., Schiereck, D., & Gaar, E. (2021). Citations and the readers' information-extracting costs of finance articles. *Journal of Banking & finance, 131*. doi:10.1016/j.jbankfin.2021.106188

Bertomeu, J. (2020). machine learning Improves accounting: Discussion, implementation and research opportunities. *Review of accounting Studies, 25*(3), 1135-1155. doi:10.1007/s11142-020-09554-9

Bertomeu, J., Cheynel, E., Floyd, E., & Pan, W. (2020). Using machine learning to detect misstatements. *Review of accounting Studies, 26*(2), 468-519. doi:10.1007/s11142-020-09563-8

Beutel, J., List, S., & von Schweinitz, G. (2019). Does machine learning help us predict banking crises? *Journal of Financial Stability, 45*. doi:10.1016/j.jfs.2019.100693

Bianchi, D., Büchner, M., & Tamoni, A. (2020). Bond risk premiums with machine learning. *The Review of Financial Studies, 34*(2), 1046-1089. doi:10.1093/rfs/hhaa062

Bisikalo, O., Vysotska, V., Lytvyn, V., Brodyak, O., Vyshemyrska, S., & Rozov, Y. (2020). Experimental investigation of significant keywords search in Ukrainian content. *Advances in Intelligent Systems and Computing*, 3-29. doi:10.1007/978-3-030-63270-0_1

Black, F., & Litterman, R. (1992). Global portfolio optimization. *Financial Analysts Journal, 48*(5), 28-43. doi:10.2469/faj.v48.n5.28

Black, F., & Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy, 81*(3), 637-654. doi:10.1086/260062

Blei, D. M., Ng, A. Y., & Jordan, M. (2003). Latent Dirichlet Allocation. *Journal of machine learning Research, 3*(4-5), 993-1022. doi:10.1162/jmlr.2003.3.4-5.993

Bonato, M., Gkillas, K., Gupta, R., & Pierdzioch, C. (2021). A note on investor happiness and the predictability of realized volatility of gold. *finance Research Letters, 39*, 101614. doi:10.1016/j.frl.2020.101614

Boot, A., Hoffmann, P., Laeven, L., & Ratnovski, L. (2021). Fintech: What's old, what's new? *Journal of Financial Stability, 53*. doi:10.1016/j.jfs.2020.100836

Booth, A., Gerding, E., & McGroarty, F. (2014). Performance-weighted ensembles of random forests for predicting price impact. *Quantitative finance, 15*(11), 1823-1835. doi:10.1080/14697688.2014.983539

Bottazzi, G., & Giachini, D. (2019). Far from the madding crowd: Collective wisdom in prediction markets. *Quantitative finance, 19*(9), 1461-1471. doi:10.1080/14697688.2019.1622285

Bouchaud, J.-P., Mézard, M., & Potters, M. (2002). Statistical properties of stock order books: Empirical results and Models. *Quantitative finance, 2*(4), 251-256. doi:10.1088/1469-7688/2/4/301

Boyack, K., & Klavans, R. (2010). Co-Citation Analysis, Bibliographic Coupling, and Direct Citation: Which Citation Approach Represents the Reaserch Front Most Accurately? *Journal of the American Society for Information Science and Technology, 61*(12), 2389-2404. doi:10.1002/asi.21419

Breiman, L. (2001). Random Forest. *machine learning, 45*, 5-32. doi:10.1023/A:1010933404324

Brown, N. C., Crowley, R. M., & Elliott, B. W. (2020). What are you saying? Using topic to detect financial misreporting. *Journal of accounting Research, 58*(1), 237-291. doi:10.1111/1475-679x.12294

Brown, P., Gallery, G., & Goei, O. (2006). Does market misvaluation help explain share market long-run underperformance following a seasoned equity issue? *accounting and finance, 46*(2), 191-219. doi:10.1111/j.1467-629X.2006.00172.x

Bubna, A., Das, S. R., & Prabhala, N. (2019). Venture Capital Communities. *Journal of Financial and Quantitative Analysis, 55*(2), 621-651. doi:10.1017/s002210901900005x

Bucci, A. (2020). Realized volatility forecasting with Neural Networks. *Journal of Financial Econometrics, 18*(3), 502-531. doi:10.1093/jjfinec/nbaa008

Buehler, H., Gonon, L., Teichmann, J., & Wood, B. (2019). Deep hedging. *Quantitative finance, 19*(8), 1271-1291. doi:10.1080/14697688.2019.1571683

Burggraf, T. (2021). Beyond risk parity – A machine learning-based hierarchical risk parity approach on cryptocurrencies. *finance Research Letters, 38*. doi:10.1016/j.frl.2020.101523

Butaru, F., Chen, Q., Clark, B., Das, S., Lo, A. W., & Siddique, A. (2016). Risk and risk management in the credit card industry. *Journal of Banking & finance, 72*, 218-239. doi:10.1016/j.jbankfin.2016.07.015

Caglayan, M., Pham, T., Talavera, O., & Xiong, X. (2020). Asset mispricing in peer-to-peer loan secondary markets. *Journal of Corporate finance, 65*. doi:10.1016/j.jcorpfin.2020.101769

Cai, C. W., Linnenluecke, M. K., Marrone, M., & Singh, A. K. (2019). machine learning and expert judgement: Analyzing emerging topics in accounting and finance research in the Asia–pacific. *Abacus, 55*(4), 709-733. doi:10.1111/abac.12179

Calomiris, C. W., & Mamaysky, H. (2019). How news and its context drive risk and returns around the world. *Journal of Financial Economics, 133*(2), 299-336. doi:10.1016/j.jfineco.2018.11.009

Carbonneau, A., & Godin, F. (2020). Equal risk pricing of derivatives with deep hedging. *Quantitative finance, 21*(4), 593-608. doi:10.1080/14697688.2020.1806343

*Chartered Assosiation of Business Schools*. (2021, October 28). Retrieved from Academic Journal Guide 2021: https://charteredabs.org/academic-journal-guide-2021/

Chen, M. A., Wu, Q., & Yang, B. (2019). How valuable is fintech innovation? *Review of Financial Studies, 32*(5), 2062-2106. doi:10.1093/rfs/hhy130

Cheng, C., Jones, S., & Moser, W. J. (2018). Abnormal trading behavior of specific types of shareholders before US firm bankruptcy and its implications for firm bankruptcy prediction. *Journal of Business finance & accounting, 45*(9-10), 1100-1138. doi:10.1111/jbfa.12338

Cheng, C.-H., & Wang, S.-H. (2015). A quarterly time-series classifier based on a reduced-dimension generated rules method for identifying financial distress. *Quantitative finance, 15*(12), 1979-1994. doi:10.1080/14697688.2015.1008029

Cheng, D., & Cirillo, P. (2018). A reinforced urn process modeling of recovery rates and Recovery Times. *Journal of Banking & finance, 96*, 1-17. doi:10.1016/j.jbankfin.2018.08.014

Cheng, X., Jin, Z., & Yang, H. (2020). Optimal Insurance Strategies: A hybrid deep learning markov chain approximation approach. *ASTIN Bulletin, 50*(2), 449-477. doi:10.1017/asb.2020.9

Ciampi, F., Giannozzi, A., Marzi, G., & Altman, E. I. (2021). Rethinking SME default prediction: a systematic literature review and future perspectives. *Scientometrics, 126*, 2141-2188. doi:10.1007/s11192-020-03856-0

Clarke, R., de Silva, H., & Thorley, S. (2011). Minimum-variance portfolio composition. *The Journal of Portfolio Management, 37*, 31-45. doi:10.3905/jpm.2010.2010.1.009

Colak, G., Fu, M., & Hasan, I. (2020). Why are some Chinese firms failing in the US Capital Markets? A machine learning approach. *Pacific-Basin finance Journal, 61*. doi:10.1016/j.pacfin.2020.101331

Collins, M., Schapire, R. E., & Singer, Y. (2002). Logistic Regression, AdaBoost and Bregman Distances. *machine learning, 48*(1-3), 253-285. doi:10.1023/A:1013912006537

Cont, R., & Kukanov, A. (2016). Optimal order placement in limit order markets. *Quantitative finance, 17*(1), 21-39. doi:10.1080/14697688.2016.1190030

Corazza, L., Truant, E., Scagnelli, S., & Mio, C. (2020). Sustainability reporting after the Costa Concordia Disaster: A multi-theory study on legitimacy, Impression Management and image restoration. *accounting, Auditing & Accountability Journal, 33*(8), 1909-1941. doi:10.1108/aaaj-05-2018-3488

Costello, A. M., Down, A. K., & Mehta, M. N. (2020). machine plus man: A field experiment on the role of discretion in augmenting AI-based lending models. *Journal of accounting & Economics, 70*(2-3). doi:10.1016/j.jacceco.2020.101360

Cowden, C., Fabozzi, F. J., & Nazemi, A. (2019). Default prediction of commercial real estate properties using machine learning Techniques. *The Journal of Portfolio Management, 45*(7), 55-67. doi:10.3905/jpm.2019.1.104

Creamer, G. (2012). Model calibration and Automated Trading Agent for euro futures. *Quantitative finance, 12*(4), 531-545. doi:10.1080/14697688.2012.664921

Creamer, G. G. (2015). Can a corporate network and news sentiment improve portfolio optimization using the black–litterman model? *Quantitative finance, 15*(2), 1405-1416. doi:10.1080/14697688.2015.1039865

Creamer, G. G., & Lee, C. (2019). A multivariate distance nonlinear causality test based on partial distance correlation: A machine learning application to energy futures. *Quantitative finance, 19*(9), 1531-1542. doi:10.1080/14697688.2019.1622300

D'Hondt, C., De Winne, R., Ghysels, E., & Raymond, S. (2020). Artificial Intelligence Alter Egos: Who might benefit from robo-investing? *Journal of Empirical finance, 59*, 278-299. doi:10.1016/j.jempfin.2020.10.002

Da, Z., Engelberg, J., & Gao, P. (2011). In search of attention. *The Journal of finance, 66*(5), 1461-1499. doi:10.1111/j.1540-6261.2011.01679.x

De Moor, L., Luitel, P., Sercu, P., & Vanpee, R. (2018). Subjectivity in sovereign credit ratings. *JOURNAL OF BANKING & FINANCE, 88*, 366-392. doi:10.1016/j.jbankfin.2017.12.014

De Spiegeleer, J., Madan, D. B., Reyners, S., & Schoutens, W. (2018). machine learning for Quantitative finance: Fast derivative pricing, hedging and fitting. *18*(10), 1635-1643. doi:10.1080/14697688.2018.1495335

Deku, S. Y., Kara, A., & Semeyutin, A. (2020). The predictive strength of MBS yield spreads during asset bubbles. *Review of Quantitative finance and accounting, 56*(1), 111-142. doi:10.1007/s11156-020-00888-8

DeMiguel, V., Garlappi, L., & Uppal, R. (2007). Optimal versus naive diversification: How inefficient is the 1/N portfolio strategy? *Review of Financial Studies, 22*(5), 1915-1953. doi:10.1093/rfs/hhm075

Devriendt, S., Antonio, K., Reynkens, T., & Verbelen, R. (2021). Sparse regression with multi-type regularized feature modeling. *Insurance: Mathematics and Economics, 96*, 248-261. doi:10.1016/j.insmatheco.2020.11.010

Ding, K., Lev, B., Peng, X., Sun, T., & Vasarhelyi, M. A. (2020). machine learning improves accounting estimates: Evidence from insurance payments. *Review of accounting Studies, 25*(3), 1098-1134. doi:10.1007/s11142-020-09546-9

Donovan, J., Jennings, J., Koharki, K., & Lee, J. (2021). Measuring credit risk using qualitative disclosure. *Review of accounting Studies, 26*(2), 815-863. doi:10.1007/s11142-020-09575-4

Duan, Y., Hsieh, T.-S., Wang, R. R., & Wang, Z. (2020). Entrepreneurs' facial trustworthiness, gender, and crowdfunding success. *Journal of Corporate finance, 64*. doi:10.1016/j.jcorpfin.2020.101693

Dumitriu, D., & Popescu, M. (2020). Artificial Intelligence Solutions for Digital Marketing. *Procedia Manufacturing, 46*, 630-636. doi:10.1016/j.promfg.2020.03.090

Dunis, C. L., Laws, J., Middleton, P. W., & Karathanasopoulos, A. (2013). Trading and hedging the corn/ethanol crush spread using time-varying leverage and nonlinear models. *The European Journal of finance, 21*(4), 352-375. doi:10.1080/1351847x.2013.830140

Easley, D., de Prado, M. L., O'Hara, M., & Zhang, Z. (2021). Microstructure in the machine Age. *Review of Financial Studies*, 3316-3363. doi:10.1093/rfs/hhaa078

El-Haj, M., Rayson, P., Walker, M., Young, S., & Simaki, V. (2019). In search of meaning: Lessons, resources and next steps for computational analysis of Financial Discourse. *Journal of Business finance & accounting, 46*(3-4), 265-306. doi:10.1111/jbfa.12378

Engle, R. F., Hansen, M. K., Karagozoglu, A. K., & Lunde, A. (2020). News and idiosyncratic volatility: The Public Information Processing Hypothesis*. *Journal of Financial Econometrics, 19*(1), 1-38. doi:10.1093/jjfinec/nbaa038

Erel, I., Stern, L. H., Tan, C., & Weisbach, M. S. (2021). Selecting directors using machine learning. *The Review of Financial Studies, 34*(7), 3226-3264. doi:10.1093/rfs/hhab050

Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and Bonds. *Journal of Financial Economics, 33*(1), 3-56. doi:10.1016/0304-405x(93)90023-5

Fama, E. F., & French, K. R. (2015). A five-factor asset pricing model. *Journal of Financial Economics, 116*(1), 1-22. doi:10.1016/j.jfineco.2014.10.010

Fang, J., Lin, J., Xia, S., Xia, Z., Hu, S., Liu, X., & Jiang, Y. (2020). Neural network-based automatic factor construction. *Quantitative finance, 20*(12), 2101-2114. doi:10.1080/14697688.2020.1814039

Félix, L., Kräussl, R., & Stork, P. (2020). Implied volatility sentiment: A tale of two tails. *Quantitative finance, 20*(5), 823-849. doi:10.1080/14697688.2019.1696018

Fischer, J. A., Pohl, P., & Ratz, D. (2020). A machine learning approach to Univariate Time Series forecasting of quarterly earnings. *Review of Quantitative finance and accounting, 55*(4), 1163-1179. doi:10.1007/s11156-020-00871-3

Foley, S., Kwan, A., McInish, T. H., & Philip, R. (2016). Director discretion and insider trading profitability. *Pacific-Basin finance Journal, 39*, 28-43. doi:10.1016/j.pacfin.2016.05.005

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences, 55*(1), 119-139. doi:10.1006/jcss.1997.1504

Friberg, R., & Seiler, T. (2021). Different ways of managing risk as reported in 10-KS: A supervised learning approach. *Financial Review, 56*(4), 773-792. doi:10.1111/fire.12268

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting. *The Annals of Statistics, 28*(2), 337-407. doi:10.1214/aos/1016218223

Fritzsch, S., Scharner, P., & Weiß, G. (2021). Estimating the relation between digitalization and the market value of insurers. *Journal of Risk and Insurance, 88*(3), 529-567. doi:10.1111/jori.12346

Funahashi, H. (2020). Artificial neural network for option pricing with and without asymptotic correction. *Quantitative finance, 21*(4), 575-592. doi:10.1080/14697688.2020.1812702

Gan, G. (2013). Application of data clustering and machine learning in Variable annuity valuation. *Insurance: Mathematics and Economics, 53*(3), 795-801. doi:10.1016/j.insmatheco.2013.09.021

Geertsema, P., & Lu, H. (2020). The correlation structure of Anomaly Strategies. *Journal of Banking & finance, 119*. doi:10.1016/j.jbankfin.2020.105934

Giglio, S., Liao, Y., & Xiu, D. (2020). Thousands of alpha tests. *The Review of Financial Studies, 34*(7), 3456-3496. doi:10.1093/rfs/hhaa111

Gkillas, K., Gupta, R., & Pierdzioch, C. (2020). Forecasting realized oil-price volatility: The role of Financial Stress and asymmetric loss. *Journal of International Money and finance, 104*, 102-137. doi:10.1016/j.jimonfin.2020.102137

Gómez Martínez, R., Prado Román, M., & Plaza Casado, P. (2018). Big Data Algorithmic Trading Systems based on investors' mood. *Journal of Behavioral finance, 20*(2), 227-238. doi:10.1080/15427560.2018.1506786

Gonon, L., Muhle-Karbe, J., & Shi, X. (2021). Asset pricing with general transaction costs: Theory and numerics. *Mathematical finance, 31*(2), 595-648. doi:10.1111/mafi.12297

Götze, T., Gürtler, M., & Witowski, E. (2020). Improving cat bond pricing models via machine learning. *Journal of Asset Management, 21*(5), 428-446. doi:10.1057/s41260-020-00167-0

Goudenège, L., Molent, A., & Zanette, A. (2020). machine learning for Pricing American options in high-dimensional Markovian and non-Markovian models. *Quantitative finance, 20*(4), 573-591. doi:10.1080/14697688.2019.1701698

Gray, G. L., Chiu, V., Liu, Q., & Li, P. (2014). The Expert Systems Life Cycle in AIS research: What does it mean for future AIS research? *International Journal of accounting Information Systems, 15*(4), 423-451. doi:10.1016/j.accinf.2014.06.001

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies, 33*(5), 2223-2273. doi:10.1093/rfs/hhaa009

Guéant, O., Manziuk, I., & Pu, J. (2020). Accelerated share repurchase and other buyback programs: What neural networks can bring. *Quantitative finance, 20*(8), 1389-1404. doi:10.1080/14697688.2020.1729397

Harvey, C., Morris, H., Kelly, A., & Rowlinson, M. (2010). Academic Journal Quality Guide.

Harzing, A. W., & Alakangas, S. (2016). Google Scholar, Scopus and the Web of Science: A longitudinal and cross-disciplinary comparison. *Scientometrics, 106*, 787-804. doi:10.1007/s11192-015-1798-9

Hayden, L. X., Chachra, R., Alemi, A. A., Ginsparg, P. H., & Sethna, J. P. (2018). Canonical sectors and evolution of firms in the US Stock Markets. *Quantitative finance, 18*(10), 1619-1634. doi:10.1080/14697688.2018.1444278

Heston, S. L. (1993). A closed-form solution for options with stochastic volatility with applications to bond and currency options. *Review of Financial Studies, 6*(2), 327-343. doi:10.1093/rfs/6.2.327

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735-1780. doi:10.1162/neco.1997.9.8.1735

Hoepner, A. G., McMillan, D., Vivian, A., & Wese Simen, C. (2020). Significance, relevance and explainability in the machine learning age: An econometrics and Financial Data Science Perspective. *The European Journal of finance, 27*(1-2), 1-7. doi:10.1080/1351847X.2020.1847725

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*(1), 55-67. doi:10.1080/00401706.1970.10488634

Holland, J. H. (1975). *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence.* University of Michigan Press.

Holopainen, M., & Sarlin, P. (2017). Toward robust early-warning models: A horse race, ensembles and model uncertainty. *Quantitative finance, 17*(12), 1933-1963. doi:10.1080/14697688.2017.1357972

Horvath, B., Muguruza, A., & Tomas, M. (2020). Deep learning volatility: A deep neural network perspective on pricing and calibration in (rough) volatility models. *Quantitative finance, 21*(1), 11-27. doi:10.1080/14697688.2020.1817974

Hou, K., Xue, C., & Zhang, L. (2015). Digesting anomalies: An investment approach. *Review of Financial Studies, 28*(3), 650-705. doi:10.1093/rfs/hhu068

Hrazdil, K., Novak, J., Rogo, R., Wiedman, C., & Zhang, R. (2019). Measuring executive personality using machine-learning algorithms: A new approach and audit fee-based Validation Tests. *Journal of Business finance & accounting, 47*(3-4), 519-544. doi:10.1111/jbfa.12406

Hsieh, T.-S., Kim, J.-B., Wang, R. R., & Wang, Z. (2020). Seeing is believing? executives' facial trustworthiness, auditor tenure, and audit fees. *Journal of accounting and Economics, 69*(1). doi:10.1016/j.jacceco.2019.101260

Hu, M. Y., & Tsoukalas, C. (1999). Combining conditional volatility forecasts using neural networks: An application to the EMS exchange rates. *Journal of International Financial Markets, Institutions and Money, 9*(4), 407-422. doi:10.1016/s1042-4431(99)00015-3

Hu, R. (2020). Deep learning for ranking response surfaces with applications to optimal stopping problems. *Quantitative finance, 20*(9), 1567-1581. doi:10.1080/14697688.2020.1741669

Hu, W., & Zastawniak, T. (2020). Pricing high-dimensional American options by kernel ridge regression. *Quantitative finance, 20*(5), 851-865. doi:10.1080/14697688.2020.1713393

Huang, D., Li, J., & Wang, L. (2021). Are disagreements agreeable? evidence from information aggregation. *Journal of Financial Economics, 141*(1), 83-101. doi:10.1016/j.jfineco.2021.02.006

Huang, F., & Vasarhelyi, M. A. (2019). Applying robotic process automation (RPA) in Auditing: A Framework. *International Journal of accounting Information Systems, 35*. doi:10.1016/j.accinf.2019.100433

Issa, H., Sun, T., & Vasarhelyi, M. A. (2016). Research ideas for artificial intelligence in auditing: The formalization of audit and workforce supplementation. *Journal of Emerging Technologies in accounting, 13*(2), 1-20. doi:10.2308/jeta-10511

Jagtiani, J., & Lemieux, C. (2019). The roles of Alternative Data and machine learning in fintech lending: Evidence from the lendingclub consumer platform. *Financial Management, 48*, 1009-1029. doi:https://doi.org/10.1111/fima.12295

Jang, H., & Lee, J. (n.d.). Generative Bayesian neural network model for risk-neutral pricing of American index options. *Quantitative finance, 19*(4), 587-603. doi:10.1080/14697688.2018.1490807

Jiang, Y., & Jones, S. (2018). Corporate distress prediction in China: A machine learning approach. *accounting & finance, 58*(4), 1063-1109. doi:10.1111/acfi.12432

Johnson, E. B., Tidwell, A., & Villupuram, S. V. (2019). Valuing Curb Appeal. *The Journal of Real Estate finance and Economics, 60*(1-2), 111-133. doi:10.1007/s11146-019-09713-z

Jones, S., & Alam, N. (2019). A machine learning analysis of citation impact among selected Pacific basin journals. *accounting & finance, 59*(4), 2509-2552. doi:10.1111/acfi.12584

Kakushadze, Z., & Yu, W. (2017). Decoding stock market with Quant Alphas. *Journal of Asset Management, 19*(1), 38-48. doi:10.1057/s41260-017-0059-2

Kaminski, M., Loniewski, I., & Marlicz, W. (2020). 'Dr. Google, i have a stomach ache' - seasonal variations in abnormal pain: a 4-year retrospective data analysis from Google Ads keywords planner. *Therapeutic Advances in Gastroenterology, 13*, 1-10. doi:10.1177/1756284820931744

Karolyi, G. A., & Van Nieuwerburgh, S. (2020). New methods for the cross-section of returns. *The Review of Financial Studies, 33*(5), 1879-1890. doi:10.1093/rfs/hhaa019

Kercheval, A. N., & Zhang, Y. (2015). Modelling high-frequency limit order book dynamics with support Vector machines. *Quantitative finance, 15*(8), 1315-1329. doi:10.1080/14697688.2015.1032546

Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & finance, 34*(11), 2767-2787. doi:10.1016/j.jbankfin.2010.06.001

Kim, S. (2019). Enhancing the momentum strategy through deep regression. *Quantitative finance, 19*(7), 1121-1133. doi:10.1080/14697688.2018.1563707

Kim, S., & Kim, S. (2019). Index tracking through Deep Latent Representation learning. *Quantitative finance, 20*(4), 639-652. doi:10.1080/14697688.2019.1683599

Knoll, J., Stübinger, J., & Grottke, M. (2018). Exploiting social media with higher-order factorization machines: Statistical arbitrage on high-frequency data of the S&P 500. *Quantitative finance, 19*(4), 571-585. doi:10.1080/14697688.2018.1521002

Kokina, J., & Davenport, T. H. (2017). The emergence of Artificial Intelligence: How Automation is Changing Auditing. *Journal of Emerging Technologies in accounting, 14*(1), 115-122. doi:https://doi.org/10.2308/jeta-51730

Konstantinov, G., Chorus, A., & Rebmann, J. (2020). A network and machine learning approach to factor, asset, and blended allocation. *The Journal of Portfolio Management, 46*(6), 54-71. doi:10.3905/jpm.2020.1.147

Korhonen, T., Selos, E., Laine, T., & Suomala, P. (2020). Exploring the programmability of management accounting work for increasing automation: An interventionist case study. *accounting, Auditing & Accountability Journal, 34*(2), 253-280. doi:10.1108/AAAJ-12-2016-2809

Kozak, S., Nagel, S., & Santosh, S. (2020). Shrinking the cross-section. *Journal of Financial Economics, 135*(2), 271-292. doi:10.1016/j.jfineco.2019.06.008

Kumar, G., Muckley, C. B., Pham, L., & Ryan, D. (2019). Can alert models for fraud protect the elderly clients of a financial institution? *SSRN Electronic Journal, 25*(17), 1683-1707. doi:10.1080/1351847X.2018.1552603

Lahmiri, S., & Bekiros, S. (2019). Can machine learning approaches predict corporate bankruptcy? evidence from a qualitative experimental design. *Quantitative finance, 19*(9), 1569-1577. doi:10.1080/14697688.2019.1588468

Law, T., & Shawe-Taylor, J. (n.d.). Practical bayesian support vector regression for financial time series prediction and Market Condition Change Detection. *Quantitative finance, 17*(9), 1403-1416. doi:10.1080/14697688.2016.1267868

Le, H. H., & Viviani, J.-L. (2018). Predicting bank failure: An improvement by implementing a machine-learning approach to classical financial ratios. *Research in International Business and finance, 44*, 16-25. doi:10.1016/j.ribaf.2017.07.104

Lee, S. C. (2020). Addressing imbalanced insurance data through zero-inflated Poisson regression with boosting. *ASTIN Bulletin, 51*(1), 27-55. doi:10.1017/asb.2020.40

Levine-Clark, M., & Gil, E. (2021). *Journal of Business & finance Librarianship, 26*, 145-163. doi:10.1080/08963568.2021.1916724

Li, F. (2010). The Information Content of Forward-Looking Statements in Corporate Filings-A Naive Bayesian machine learning Approach. *Journal of accounting Research, 48*(5), 1049-1102. doi:10.1111/j.1475-679X.2010.00382.x

Li, H., Song, Q., & Su, J. (2021). Robust estimates of insurance misrepresentation through kernel quantile regression mixtures. *Journal of Risk and Insurance, 88*(3), 625-663. doi:10.1111/jori.12358

Li, K., Mai, F., Shen, R., & Yan, X. (2020). Measuring corporate culture using machine learning. *The Review of Financial Studies, 34*(7), 3265-3315. doi:10.1093/rfs/hhaa079

Li, T., Chen, K., Feng, Y., & Ying, Z. (2016). Binary switch portfolio. *Quantitative finance, 17*(5), 763-780. doi:10.1080/14697688.2016.1223337

Li, W., & Mei, F. (2020). Asset returns in deep learning methods: An empirical analysis on SSE 50 and CSI 300. *Research in International Business and finance, 54*. doi:10.1016/j.ribaf.2020.101291

Liang, J., Xu, Z., & Li, P. (2021). Deep learning-based least squares forward-backward stochastic differential equation solver for high-dimensional derivative pricing. *Quantitative finance, 21*(8). doi:10.1080/14697688.2021.1881149

Liu, F., Pantelous, A. A., & von Mettenheim, H.-J. (2018). Forecasting and trading high frequency volatility on large indices. *Quantitative finance, 18*(5), 737-748. doi:10.1080/14697688.2017.1414489

Liu, H., Mulvey, J., & Zhao, T. (2015). A semiparametric graphical modelling approach for large-scale equity selection. *Quantitative finance, 16*(7), 1053-1067. doi:10.1080/14697688.2015.1101149

Liu, M., Li, G., Li, J., Zhu, X., & Yao, Y. (2021). Forecasting the price of bitcoin using Deep learning. *finance Research Letters, 40*. doi:10.1016/j.frl.2020.101755

Liu, Q., Wang, C., Zhang, P., & Zheng, K. (2021). Detecting stock market manipulation via machine learning: Evidence from China Securities Regulatory Commission punishment cases. *International Review of Financial Analysis, 78*. doi:10.1016/j.irfa.2021.101887

Loisel, S., Piette, P., & Tsai, C.-H. J. (2021). Applying economic measures to lapse risk management with machine learning approaches. *ASTIN Bulletin, 51*(3), 839-871. doi:10.1017/asb.2021.10

Loughran, T., & McDonald, B. (2016). Textual analysis in accounting and finance: A survey. *Journal of accounting Research, 54*(4), 1187-1230. doi:10.1111/1475-679x.12123

M'ng, J., & Jer, H. (2021). Do economic statistics contain information to predict stock indexes futures prices and returns? evidence from Asian Equity Futures Markets. *Review of Quantitative finance and accounting, 57*(3), 1033-1060. doi:10.1007/s11156-021-00969-2

Madani, F. (2015). 'Technology Mining' bibliometrics analysis: applying network analysis and cluster analysis. *Scientometrics, 105*(1), 323-335. doi:10.1007/s11192-015-1685-4

Mäkinen, Y., Kanniainen, J., Gabbouj, M., & Iosifidis, A. (2019). Forecasting jump arrivals in stock prices: New attention-based network architecture using Limit Order Book Data. *Quantitative finance, 19*(12), 2033-2050. doi:10.1080/14697688.2019.1634277

Manahov, V., & Urquhart, A. (2021). The efficiency of bitcoin: A strongly typed genetic programming approach to Smart Electronic Bitcoin markets. *International Review of Financial Analysis, 73*. doi:10.1016/j.irfa.2020.101629

Manela, A., & Moreira, A. (2017). News implied volatility and disaster concerns. *Journal of Financial Economics, 123*(1), 137-162. doi:10.1016/j.jfineco.2016.01.032

Markowitz, H. (1952). Portfolio selection. *The Journal of finance, 7*(1), 77-91. doi:10.2307/2975974

Marrone, M., Linnenluecke, M. K., Richardson, G., & Smith, T. (2020). Trends in environmental accounting research within and outside of the accounting discipline. *accounting, Auditing & Accountability Journal, 33*(8), 2167-2193. doi:10.1108/aaaj-03-2020-4457

McCulloch, W. S., & Pitts, W. (1943). A Logical Calculus of the Ideas Immanent in Nervous Activity. *Bulletin of Mathematical Biophysics, 5*(4), 115-133. doi:10.1007/bf02478259

McInish, T. H., Nikolsko-Rzhevska, O., Nikolsko-Rzhevskyy, A., & Panovska, I. (2019). Fast and slow cancellations and trader behavior. *Financial Management, 49*(4), 973-996. doi:10.1111/fima.12298

Möller, K., Schäffer, U., & Verbeeten, F. (2020). Digitalization in Management accounting and control: An Editorial. *Journal of Management Control, 31*(1-2), 1-8. doi:10.1007/s00187-020-00300-5

Mulvey, J. M., Lu, N., & Sweemer, J. (2001). Rebalancing strategies for multi-period asset allocation. *The Journal of Wealth Management, 4*(2), 51-58. doi:10.3905/jwm.2001.320412

Murphy, K. P. (2012). *machine learning: A Probabilistic Perspective.* London: The MIT Press.

Nian, K., Coleman, T. F., & Li, Y. (2018). learning minimum variance discrete hedging directly from the market. *Quantitative finance, 18*(7), 1115-1128. doi:10.1080/14697688.2017.1413245

Nian, K., Coleman, T. F., & Li, Y. (2021). learning sequential option hedging models from Market Data. *Journal of Banking & finance, 133*. doi:10.1016/j.jbankfin.2021.106277

Novak, M. G., & Velušček, D. (2015). Prediction of stock price movement based on daily high prices. *Quantitative finance, 16*(5), 793-826. doi:10.1080/14697688.2015.1070960

Nowak, A. D., Price, B. S., & Smith, P. S. (2019). Real estate dictionaries across space and Time. *The Journal of Real Estate finance and Economics, 62*(1), 139-163. doi:10.1007/s11146-019-09740-w

Olson, L. M., Qi, M., Zhang, X., & Zhao, X. (2021). machine learning loss given default for corporate debt. *Journal of Empirical finance, 64*, 144-159. doi:10.1016/j.jempfin.2021.08.009

Onorante, L., & Raftery, A. E. (2016). Dynamic model averaging in large model spaces using dynamic occam's window. *European Economic Review, 81*, 2-14. doi:10.1016/j.euroecorev.2015.07.013

Oz, I. O., Yelkenci, T., & Meral, G. (2021). The role of Earnings Components and machine learning on the revelation of deteriorating firm performance. *International Review of Financial Analysis, 77*. doi:10.1016/j.irfa.2021.101797

Pace, R. K., & Hayunga, D. (2019). Examining the information content of residuals from hedonic and spatial models using trees and forests. *The Journal of Real Estate finance and Economics, 60*(1-2), 170-180. doi:10.1007/s11146-019-09724-w

Perols, J. (2011). Financial statement fraud detection: An analysis of statistical and machine learning algorithms. *AUDITING: A Journal of Practice & Theory, 30*(2), 19-50. doi:10.2308/ajpt-50009

Pierdzioch, C., & Risse, M. (2018). A machine-learning analysis of the rationality of Aggregate Stock Market forecasts. *International Journal of finance & Economics, 23*(4), 642-654. doi:10.1002/ijfe.1641

Plakandaras, V., Gupta, R., & Wohar, M. E. (2017). The depreciation of the pound post-Brexit: Could it have been predicted? *finance Research Letters, 21*, 206-213. doi:10.1016/j.frl.2016.12.003

Poloz, S. S. (2021). Technological Progress and Monetary Policy: Managing the fourth industrial revolution. *Journal of International Money and finance*, 114. doi:10.1016/j.jimonfin.2021.102373

Polyzos, S., Samitas, A., & Katsaiti, M.-S. (2020). Who is unhappy for Brexit? A machine-learning, agent-based study on financial instability. *International Review of Financial Analysis, 72*. doi:10.1016/j.irfa.2020.101590

Pun, C. S. (2021). A sparse learning approach to relative-volatility-managed portfolio selection. *SIAM Journal on Financial Mathematics, 12*(1), 410-445. doi:10.1137/19M1291674

Pun, C. S., & Wang, L. (2020). A cost-effective approach to portfolio construction with range-based risk measures. *Quantitative finance, 21*(3), 431-447. doi:10.1080/14697688.2020.1781237

Pyo, S., & Lee, J. (2018). Exploiting the low-risk anomaly using machine learning to enhance the black–litterman framework: Evidence from South Korea. *Pacific-Basin finance Journal, 51*, 1-12. doi:10.1016/j.pacfin.2018.06.002

Rasekhschaffe, K. C., & Jones, R. C. (2019). machine learning for Stock Selection. *Financial Analysts Journal, 75*(3), 70-88. doi:10.1080/0015198x.2019.1596678

Renault, T. (2017). Intraday online investor sentiment and return patterns in the U.S. stock market. *Journal of Banking & finance, 84*, 25-40. doi:10.1016/j.jbankfin.2017.07.002

Rialti, R., Marzi, G., Ciappei, G., & Busso, D. (2019). Big Data and Dynamic Capabilities: A Bibliometric Analysis and Systematic Literature Review. *Management Decision, 57*, 2052-2068. doi:10.1108/MD-07-2018-0821

Risk, J., & Ludkovski, M. (2018). Sequential Design and spatial modeling for portfolio tail risk measurement. *SIAM Journal on Financial Mathematics, 9*(4), 1137-1174. doi:10.1137/17M1158380

Risse, M., & Ohl, L. (2017). Using dynamic model averaging in state space representation with Dynamic Occam's window and applications to the stock and gold market. *Journal of Empirical finance, 44*, 158-176. doi:10.1016/j.jempfin.2017.09.005

Routledge, B. R. (2019). machine learning and Asset Allocation. *Financial Management, 48*(4), 1069-1094. doi:10.1111/fima.12303

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1985). learning Internal Representations by Error Propagation. doi:10.21236/ada164453

Rybinski, K. (2020). The forecasting power of the multi-language narrative of sell-side research: A machine learning evaluation. *finance Research Letters, 34*. doi:10.1016/j.frl.2019.08.009

*S&P Global Market Intelligence delivers essential information so you can make decisions with conviction.* (2021, 12 6). Retrieved from S&P Global: https://www.spglobal.com/marketintelligence/en/documents/112727-gics-mapbook_2018_v3_letter_digitalspreads.pdf

Samitas, A., Kampouris, E., & Kenourgios, D. (2020). machine learning as an early warning system to predict financial crisis. *International Review of Financial Analysis, 71*. doi:10.1016/j.irfa.2020.101507

Samuel, A. (1959). Some Studies in machine learning Using the Game of Checkers. *IBM J. Res Dev*(3), 535–554.

Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J., & Napolitano, A. (2010). RUSBoost: A hybrid approach to alleviating class imbalance. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, 40*(1), 185-197. doi:10.1109/tsmca.2009.2029559

Sermpinis, G., Laws, J., & Dunis, C. L. (2013). Modelling Commodity Value at Risk with psi sigma neural networks using open–high–low–close data. *The European Journal of finance, 21*(4), 316-336. doi:10.1080/1351847x.2012.744763

Sirignano, J. A. (2018). Deep learning for Limit Order Books. *Quantitative finance, 19*(4), 549-570. doi:10.1080/14697688.2018.1546053

Sirignano, J., & Cont, R. (2019). Universal features of Price Formation in financial markets: Perspectives from Deep learning. *Quantitative finance, 19*(9), 1449-1459. doi:10.1080/14697688.2019.1622295

Sirignano, J., & Spiliopoulos, K. (2017). Stochastic Gradient Descent in Continuous Time. *SIAM Journal on Financial Mathematics, 8*(1), 933-961. doi:10.1137/17m1126825

Slapnik, U., & Lončarski, I. (2021). On the information content of sovereign credit rating reports: Improving the predictability of rating transitions. *Journal of International Financial Markets, Institutions and Money, 73*. doi:10.1016/j.intfin.2021.101344

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society fo rInformation Science, 24*(4), 265-269. doi:10.1002/asi.4630240406

Soto, P. E. (2020). Breaking the word bank: Measurement and effects of bank level uncertainty. *Journal of Financial Services Research, 59*(1-2), 1-45.

Sun, T. (. (2019). Applying deep learning to audit procedures: An illustrative framework. *accounting Horizons, 33*(3), 89-109. doi:10.2308/acch-52455

Sutton, S., Holt, M., & Arnold, V. (2016). "The reports of my death are greatly exaggerated"—Artificial intelligence research in accounting. *International Journal of accounting Information Systems, 22*, 60-73. doi:10.1016/j.accinf.2016.07.005

Tantri, P. (2020). Fintech for the poor: Financial intermediation without discrimination*. *Review of finance*, 561-593. doi:10.1093/rof/rfaa039

Tashiro, D., Matsushima, H., Izumi, K., & Sakaji, H. (2019). Encoding of high-frequency order information and prediction of short-term stock price by Deep learning. *Quantitative finance, 19*(9), 1499-1506. doi:10.1080/14697688.2019.1622314

Thomaidou, S., Liakopoulos, K., & Vazirgiannis, M. (2014). Toward an integrated framework for automated development and optimization of online advertising campaigns. *Intelligent Data Analysis, 18*, 1199-1227. doi:10.3233/IDA-140691

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological), 58*(1), 267-288. doi:10.1111/j.2517-6161.1996.tb02080.x
Tsang, K. H., & Wong, H. Y. (2020). Deep-learning solution to portfolio selection with serially dependent returns. *SIAM Journal on Financial Mathematics, 11*(2), 593-619. doi:10.1137/19m1274924
Turing, A. (1950). Computing machinery and intelligence. *Mind*(236), 433–460. doi:10.1093/mind/LIX.236.433
Van Eck, N., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics, 84*(2), 523-538. doi:10.1007/s11192-009-0146-3
Varetto, F. (1998). Genetic algorithms applications in the analysis of insolvency risk. *Journal of Banking & finance, 22*(10-11), 1421-1439. doi:10.1016/s0378-4266(98)00059-4
Viebig, J. (2019). Exuberance in financial markets: Evidence from machine learning algorithms. *Journal of Behavioral finance, 21*(2), 128-135. doi:10.1080/15427560.2019.1663849
Viswanathan, P. K., Srinivasan, S., & Hariharan, N. (2020). Predicting financial health of banks for investor guidance using machine learning algorithms. *Journal of Emerging Market finance, 19*(2), 226-261. doi:10.1177/0972652720913478
*VOSviewer - Visualizing scientific landscapes*. (2021, 11 29). Retrieved from VOSviewer - Visualizing scientific landscapes: https://www.vosviewer.com/
Vrontos, S. D., Galakis, J., & Vrontos, I. D. (2021). Implied volatility directional forecasting: A machine learning approach. *Quantitative finance, 21*(10), 1687-1706. doi:10.1080/14697688.2021.1905869
Wall, L. D. (2018). Some financial regulatory implications of Artificial Intelligence. *Journal of Economics and Business, 100*, 55-63. doi:10.1016/j.jeconbus.2018.05.003
Wang, C., Zhang, Y., Zhang, W., & Gong, X. (2021). Textual sentiment of comments and collapse of P2P platforms: Evidence from China's P2P market. *Research in International Business and finance, 58*. doi:10.1016/j.ribaf.2021.101448
Wang, C.-W., Zhang, J., & Zhu, W. (2021). Neighbouring prediction for mortality. *ASTIN Bulletin, 51*(3), 689-718. doi:10.1017/asb.2021.13
Wang, H., & Zhou, X. Y. (2020). Continuous-time mean–variance portfolio selection: A reinforcement learning framework. *Mathematical finance, 30*(4). doi:10.1111/mafi.12281
Weigand, A. (2019). machine learning in empirical asset pricing. *Financial Markets and Portfolio Management, 33*(1), 93-104. doi:10.1007/s11408-019-00326-3
Wiese, M., Knobloch, R., Korn, R., & Kretschmer, P. (2020). Quant Gans: Deep generation of financial time series. *Quantitative finance, 20*(9), 1419-1440. doi:10.1080/14697688.2020.1730426
Wolff, D., & Neugebauer, U. (2019). Tree-based machine learning approaches for equity market predictions. *Journal of Asset Management, 20*(4), 273-288. doi:10.1057/s41260-019-00125-5
*WordStream Online Advertising Made Easy*. (2021, 9 30). Retrieved from Wordstream: https://www.wordstream.com/

Wu, P., & Siwasarit, W. (2019). Capturing the order imbalance with Hidden Markov Model: A case of set50 and KOSPI50. *Asia-Pacific Financial Markets, 27*(1), 115-144. doi:10.1007/s10690-019-09285-1

Wu, W., Lee, V. C., & Tan, T. Y. (2006). Data preprocessing and data parsimony in corporate failure forecast models: Evidence from Australian Materials Industry. *accounting and finance, 46*(2), 327-345. doi:10.1111/j.1467-629x.2006.00170.x

Xue, F., Li, X., Zhang, T., & Hu, N. (2021). Stock market reactions to the COVID-19 pandemic: The moderating role of Corporate Big Data Strategies based on word2vec. *Pacific-Basin finance Journal, 68*. doi:10.1016/j.pacfin.2021.101608

Yang, S. Y., Qiao, Q., Beling, P. A., Scherer, W. T., & Kirilenko, A. A. (2015). Gaussian process-based algorithmic trading strategy identification. *Quantitative finance, 15*(10), 1683-1703. doi:10.1080/14697688.2015.1011684

Zaremba, A., Kizys, R., Tzouvanas, P., Aharon, D. Y., & Demir, E. (2021). The quest for Multidimensional Financial Immunity to the COVID-19 pandemic: Evidence from international stock markets. *Journal of International Financial Markets, Institutions and Money, 71*. doi:10.1016/j.intfin.2021.101284

Zhang, J., & Huang, W. (2021). Option hedging using LSTM-RNN: An empirical analysis. *Quantitative finance, 21*(10), 1753-1772. doi:10.1080/14697688.2021.1905171

Zhang, Y., Chu, G., & Shen, D. (2021). The role of investor attention in predicting stock prices: The long short-term memory networks perspective. *finance Research Letters, 38*. doi:10.1016/j.frl.2020.101484

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67*(2), 301-320. doi:10.1111/j.1467-9868.2005.00503.x

# Appendix A: "Search Query"

Initial Keywords (for accounting & finance)

| Keywords | Grouping |
|---|---|
| finance | financ* |
| financial leverage | leverage |
| financial institution | financial institution |
| deficit financing | deficit financing |
| financial capital | capital |
| financial transaction | financial transaction |
| accounting | accounting |
| cpa | cpa |
| accountant | accountant |
| bookkeeping | bookkeeping |
| chartered accountant | accountant |
| management accounting | accounting |
| certified management accountant | accountant |
| trial balance | trial balance |
| retained earnings | retained earnings |
| financial accounting | financial accounting |
| petty cash | petty cash |
| irs account | irs |
| cost accounting | accounting |
| accountancy | accountancy |
| ifrs 9 | ifrs |
| general ledger | ledger |
| account receivable | account receivable |
| cpa near me | cpa |
| forensic accounting | accounting |
| iasb | iasb |
| accumulated depreciation | accumulated depreciation |
| accrual accounting | accounting |
| tax preparer near me | tax |
| accountants near me | accountant |
| profit and loss account | profit and loss account |
| prepaid expenses | prepaid expenses |
| managerial accounting | accounting |
| account payable | account payable |
| payable | payable |
| tax accountant near me | tax |
| ledger account | ledger |
| receivable | receivable |
| tax account | tax |
| t account | t account |
| capital account | capital account |

| | |
|---|---|
| certified public accountant | accountant |
| tax advisor | tax |
| accounting services | accounting |
| bookkeeping services | bookkeeping |
| corporate accounting | accounting |
| basic accounting | accounting |
| accounting firm | accounting |
| vat payment | vat |
| business accounting | accounting |
| account reconciliation | account reconciliation |
| non current asset | asset |
| accounting standard | accounting |
| income tax | tax |
| quickbooks | quickbooks |
| account | accounting |
| tax | tax |
| audit | audit* |
| tax return | tax |
| payroll | payroll |
| auditor | audit* |
| accountant salary | accountant |
| accounts payable | accounts payable |
| accounting software | accounting |
| corporation tax | tax |
| what is accounting | accounting |
| tax preparation | tax |
| tax services | tax |
| accounting jobs | accounting |
| accounting principles | accounting |
| accounting standards | accounting |
| accounting equation | accounting |
| financial statements | financ* |
| tax accountant | accountant |
| payroll tax | tax |
| accounting concepts | accounting |
| accounting firms | accounting |
| accounting app | accounting |
| accounting definition | accounting |
| bookkeeping course | bookkeeping |
| free accounting software | accounting |
| bookkeeping jobs | bookkeeping |
| account management | accounting |
| international accounting standards | accounting |
| the accountant online | accountant |
| accounting courses | accounting |
| accounting degree | accounting |
| book keeping | bookkeeping |

| | |
|---|---|
| tax planning | tax |
| tax consultant | tax |
| principles of accounting | accounting |
| stock market | stock |
| stock | stock |
| stock quotes | stock |
| ministry of finance | financ* |
| stock prices | stock |
| financial news | financ* |
| efinance | financ* |
| financial markets | financ* |
| american financing | financ* |
| business finance | financ* |
| stock market quotes | stock |
| sources of finance | financ* |

Finalized Query:

ALL=(( "financ*" or "leverage" or "financial institution" or "deficit financing" or "capital" or "financial transaction" or "accounting" or "cpa" or "accountant" or "bookkeeping" or "trial balance" or "retained earnings" or "financial accounting" or "petty cash" or "irs" or "accountancy" or "ifrs" or "ledger" or "account receivable" or "iasb" or "accumulated depreciation" or "tax" or "profit and loss account" or "prepaid expenses" or "account payable" or "payable" or "receivable" or "t account" or "capital account" or "vat" or "account reconciliation" or "asset" or "quickbooks" or "audit*" or "payroll" or

"accounts payable" or "stock") and ("machine learning" or "Artificial Intelligence" or "Deep learning" or "Supervised")) and Articles or Review Articles (Document Types) and Economics or Business finance (Web of Science Categories)

# Appendix B: "Corpus"

| Authors | Article Title | Journal | Pub. Year | ML Cat. | Field | Grouping | # of ML Algo. | Best Algo. | Continent | Cluster |
|---|---|---|---|---|---|---|---|---|---|---|
| Agrawal, S; Azar, PD; Lo, AW; Singh, T | Momentum, Mean-Reversion, and Social Media: Evidence from StockTwits and Twitter | JOURNAL OF PORTFOLIO MANAGEMENT | 2018 | Supervised | Fin | Markets & Predictions | 2 | N/A | North America | Yellow |
| Ahmed, R | Monetary policy spillovers under intermediate exchange rate regimes | JOURNAL OF INTERNATIONAL MONEY AND FINANCE | 2021 | Supervised | Fin | Markets & Predictions | 1 | N/A | Worldwide | Purple |
| Ahn, Y; Tsai, SCA | What factors are associated with stock price jumps in high frequency? | PACIFIC-BASIN FINANCE JOURNAL | 2021 | Supervised | Fin | Markets & Predictions | 2 | N/A | Asia | Red |
| Akbari, A; Ng, L; Solnik, B | Drivers of economic and financial integration: A machine learning approach | JOURNAL OF EMPIRICAL FINANCE | 2021 | Supervised | Fin | Market Variables & Crisis Prediction | 1 | N/A | Worldwide | Purple |
| Albanese, C; Crepey, S; Hoskinson, R; Saadeddine, B | XVA analysis from the balance sheet | QUANTITATIVE FINANCE | 2021 | Supervised | Fin | Other | 1 | N/A | Worldwide | Red |
| Amat, C; Michalski, T; Stoltz, G | Fundamentals and exchange rate forecastability with simple machine learning methods | JOURNAL OF INTERNATIONAL MONEY AND FINANCE | 2018 | Supervised | Fin | Markets & Predictions | 2 | Ridge Regression | Worldwide | Green |
| Amin, MH; Mohamed, EKA; Elragal, A | CSR disclosure on Twitter: Evidence from the UK | INTERNATIONAL JOURNAL OF ACCOUNTING INFORMATION SYSTEMS | 2021 | Unsupervised | Acc | Corporate | 1 | N/A | Europe | Yellow |
| Amini, S; Elmore, R; Oztekin, OO; Strauss, J | Can machines learn capital structure dynamics?* | JOURNAL OF CORPORATE FINANCE | 2021 | Supervised | Fin | Corporate | 5 | Random Forest | North America | Blue |

| Authors | Article Title | Journal | Pub. Year | ML Cat. | Field | Grouping | # of ML Algo. | Best Algo. | Continent | Cluster |
|---|---|---|---|---|---|---|---|---|---|---|
| An, JF; Rau, R | Finance, technology and disruption | EUROPEAN JOURNAL OF FINANCE | 2021 | N/A | Fin | Market Variables & Crisis Prediction | 0 | N/A | N/A | Cyan |
| Anghel, DG | A reality check on trading rule performance in the cryptocurrency market: Machine learning vs. technical analysis | FINANCE RESEARCH LETTERS | 2021 | Supervised | Fin | Markets & Predictions | 5 | N/A | Worldwide | Green |
| Appelbaum, D; Showalter, DS; Sun, T; Vasarhelyi, MA | A Framework for Auditor Data Literacy: A Normative Position | ACCOUNTING HORIZONS | 2021 | N/A | Acc | Other | 0 | N/A | North America | Cyan |
| Arakelian, V; Dellaportas, P; Savona, R; Vezzoli, M | Sovereign risk zones in Europe during and after the debt crisis | QUANTITATIVE FINANCE | 2019 | Supervised | Fin | Bankruptcy Prediction\Default | 1 | N/A | Europe | Purple |
| Avanzi, B; Taylor, G; Wang, M; Wong, B | SynthETIC: An individual insurance claim simulator with feature control | INSURANCE MATHEMATICS & ECONOMICS | 2021 | N/A | Fin | Algorithm\Methodology Creation | 0 | N/A | Worldwide | Red |
| Avramov, D; Li, MW; Wang, H | Predicting corporate policies using downside risk: A machine learning approach | JOURNAL OF EMPIRICAL FINANCE | 2021 | Supervised | Fin | Market Variables & Crisis Prediction | 1 | N/A | North America | Yellow |
| Baek, S; Mohanty, SK; Glambosky, M | COVID-19 and stock market volatility: An industry level analysis | FINANCE RESEARCH LETTERS | 2020 | Supervised | Fin | Market Variables & Crisis Prediction | 3 | N/A | North America | Cyan |
| Bao, Y; Ke, B; Li, B; Yu, YJ; Zhang, J | Detecting Accounting Fraud in Publicly Traded US Firms Using a Machine Learning Approach | JOURNAL OF ACCOUNTING RESEARCH | 2020 | Supervised | Acc | Financial Statements Misreporting | 3 | RusBoost | North America | Blue |

| Authors | Article Title | Journal | Pub. Year | ML Cat. | Field | Grouping | # of ML Algo. | Best Algo. | Continent | Cluster |
|---|---|---|---|---|---|---|---|---|---|---|
| Bender, M; Panz, S | A general framework for the identification and categorization of risks: an application to the context of financial markets | JOURNAL OF RISK | 2021 | Unsupervised | Fin | Market Variables & Crisis Prediction | 2 | N/A | Worldwide | Green |
| Berninger, M; Kiesel, F; Schiereck, D; Gaar, E | Citations and the readers' information-extracting costs of finance articles | JOURNAL OF BANKING & FINANCE | 2021 | Unsupervised | Fin | Literature, Corpus Analysis | 1 | N/A | Worldwide | Yellow |
| Bertomeu, J | Machine learning improves accounting: discussion, implementation and research opportunities | REVIEW OF ACCOUNTING STUDIES | 2020 | N/A | Acc | ML Critisism | 0 | N/A | Worldwide | Cyan |
| Bertomeu, J; Cheynel, E; Floyd, E; Pan, WQ | Using machine learning to detect misstatements | REVIEW OF ACCOUNTING STUDIES | 2021 | Supervised | Acc | Financial Statements Misreporting | 4 | RusBoost | North America | Blue |
| Beutel, J; List, S; von Schweinitz, G | Does machine learning help us predict banking crises? | JOURNAL OF FINANCIAL STABILITY | 2019 | Supervised | Fin | Market Variables & Crisis Prediction | 6 | Logistic Regression | Europe | Blue |
| Bianchi, D; Buchner, M; Tamoni, A | Bond Risk Premiums with Machine Learning | REVIEW OF FINANCIAL STUDIES | 2021 | Supervised | Fin | Markets & Predictions | 7 | Neural Networks | North America | Purple |
| Bonato, M; Gkillas, K; Gupta, R; Pierdzioch, C | A note on investor happiness and the predictability of realized volatility of gold | FINANCE RESEARCH LETTERS | 2021 | Supervised | Fin | Markets & Predictions | 1 | N/A | Worldwide | Green |
| Boot, A; Hoffmann, P; Laeven, L; Ratnovski, L | Fintech: what's old, what's new? | JOURNAL OF FINANCIAL STABILITY | 2021 | N/A | Fin | Other | 0 | N/A | Worldwide | Yellow |

| Authors | Article Title | Journal | Pub. Year | ML Cat. | Field | Grouping | # of ML Algo. | Best Algo. | Continent | Cluster |
|---|---|---|---|---|---|---|---|---|---|---|
| Booth, A; Gerding, E; Mcgroarty, F | Performance-weighted ensembles of random forests for predicting price impact | QUANTITATIVE FINANCE | 2015 | Supervised | Fin | Markets & Predictions | 6 | Neural Network | Worldwide | Red |
| Bottazzi, G; Giachini, D | Far from the madding crowd: collective wisdom in prediction markets | QUANTITATIVE FINANCE | 2019 | N/A | Fin | Other | 0 | N/A | Worldwide | Blue |
| Brown, NC; Crowley, RM; Elliott, WB | What Are You Saying? Using topic to Detect Financial Misreporting | JOURNAL OF ACCOUNTING RESEARCH | 2020 | Unsupervised | Acc | Financial Statements Misreporting | 1 | N/A | North America | Yellow |
| Brown, P; Gallery, G; Goei, O | Does market misvaluation help explain share market long-run underperformance following a seasoned equity issue? | ACCOUNTING AND FINANCE | 2006 | Supervised | Fin | Markets & Predictions | 1 | N/A | Oceania | Purple |
| Bubna, A; Das, SR; Prabhala, N | Venture Capital Communities | JOURNAL OF FINANCIAL AND QUANTITATIVE ANALYSIS | 2020 | Supervised & Unsupervised | Fin | Corporate | 3 | N/A | North America | Cyan |
| Bucci, A | Realized Volatility Forecasting with Neural Networks | JOURNAL OF FINANCIAL ECONOMETRICS | 2020 | Supervised | Fin | Markets & Predictions | 6 | LSTM | North America | Green |
| Buehler, H; Gonon, L; Teichmann, J; Wood, B | Deep hedging | QUANTITATIVE FINANCE | 2019 | Reinforcement | Fin | Algorithm\Methodology Creation | 1 | N/A | North America | Red |
| Burggraf, T | Beyond risk parity - A machine learning-based hierarchical risk parity approach on cryptocurrencies | FINANCE RESEARCH LETTERS | 2021 | Unsupervised | Fin | Markets & Predictions | 1 | N/A | Worldwide | Green |

| Authors | Article Title | Journal | Pub. Year | ML Cat. | Field | Grouping | # of ML Algo. | Best Algo. | Continent | Cluster |
|---|---|---|---|---|---|---|---|---|---|---|
| Butaru, F; Chen, QP; Clark, B; Das, S; Lo, AW; Siddique, A | Risk and risk management in the credit card industry | JOURNAL OF BANKING & FINANCE | 2016 | Supervised | Fin | Bankruptcy Prediction\ Default | 3 | Decision Tree | North America | Blue |
| Caglayan, M; Pham, T; Talavera, O; Xiong, X | Asset mispricing in peer-to-peer loan secondary markets | JOURNAL OF CORPORATE FINANCE | 2020 | Supervised | Fin | Market Variables & Crisis Prediction | 3 | LASSO | Europe | Purple |
| Cai, CW; Linnenluecke, MK; Marrone, M; Singh, AK | Machine Learning and Expert Judgement: Analyzing Emerging Topics in Accounting and Finance Research in the Asia-Pacific | ABACUS-A JOURNAL OF ACCOUNTING FINANCE AND BUSINESS STUDIES | 2019 | Unsupervised | Acc | Literature, Corpus Analysis | 1 | Own Creation | Worldwide | Yellow |
| Calomiris, CW; Mamaysky, H | How news and its context drive risk and returns around the world | JOURNAL OF FINANCIAL ECONOMICS | 2019 | Supervised & Unsupervised | Fin | Markets & Predictions | 2 | N/A | Worldwide | Yellow |
| Carbonneau, A; Godin, F | Equal risk pricing of derivatives with deep hedging | QUANTITATIVE FINANCE | 2021 | Reinforcement | Fin | Markets & Predictions | 1 | N/A | North America | Red |
| Chen, MA; Wu, QX; Yang, BZ | How Valuable Is FinTech Innovation? | REVIEW OF FINANCIAL STUDIES | 2019 | Supervised | Fin | Other | 6 | Combination of SVM & Neural network | North America | Cyan |
| Cheng, C; Jones, S; Moser, WJ | Abnormal trading behavior of specific types of shareholders before US firm bankruptcy and its implications for firm bankruptcy prediction | JOURNAL OF BUSINESS FINANCE & ACCOUNTING | 2018 | Supervised | Acc | Bankruptcy Prediction\ Default | 1 | N/A | North America | Blue |
| Cheng, CH; Wang, SH | A quarterly time-series classifier based on a reduced-dimension generated rules method for identifying financial distress | QUANTITATIVE FINANCE | 2015 | Supervised | Fin | Bankruptcy Prediction\ Default | 4 | Neural Network | Asia | Blue |

| Authors | Article Title | Journal | Pub. Year | ML Cat. | Field | Grouping | # of ML Algo. | Best Algo. | Continent | Cluster |
|---|---|---|---|---|---|---|---|---|---|---|
| Cheng, D; Cirillo, P | A reinforced urn process modeling of recovery rates and recovery times | JOURNAL OF BANKING & FINANCE | 2018 | Reinforcement | Fin | Bankruptcy Prediction\ Default | 1 | N/A | North America | Blue |
| Cheng, X; Jin, Z; Yang, HL | OPTIMAL INSURANCE STRATEGIES: A HYBRID DEEP LEARNING MARKOV CHAIN APPROXIMATION APPROACH | ASTIN BULLETIN | 2020 | Unsupervised | Fin | Algorithm\ Methodology Creation | 1 | N/A | Worldwide | Red |
| Colak, G; Fu, MC; Hasan, I | Why are some Chinese firms failing in the US capital markets? A machine learning approach | PACIFIC-BASIN FINANCE JOURNAL | 2020 | Supervised | Fin | Markets & Predictions | 4 | Gradient Boosting | Worldwide | Blue |
| Cont, R; Kukanov, A | Optimal order placement in limit order markets | QUANTITATIVE FINANCE | 2017 | N/A | Fin | Algorithm\ Methodology Creation | 0 | N/A | Worldwide | Red |
| Corazza, L; Truant, E; Scagnelli, SD; Mio, C | Sustainability reporting after the Costa Concordia disaster: a multi-theory study on legitimacy, impression management and image restoration | ACCOUNTING AUDITING & ACCOUNTABILITY JOURNAL | 2020 | Unknown | Acc | Corporate | 1 | N/A | N/A | Yellow |
| Costello, AM; Down, AK; Mehta, MN | Machine plus man: A field experiment on the role of discretion in augmenting AI-based lending models | JOURNAL OF ACCOUNTING & ECONOMICS | 2020 | N/A | Acc | Other | 0 | N/A | North America | Yellow |
| Cowden, C; Fabozzi, FJ; Nazemi, A | Default Prediction of Commercial Real Estate Properties Using Machine Learning Techniques | JOURNAL OF PORTFOLIO MANAGEMENT | 2019 | Supervised | Fin | Bankruptcy Prediction\ Default | 4 | Support Vector Machine | North America | Blue |
| Creamer, G | Model calibration and automated trading agent for Euro futures | QUANTITATIVE FINANCE | 2012 | Supervised | Fin | Markets & Predictions | 4 | LogitBoost | Europe | Green |

| Authors | Article Title | Journal | Pub. Year | ML Cat. | Field | Grouping | # of ML Algo. | Best Algo. | Continent | Cluster |
|---|---|---|---|---|---|---|---|---|---|---|
| Creamer, G; Freund, Y | Automated trading with boosting and expert weighting | QUANTITATIVE FINANCE | 2010 | Supervised | Fin | Markets & Predictions | 1 | N/A | North America | Green |
| Creamer, GG | Can a corporate network and news sentiment improve portfolio optimization using the Black-Litterman model? | QUANTITATIVE FINANCE | 2015 | Supervised & Unsupervised | Fin | Markets & Predictions | 4 | N/A | Worldwide | Green |
| Creamer, GG; Lee, C | A multivariate distance nonlinear causality test based on partial distance correlation: a machine learning application to energy futures | QUANTITATIVE FINANCE | 2019 | Supervised | Fin | Markets & Predictions | 2 | N/A | Worldwide | Blue |
| De Moor, L; Luitel, P; Sercu, P; Vanpee, R | Subjectivity in sovereign credit ratings | JOURNAL OF BANKING & FINANCE | 2018 | Supervised | Fin | Bankruptcy Prediction\Default | 2 | Random Forest | Worldwide | Yellow |
| De Spiegeleer, J; Madan, DB; Reyners, S; Schoutens, W | Machine learning for quantitative finance: fast derivative pricing, hedging and fitting | QUANTITATIVE FINANCE | 2018 | Supervised | Fin | Markets & Predictions | 1 | N/A | North America | Red |
| Deku, SY; Kara, A; Semeyutin, A | The predictive strength of MBS yield spreads during asset bubbles | REVIEW OF QUANTITATIVE FINANCE AND ACCOUNTING | 2021 | Supervised | Fin | Markets & Predictions | 5 | Random Forest | Europe | Red |
| Devriendt, S; Antonio, K; Reynkens, T; Verbelen, R | Sparse regression with Multi-type Regularized Feature modeling | INSURANCE MATHEMATICS & ECONOMICS | 2021 | Supervised | Fin | Algorithm\Methodology Creation | 1 | N/A | Europe | Purple |
| D'Hondt, C; De Winne, R; Ghysels, E; Raymond, S | Artificial Intelligence Alter Egos: Who might benefit from robo-investing? | JOURNAL OF EMPIRICAL FINANCE | 2020 | Supervised | Fin | Markets & Predictions | 4 | Neural Network | Worldwide | Green |

| Authors | Article Title | Journal | Pub. Year | ML Cat. | Field | Grouping | # of ML Algo. | Best Algo. | Continent | Cluster |
|---|---|---|---|---|---|---|---|---|---|---|
| Ding, KX; Lev, B; Peng, X; Sun, T; Vasarhelyi, MA | Machine learning improves accounting estimates: evidence from insurance payments | REVIEW OF ACCOUNTING STUDIES | 2020 | Supervised | Acc | Financial Statements Misreporti | 4 | Random Forest | North America | Blue |
| Donovan, J; Jennings, J; Koharki, K; Lee, J | Measuring credit risk using qualitative disclosure | REVIEW OF ACCOUNTING STUDIES | 2021 | Supervised | Acc | Bankruptcy Prediction \Default | 3 | N/A | North America | Yellow |
| Duan, Y; Hsieh, TS; Wang, RR; Wang, ZH | Entrepreneurs' facial trustworthiness, gender, and crowdfunding success | JOURNAL OF CORPORATE FINANCE | 2020 | Supervised | Fin | Corporate | 2 | N/A | Worldwide | Yellow |
| Dunis, CL; Laws, J; Middleton, PW; Karathanasopoulos, A | Trading and hedging the corn/ethanol crush spread using time-varying leverage and nonlinear models | EUROPEAN JOURNAL OF FINANCE | 2015 | Supervised | Fin | Markets & Predictions | 2 | Genetic Algorithm | Worldwide | Green |
| Easley, D; de Prado, ML; O'Hara, M; Zhang, ZB | Microstructure in the Machine Age | REVIEW OF FINANCIAL STUDIES | 2021 | Supervised | Fin | Market Variables & Crisis Prediction | 2 | Random Forest | North America | Red |
| El-Haj, M; Rayson, P; Walker, M; Young, S; Simaki, V | In search of meaning: Lessons, resources and next steps for computational analysis of financial discourse | JOURNAL OF BUSINESS FINANCE & ACCOUNTING | 2019 | N/A | Acc | Literature, Corpus Analysis | 0 | N/A | N/A | Yellow |
| Engle, RF; Hansen, MK; Karagozoglu, AK; Lunde, A | News and Idiosyncratic Volatility: The Public Information Processing Hypothesis | JOURNAL OF FINANCIAL ECONOMETRICS | 2021 | Supervised | Fin | Markets & Predictions | 1 | N/A | North America | Red |
| Erel, I; Stern, LH; Tan, CH; Weisbach, MS | Selecting Directors Using Machine Learning | REVIEW OF FINANCIAL STUDIES | 2021 | Supervised | Fin | Corporate | 4 | Gradient Boosting | North America | Cyan |

| Authors | Article Title | Journal | Pub. Year | ML Cat. | Field | Grouping | # of ML Algo. | Best Algo. | Continent | Cluster |
|---|---|---|---|---|---|---|---|---|---|---|
| Fang, J; Lin, JW; Xia, ST; Xia, ZK; Hu, SL; Liu, X; Jiang, Y | Neural network-based automatic factor construction | QUANTITATIVE FINANCE | 2020 | Supervised | Fin | Algorithm\ Methodology Creation | 7 | Neural Network | Asia | Purple |
| Felix, L; Kraussl, R; Stork, P | Implied volatility sentiment: a tale of two tails | QUANTITATIVE FINANCE | 2020 | Supervised | Fin | Markets & Predictions | 4 | Ridge Regression | North America | Purple |
| Fischer, JA; Pohl, P; Ratz, D | A machine learning approach to univariate time series forecasting of quarterly earnings | REVIEW OF QUANTITATIVE FINANCE AND ACCOUNTING | 2020 | Supervised | Fin | Algorithm\ Methodology Creation | 1 | N/A | North America | Blue |
| Foley, S; Kwan, A; McInish, TH; Philip, R | Director discretion and insider trading profitability | PACIFIC-BASIN FINANCE JOURNAL | 2016 | Supervised | Fin | Markets & Predictions | 1 | N/A | Oceania | Blue |
| Friberg, R; Seiler, T | Different ways of managing risk as reported in 10-Ks: A supervised learning approach | FINANCIAL REVIEW | 2021 | Supervised | Fin | Other | 2 | N/A | North America | Yellow |
| Fritzsch, S; Scharner, P; Weiss, G | Estimating the relation between digitalization and the market value of insurers | JOURNAL OF RISK AND INSURANCE | 2021 | Unsupervised | Fin | Insurance Predictions | 1 | N/A | North America | Yellow |
| Funahashi, H | Artificial neural network for option pricing with and without asymptotic correction | QUANTITATIVE FINANCE | 2021 | Supervised | Fin | Markets & Predictions | 1 | N/A | Europe | Red |
| Gan, GJ | Application of data clustering and machine learning in variable annuity valuation | INSURANCE MATHEMATICS & ECONOMICS | 2013 | Supervised & Unsupervised | Fin | Insurance Predictions | 2 | N/A | Unknown | Red |

| Authors | Article Title | Journal | Pub. Year | ML Cat. | Field | Grouping | # of ML Algo. | Best Algo. | Continent | Cluster |
|---|---|---|---|---|---|---|---|---|---|---|
| Geertsema, P; Lu, H | The correlation structure of anomaly strategies | JOURNAL OF BANKING & FINANCE | 2020 | Unsupervised | Fin | Markets & Predictions | 1 | N/A | North America | Purple |
| Giglio, S; Liao, Y; Xiu, DC | Thousands of Alpha Tests | REVIEW OF FINANCIAL STUDIES | 2021 | N/A | Fin | Algorithm\Methodology Creation | 0 | N/A | North America | Purple |
| Gkillas, K; Gupta, R; Pierdzioch, C | Forecasting realized oil-price volatility: The role of financial stress and asymmetric loss | JOURNAL OF INTERNATIONAL MONEY AND FINANCE | 2020 | Supervised | Fin | Markets & Predictions | 1 | N/A | Worldwide | Green |
| Gonon, L; Muhle-Karbe, J; Shi, XF | Asset pricing with general transaction costs: Theory and numerics | MATHEMATICAL FINANCE | 2021 | Supervised | Fin | Markets & Predictions | 1 | N/A | North America | Red |
| Gotze, T; Gurtler, M; Witowski, E | Improving CAT bond pricing models via machine learning | JOURNAL OF ASSET MANAGEMENT | 2020 | Supervised | Fin | Markets & Predictions | 4 | Random Forest | North America | Purple |
| Goudenege, L; Molent, A; Zanette, A | Machine learning for pricing American options in high-dimensional Markovian and non-Markovian models | QUANTITATIVE FINANCE | 2020 | Supervised | Fin | Markets & Predictions | 2 | | North America | Red |
| Gray, GL; Chiu, V; Liu, Q; Li, P | The expert systems life cycle in AIS research: What does it mean for future AIS research? | INTERNATIONAL JOURNAL OF ACCOUNTING INFORMATION SYSTEMS | 2014 | N/A | Acc | Literature, Corpus Analysis | 0 | N/A | Worldwide | Cyan |
| Gu, SH; Kelly, B; Xiu, DC | Empirical Asset Pricing via Machine Learning | REVIEW OF FINANCIAL STUDIES | 2020 | Supervised | Fin | Markets & Predictions | 6 | Neural Networks | North America | Purple |

| Authors | Article Title | Journal | Pub. Year | ML Cat. | Field | Grouping | # of ML Algo. | Best Algo. | Continent | Cluster |
|---|---|---|---|---|---|---|---|---|---|---|
| Gueant, O; Manziuk, I; Pu, J | Accelerated share repurchase and other buyback programs: what neural networks can bring | QUANTITATIVE FINANCE | 2020 | Reinforcement | Fin | Market Variables & Crisis Prediction | 1 | N/A | Europe | Red |
| Hayden, LX; Chachra, R; Alemi, AA; Ginsparg, PH; Sethna, JP | Canonical sectors and evolution of firms in the US stock markets | QUANTITATIVE FINANCE | 2018 | Unsupervised | Fin | Markets & Predictions | 1 | N/A | North America | Purple |
| Hoepner, AGF; McMillan, D; Vivian, A; Simen, CW | Significance, relevance and explainability in the machine learning age: an econometrics and financial data science perspective | EUROPEAN JOURNAL OF FINANCE | 2021 | N/A | Fin | ML Critisism | 0 | N/A | Worldwide | Purple |
| Holopainen, M; Sarlin, P | Toward robust early-warning models: a horse race, ensembles and model uncertainty | QUANTITATIVE FINANCE | 2017 | Supervised | Fin | Market Variables & Crisis Prediction | 12 | K-nearest-neighbors | Europe | Blue |
| Horvath, B; Muguruza, A; Tomas, M | Deep learning volatility: a deep neural network perspective on pricing and calibration in (rough) volatility models | QUANTITATIVE FINANCE | 2021 | Supervised | Fin | Algorithm\Methodology Creation | 1 | N/A | North America | Red |
| Hrazdil, K; Novak, J; Rogo, R; Wiedman, C; Zhang, R | Measuring executive personality using machine-learning algorithms: A new approach and audit fee-based validation tests | JOURNAL OF BUSINESS FINANCE & ACCOUNTING | 2020 | Unknown | Acc | Risk tolerance and Auditing Charges | 1 | N/A | North America | Yellow |
| Hsieh, TS; Kim, JB; Wang, RR; Wang, ZH | Seeing is believing? Executives' facial trustworthiness, auditor tenure, and audit fees | JOURNAL OF ACCOUNTING & ECONOMICS | 2020 | Supervised | Acc | Financial Statements Misreporting | 2 | Neural Networks | North America | Yellow |
| Hu, RM | Deep learning for ranking response surfaces with applications to optimal stopping problems | QUANTITATIVE FINANCE | 2020 | Supervised | Fin | Algorithm\Methodology Creation | 1 | N/A | Worldwide | Red |

| Authors | Article Title | Journal | Pub. Year | ML Cat. | Field | Grouping | # of ML Algo. | Best Algo. | Continent | Cluster |
|---|---|---|---|---|---|---|---|---|---|---|
| Hu, WB; Zastawniak, T | Pricing high-dimensional American options by kernel ridge regression | QUANTITATIVE FINANCE | 2020 | N/A | Fin | Markets & Predictions | 0 | N/A | North America | Red |
| Huang, DS; Li, JY; Wang, LY | Are disagreements agreeable? Evidence from information aggregation | JOURNAL OF FINANCIAL ECONOMICS | 2021 | Supervised | Fin | Markets & Predictions | 2 | ElasticNet | North America | Purple |
| Huang, FQ; Vasarhelyi, MA | Applying robotic process automation (RPA) in auditing: A framework | INTERNATIONAL JOURNAL OF ACCOUNTING INFORMATION SYSTEMS | 2019 | N/A | Acc | Other | 0 | N/A | Worldwide | Cyan |
| Jagtiani, J; Lemieux, C | The roles of alternative data and machine learning in fintech lending: Evidence from the LendingClub consumer platform | FINANCIAL MANAGEMENT | 2019 | N/A | Fin | Bankruptcy Prediction\ Default | 0 | N/A | North America | Yellow |
| Jang, H; Lee, J | Generative Bayesian neural network model for risk-neutral pricing of American index options | QUANTITATIVE FINANCE | 2019 | Supervised | Fin | Markets & Predictions | 4 | Generative Bayesian neural network | North America | Red |
| Jiang, Y; Jones, S | Corporate distress prediction in China: a machine learning approach | ACCOUNTING AND FINANCE | 2018 | Supervised | Acc | Bankruptcy Prediction\ Default | 1 | N/A | Asia | Blue |
| Johnson, EB; Tidwell, A; Villupuram, SV | Valuing Curb Appeal | JOURNAL OF REAL ESTATE FINANCE AND ECONOMICS | 2020 | Supervised | Fin | Real Estate | 2 | N/A | North America | Orange |

| Authors | Article Title | Journal | Pub. Year | ML Cat. | Field | Grouping | # of ML Algo. | Best Algo. | Continent | Cluster |
|---|---|---|---|---|---|---|---|---|---|---|
| Jones, S; Alam, N | A machine learning analysis of citation impact among selected Pacific Basin journals | ACCOUNTING AND FINANCE | 2019 | Supervised | Acc | Literature, Corpus Analysis | 1 | Gradient Boosting | Worldwide | Blue |
| Kakushadze, Z; Yu, W | Decoding stock market with quant alphas | JOURNAL OF ASSET MANAGEMENT | 2018 | Supervised | Fin | Algorithm\ Methodology Creation | 1 | N/A | North America | Green |
| Karolyi, GA; Van Nieuwerburgh, S | New Methods for the Cross-Section of Returns | REVIEW OF FINANCIAL STUDIES | 2020 | N/A | Fin | Literature, Corpus Analysis | 0 | N/A | N/A | Purple |
| Kercheval, AN; Zhang, Y | Modelling high-frequency limit order book dynamics with support vector machines | QUANTITATIVE FINANCE | 2015 | Supervised | Fin | Markets & Predictions | 1 | N/A | North America | Red |
| Khandani, AE; Kim, AJ; Lo, AW | Consumer credit-risk models via machine-learning algorithms | JOURNAL OF BANKING & FINANCE | 2010 | Supervised | Fin | Bankruptcy Prediction\ Default | 1 | N/A | North America | Blue |
| Kim, S | Enhancing the momentum strategy through deep regression | QUANTITATIVE FINANCE | 2019 | Supervised | Fin | Markets & Predictions | 2 | N/A | North America | Green |
| Kim, S; Kim, S | Index tracking through deep latent representation learning | QUANTITATIVE FINANCE | 2020 | Unsupervised | Fin | Markets & Predictions | 1 | N/A | Worldwide | Green |

| Authors | Article Title | Journal | Pub. Year | ML Cat. | Field | Grouping | # of ML Algo. | Best Algo. | Continent | Cluster |
|---|---|---|---|---|---|---|---|---|---|---|
| Knoll, J; Stubinger, J; Grottke, M | Exploiting social media with higher-order Factorization Machines: statistical arbitrage on high-frequency data of the S&P 500 | QUANTITATIVE FINANCE | 2019 | Supervised | Fin | Markets & Predictions | 2 | Factorization Machines | North America | Yellow |
| Kokina, J; Blanchette, S | Early evidence of digital labor in accounting: Innovation with Robotic Process Automation | INTERNATIONAL JOURNAL OF ACCOUNTING INFORMATION SYSTEMS | 2019 | N/A | Acc | Other | 0 | N/A | Worldwide | Cyan |
| Konstantinov, G; Chorus, A; Rebmann, J | A Network and Machine Learning Approach to Factor, Asset, and Blended Allocation | JOURNAL OF PORTFOLIO MANAGEMENT | 2020 | Supervised | Fin | Markets & Predictions | 1 | N/A | Worldwide | Green |
| Korhonen, T; Selos, E; Laine, T; Suomala, P | Exploring the programmability of management accounting work for increasing automation: an interventionist case study | ACCOUNTING AUDITING & ACCOUNTABILITY JOURNAL | 2021 | N/A | Acc | Other | 0 | N/A | Worldwide | Cyan |
| Kozak, S; Nagel, S; Santosh, S | Shrinking the cross-section | JOURNAL OF FINANCIAL ECONOMICS | 2020 | N/A | Fin | Algorithm\Methodology Creation | 0 | N/A | North America | Purple |
| Kumar, G; Muckley, CB; Pham, L; Ryan, D | Can alert models for fraud protect the elderly clients of a financial institution? | EUROPEAN JOURNAL OF FINANCE | 2019 | Supervised | Fin | Other | 3 | Random Forest | Europe | Blue |
| Lahmiri, S; Bekiros, S | Can machine learning approaches predict corporate bankruptcy? Evidence from a qualitative experimental design | QUANTITATIVE FINANCE | 2019 | Supervised | Fin | Bankruptcy Prediction\Default | 5 | GRNN | North America | Blue |
| Law, T; Shawe-Taylor, J | Practical Bayesian support vector regression for financial time series prediction and market condition change detection | QUANTITATIVE FINANCE | 2017 | Supervised | Fin | Algorithm\Methodology Creation | 1 | N/A | North America | Green |

| Authors | Article Title | Journal | Pub. Year | ML Cat. | Field | Grouping | # of ML Algo. | Best Algo. | Continent | Cluster |
|---|---|---|---|---|---|---|---|---|---|---|
| Le, HH; Viviani, JL | Predicting bank failure: An improvement by implementing a machine-learning approach to classical financial ratios | RESEARCH IN INTERNATIONAL BUSINESS AND FINANCE | 2018 | Supervised | Fin | Bankruptcy Prediction\Default | 4 | Neural Network | North America | Blue |
| Lee, SCK | ADDRESSING IMBALANCED INSURANCE DATA THROUGH ZERO-INFLATED POISSON REGRESSION WITH BOOSTING | ASTIN BULLETIN | 2021 | Supervised | Fin | Algorithm\Methodology Creation | 1 | N/A | N/A | Blue |
| Li, F | The Information Content of Forward-Looking Statements in Corporate Filings-A Naive Bayesian Machine Learning Approach | JOURNAL OF ACCOUNTING RESEARCH | 2010 | Supervised | Acc | Markets & Predictions | 1 | N/A | North America | Yellow |
| Li, H; Song, QF; Su, JX | Robust estimates of insurance misrepresentation through kernel quantile regression mixtures | JOURNAL OF RISK AND INSURANCE | 2021 | N/A | Fin | Insurance Predictions | 0 | N/A | North America | Red |
| Li, K; Mai, F; Shen, R; Yan, XY | Measuring Corporate Culture Using Machine Learning | REVIEW OF FINANCIAL STUDIES | 2021 | Unsupervised | Fin | Corporate | 1 | N/A | North America | Yellow |
| Li, TF; Chen, K; Feng, Y; Ying, ZL | Binary switch portfolio | QUANTITATIVE FINANCE | 2017 | Supervised | Fin | Markets & Predictions | 1 | N/A | North America | Green |
| Li, WP; Mei, F | Asset returns in deep learning methods: An empirical analysis on SSE 50 and CSI 300 | RESEARCH IN INTERNATIONAL BUSINESS AND FINANCE | 2020 | Supervised | Fin | Markets & Predictions | 4 | NONE | Asia | Green |
| Liang, J; Xu, Z; Li, P | Deep learning-based least squares forward-backward stochastic differential equation solver for high-dimensional derivative pricing | QUANTITATIVE FINANCE | 2021 | Supervised | Fin | Algorithm\Methodology Creation | 1 | N/A | Worldwide | Red |

| Authors | Article Title | Journal | Pub. Year | ML Cat. | Field | Grouping | # of ML Algo. | Best Algo. | Continent | Cluster |
|---|---|---|---|---|---|---|---|---|---|---|
| Liu, F; Pantelous, AA; von Mettenheim, HJ | Forecasting and trading high frequency volatility on large indices | QUANTITATIVE FINANCE | 2018 | Supervised | Fin | Markets & Predictions | 1 | N/A | North America | Green |
| Liu, H; Mulvey, J; Zhao, TQ | A semiparametric graphical modelling approach for large-scale equity selection | QUANTITATIVE FINANCE | 2016 | Supervised | Fin | Algorithm\Methodology Creation | 1 | N/A | North America | Green |
| Liu, MX; Li, GW; Li, JP; Zhu, XQ; Yao, YH | Forecasting the price of Bitcoin using deep learning | FINANCE RESEARCH LETTERS | 2021 | Supervised & Unsupervised | Fin | Markets & Predictions | 3 | Stacked Denoising Autoencoder | Worldwide | Green |
| Liu, QB; Wang, CJ; Zhang, P; Zheng, KX | Detecting stock market manipulation via machine learning: Evidence from China Securities Regulatory Commission punishment cases | INTERNATIONAL REVIEW OF FINANCIAL ANALYSIS | 2021 | Supervised | Fin | Markets & Predictions | 2 | Support Vector Machine | Asia | Blue |
| Loisel, S; Piette, P; Tsai, CHJ | APPLYING ECONOMIC MEASURES TO LAPSE RISK MANAGEMENT WITH MACHINE LEARNING APPROACHES | ASTIN BULLETIN | 2021 | Supervised | Fin | Insurance Predictions | 4 | Extreme Gradient Boosting | Asia | Blue |
| Makinen, Y; Kanniainen, J; Gabbouj, M; Iosifidis, A | Forecasting jump arrivals in stock prices: new attention-based network architecture using limit order book data | QUANTITATIVE FINANCE | 2019 | Supervised | Fin | Markets & Predictions | 4 | CNN-LSTM-Attention | North America | Red |
| Manahov, V; Urquhart, A | The efficiency of Bitcoin: A strongly typed genetic programming approach to smart electronic Bitcoin markets | INTERNATIONAL REVIEW OF FINANCIAL ANALYSIS | 2021 | Supervised | Fin | Markets & Predictions | 1 | N/A | Worldwide | Green |
| Manela, A; Moreira, A | News implied volatility and disaster concerns | JOURNAL OF FINANCIAL ECONOMICS | 2017 | Supervised | Fin | Markets & Predictions | 1 | N/A | North America | Yellow |

| Authors | Article Title | Journal | Pub. Year | ML Cat. | Field | Grouping | # of ML Algo. | Best Algo. | Continent | Cluster |
|---|---|---|---|---|---|---|---|---|---|---|
| Marrone, M; Linnenluecke, MK; Richardson, G; Smith, T | Trends in environmental accounting research within and outside of the accounting discipline | ACCOUNTING AUDITING & ACCOUNTABILITY JOURNAL | 2020 | Unsupervised | Acc | Literature, Corpus Analysis | 1 | N/A | Worldwide | Yellow |
| Martinez, RG; Roman, MP; Casado, PP | Big Data Algorithmic Trading Systems Based on Investors' Mood | JOURNAL OF BEHAVIORAL FINANCE | 2019 | Supervised | Fin | Markets & Predictions | 1 | N/A | Europe | Green |
| McInish, TH; Nikolsko-Rzhevska, O; Nikolsko-Rzhevskyy, A; Panovska, I | Fast and slow cancellations and trader behavior | FINANCIAL MANAGEMENT | 2020 | Supervised | Fin | Markets & Predictions | 1 | N/A | North America | Red |
| M'ng, JCP; Jer, HY | Do economic statistics contain information to predict stock indexes futures prices and returns? Evidence from Asian equity futures markets | REVIEW OF QUANTITATIVE FINANCE AND ACCOUNTING | 2021 | Supervised | Fin | Markets & Predictions | 2 | N/A | Asia | Green |
| Nian, K; Coleman, TF; Li, YY | Learning minimum variance discrete hedging directly from the market | QUANTITATIVE FINANCE | 2018 | N/A | Fin | Algorithm\ Methodolog y Creation | 0 | N/A | North America | Red |
| Nian, K; Coleman, TF; Li, YY | Learning sequential option hedging models from market data | JOURNAL OF BANKING & FINANCE | 2021 | Supervised | Fin | Markets & Predictions | 1 | N/A | North America | Red |

| Authors | Article Title | Journal | Pub. Year | ML Cat. | Field | Grouping | # of ML Algo. | Best Algo. | Continent | Cluster |
|---|---|---|---|---|---|---|---|---|---|---|
| Novak, MG; Veluscek, D | Prediction of stock price movement based on daily high prices | QUANTITATIVE FINANCE | 2016 | Supervised & Unsupervised | Fin | Markets & Predictions | 3 | SVM | North America | Green |
| Nowak, AD; Price, BS; Smith, PS | Real Estate Dictionaries Across Space and Time | JOURNAL OF REAL ESTATE FINANCE AND ECONOMICS | 2021 | Supervised | Fin | Real Estate | 1 | N/A | North America | Yellow |
| Olson, LM; Qi, M; Zhang, XF; Zhao, XL | Machine learning loss given default for corporate debt | JOURNAL OF EMPIRICAL FINANCE | 2021 | Supervised | Fin | Bankruptcy Prediction\Default | 6 | K-nearest-neighbors | North America | Blue |
| Oz, IO; Yelkenci, T; Meral, G | The role of earnings components and machine learning on the revelation of deteriorating firm performance | INTERNATIONAL REVIEW OF FINANCIAL ANALYSIS | 2021 | Supervised | Fin | Markets & Predictions | 7 | Neural Networks | Worldwide | Blue |
| Pace, RK; Hayunga, D | Examining the Information Content of Residuals from Hedonic and Spatial Models Using Trees and Forests | JOURNAL OF REAL ESTATE FINANCE AND ECONOMICS | 2020 | Supervised | Fin | Real Estate | 4 | Bagging | North America | Purple |
| Perols, J | Financial Statement Fraud Detection: An Analysis of Statistical and Machine Learning Algorithms | AUDITING-A JOURNAL OF PRACTICE & THEORY | 2011 | Supervised | Acc | Financial Statements Misreporting | 6 | Support Vector Machine | North America | Blue |
| Pierdzioch, C; Risse, M | A machine-learning analysis of the rationality of aggregate stock market forecasts | INTERNATIONAL JOURNAL OF FINANCE & ECONOMICS | 2018 | Supervised | Fin | Markets & Predictions | 1 | N/A | North America | Blue |
| Plakandaras, V; Gupta, R; Wohar, ME | The depreciation of the pound post-Brexit: Could it have been predicted? | FINANCE RESEARCH LETTERS | 2017 | Supervised | Fin | Markets & Predictions | 1 | N/A | Europe | Green |

| Authors | Article Title | Journal | Pub. Year | ML Cat. | Field | Grouping | # of ML Algo. | Best Algo. | Continent | Cluster |
|---|---|---|---|---|---|---|---|---|---|---|
| Poloz, SS | Technological progress and monetary policy: Managing the fourth industrial revolution | JOURNAL OF INTERNATIONAL MONEY AND FINANCE | 2021 | N/A | Fin | Market Variables & Crisis Prediction | 0 | N/A | North America | Cyan |
| Polyzos, S; Samitas, A; Katsaiti, MS | Who is unhappy for Brexit? A machine-learning, agent-based study on financial instability | INTERNATIONAL REVIEW OF FINANCIAL ANALYSIS | 2020 | Supervised | Fin | Market Variables & Crisis Prediction | 2 | N/A | Europe | Red |
| Pun, CS | A Sparse Learning Approach to Relative-Volatility-Managed Portfolio Selection | SIAM JOURNAL ON FINANCIAL MATHEMATICS | 2021 | Supervised | Fin | Algorithm\Methodology Creation | 1 | N/A | North America | Green |
| Pun, CS; Wang, L | A cost-effective approach to portfolio construction with range-based risk measures | QUANTITATIVE FINANCE | 2021 | Supervised | Fin | Markets & Predictions | 3 | N/A | North America | Green |
| Pyo, S; Lee, J | Exploiting the low-risk anomaly using machine learning to enhance the Black-Litterman framework: Evidence from South Korea | PACIFIC-BASIN FINANCE JOURNAL | 2018 | Supervised | Fin | Markets & Predictions | 3 | Neural Network | Asia | Green |
| Rasekhschaffe, KC; Jones, RC | Machine Learning for Stock Selection | FINANCIAL ANALYSTS JOURNAL | 2019 | Supervised | Fin | Markets & Predictions | 4 | Gradient Boosting | Worldwide | Purple |
| Renault, T | Intraday online investor sentiment and return patterns in the US stock market | JOURNAL OF BANKING & FINANCE | 2017 | Supervised | Fin | Markets & Predictions | 2 | N/A | North America | Yellow |
| Risk, J; Ludkovski, M | Sequential Design and Spatial Modeling for Portfolio Tail Risk Measurement | SIAM JOURNAL ON FINANCIAL MATHEMATICS | 2018 | Supervised | Fin | Algorithm\Methodology Creation | 1 | N/A | Unknown | Red |

| Authors | Article Title | Journal | Pub. Year | ML Cat. | Field | Grouping | # of ML Algo. | Best Algo. | Continent | Cluster |
|---|---|---|---|---|---|---|---|---|---|---|
| Risse, M; Ohl, L | Using dynamic model averaging in state space representation with dynamic Occam's window and applications to the stock and gold market | JOURNAL OF EMPIRICAL FINANCE | 2017 | Supervised | Fin | Markets & Predictions | 2 | N/A | North America | Purple |
| Routledge, BR | Machine learning and asset allocation | FINANCIAL MANAGEMENT | 2019 | Supervised | Fin | Markets & Predictions | 1 | N/A | North America | Purple |
| Rybinski, K | The forecasting power of the multi-language narrative of sell-side research: A machine learning evaluation | FINANCE RESEARCH LETTERS | 2020 | Supervised | Fin | Market Variables & Crisis Prediction | 2 | VAR | Europe | Yellow |
| Samitas, A; Kampouris, E; Kenourgios, D | Machine learning as an early warning system to predict financial crisis | INTERNATIONAL REVIEW OF FINANCIAL ANALYSIS | 2020 | Supervised | Fin | Market Variables & Crisis Prediction | 5 | Support Vector Machine | Worldwide | Purple |
| Sermpinis, G; Laws, J; Dunis, CL | Modelling commodity value at risk with Psi Sigma neural networks using open-high-low-close data | EUROPEAN JOURNAL OF FINANCE | 2015 | Supervised | Fin | Markets & Predictions | 2 | PSI Neural Network | Worldwide | Green |
| Sirignano, J; Cont, R | Universal features of price formation in financial markets: perspectives from deep learning | QUANTITATIVE FINANCE | 2019 | Supervised | Fin | Markets & Predictions | 1 | N/A | North America | Red |
| Sirignano, J; Spiliopoulos, K | Stochastic Gradient Descent in Continuous Time | SIAM JOURNAL ON FINANCIAL MATHEMATICS | 2017 | Reinforcement | Fin | Algorithm\ Methodolog y Creation | 1 | N/A | North America | Red |
| Sirignano, JA | Deep learning for limit order books | QUANTITATIVE FINANCE | 2019 | Supervised | Fin | Algorithm\ Methodolog y Creation | 2 | Neural Networks | North America | Red |

| Authors | Article Title | Journal | Pub. Year | ML Cat. | Field | Grouping | # of ML Algo. | Best Algo. | Continent | Cluster |
|---|---|---|---|---|---|---|---|---|---|---|
| Slapnik, U; Loncarski, I | On the information content of sovereign credit rating reports: Improving the predictability of rating transitions | JOURNAL OF INTERNATIONAL FINANCIAL MARKETS INSTITUTIONS & MONEY | 2021 | Supervised | Fin | Market Variables & Crisis Prediction | 2 | N/A | Worldwide | Yellow |
| Soto, PE | Breaking the Word Bank: Measurement and Effects of Bank Level Uncertainty | JOURNAL OF FINANCIAL SERVICES RESEARCH | 2021 | Supervised & Unsupervised | Fin | Market Variables & Crisis Prediction | 3 | N/A | North America | Yellow |
| Sun, T | Applying Deep Learning to Audit Procedures: An Illustrative Framework | ACCOUNTING HORIZONS | 2019 | N/A | Acc | Other | 1 | N/A | N/A | Cyan |
| Sutton, SG; Holt, M; Arnold, V | The reports of my death are greatly exaggerated-Artificial intelligence research in accounting | INTERNATIONAL JOURNAL OF ACCOUNTING INFORMATION SYSTEMS | 2016 | N/A | Acc | Literature, Corpus Analysis | 0 | N/A | Worldwide | Cyan |
| Tantri, P | Fintech for the Poor: Financial Intermediation Without Discrimination | REVIEW OF FINANCE | 2021 | Supervised | Fin | Bankruptcy Prediction\Default | 4 | Gradient Boosting | Asia | Yellow |
| Tashiro, D; Matsushima, H; Izumi, K; Sakaji, H | Encoding of high-frequency order information and prediction of short-term stock price by deep learning | QUANTITATIVE FINANCE | 2019 | Supervised | Fin | Algorithm\Methodology Creation | 3 | Convolutional Neural Network | Asia | Red |
| Tsang, KH; Wong, HY | Deep-Learning Solution to Portfolio Selection with Serially Dependent Returns | SIAM JOURNAL ON FINANCIAL MATHEMATICS | 2020 | Supervised | Fin | Markets & Predictions | 1 | N/A | North America | Red |
| Varetto, F | Genetic algorithms applications in the analysis of insolvency risk | JOURNAL OF BANKING & FINANCE | 1998 | Supervised | Fin | Bankruptcy Prediction\Default | 2 | LDA | Europe | Green |

| Authors | Article Title | Journal | Pub. Year | ML Cat. | Field | Grouping | # of ML Algo. | Best Algo. | Continent | Cluster |
|---|---|---|---|---|---|---|---|---|---|---|
| Viebig, J | Exuberance in Financial Markets: Evidence from Machine Learning Algorithms | JOURNAL OF BEHAVIORAL FINANCE | 2020 | Supervised | Fin | Markets & Predictions | 1 | N/A | Worldwide | Blue |
| Viswanathan, PK; Srinivasan, S; Hariharan, N | Predicting Financial Health of Banks for Investor Guidance Using Machine Learning Algorithms | JOURNAL OF EMERGING MARKET FINANCE | 2020 | Supervised | Fin | Bankruptcy Prediction\Default | 4 | Random Forest | Asia | Blue |
| Vrontos, SD; Galakis, J; Vrontos, ID | Implied volatility directional forecasting: a machine learning approach | QUANTITATIVE FINANCE | 2021 | Supervised | Fin | Markets & Predictions | 10 | Naive Bayes | North America | Blue |
| Wall, LD | Some financial regulatory implications of artificial intelligence | JOURNAL OF ECONOMICS AND BUSINESS | 2018 | N/A | Fin | Other | 0 | N/A | N/A | Cyan |
| Wang, C; Zhang, Y; Zhang, WG; Gong, X | Textual sentiment of comments and collapse of P2P platforms: Evidence from China's P2P market | RESEARCH IN INTERNATIONAL BUSINESS AND FINANCE | 2021 | Supervised | Fin | Bankruptcy Prediction\Default | 2 | N/A | Asia | Yellow |
| Wang, CW; Zhang, JG; Zhu, WJ | NEIGHBOURING PREDICTION FOR MORTALITY | ASTIN BULLETIN | 2021 | Supervised | Fin | Insurance Predictions | 2 | Convolutional Neural Network | Worldwide | Cyan |
| Wang, HR; Zhou, XY | Continuous-time mean-variance portfolio selection: A reinforcement learning framework | MATHEMATICAL FINANCE | 2020 | Reinforcement | Fin | Algorithm\Methodology Creation | 1 | N/A | North America | Green |
| Weigand, A | Machine learning in empirical asset pricing | FINANCIAL MARKETS AND PORTFOLIO MANAGEMENT | 2019 | N/A | Fin | Literature, Corpus Analysis | 0 | N/A | N/A | Purple |

| Authors | Article Title | Journal | Pub. Year | ML Cat. | Field | Grouping | # of ML Algo. | Best Algo. | Continent | Cluster |
|---|---|---|---|---|---|---|---|---|---|---|
| Wiese, M; Knobloch, R; Korn, R; Kretschmer, P | Quant GANs: deep generation of financial time series | QUANTITATIVE FINANCE | 2020 | Supervised | Fin | Algorithm\ Methodology Creation | 1 | N/A | North America | Red |
| Wolff, D; Neugebauer, U | Tree-based machine learning approaches for equity market predictions | JOURNAL OF ASSET MANAGEMENT | 2019 | Supervised | Fin | Markets & Predictions | 9 | ElasticNet Regression | North America | Purple |
| Wu, PL; Siwasarit, W | Capturing the Order Imbalance with Hidden Markov Model: A Case of SET50 and KOSPI50 | ASIA-PACIFIC FINANCIAL MARKETS | 2020 | Unsupervised | Fin | Markets & Predictions | 1 | N/A | Asia | Green |
| Wu, WP; Lee, VCS; Tan, TY | Data preprocessing and data parsimony in corporate failure forecast models: evidence from Australian materials industry | ACCOUNTING AND FINANCE | 2006 | Supervised | Acc | Bankruptcy Prediction\ Default | 4 | Neural Networks | Oceania | Blue |
| Xue, FJ; Li, XY; Zhang, T; Hu, N | Stock market reactions to the COVID-19 pandemic: The moderating role of corporate big data strategies based on Word2Vec* | PACIFIC-BASIN FINANCE JOURNAL | 2021 | Supervised & Unsupervised | Fin | Market Variables & Crisis Prediction | 1 | N/A | Asia | Cyan |
| Yang, SY; Qiao, QF; Beling, PA; Scherer, WT; Kirilenko, AA | Gaussian process-based algorithmic trading strategy identification | QUANTITATIVE FINANCE | 2015 | Reinforcement | Fin | Algorithm\ Methodology Creation | 3 | GPIRL | North America | Red |
| Zaremba, A; Kizys, R; Tzouvanas, P; Aharon, DY; Demir, E | The quest for multidimensional financial immunity to the COVID-19 pandemic: Evidence from international stock markets | JOURNAL OF INTERNATIONAL FINANCIAL MARKETS INSTITUTIONS & MONEY | 2021 | Supervised | Fin | Market Variables & Crisis Prediction | 1 | N/A | Worldwide | Cyan |
| Zhang, JH; Huang, WJ | Option hedging using LSTM-RNN: an empirical analysis | QUANTITATIVE FINANCE | 2021 | Reinforcement | Fin | Markets & Predictions | 1 | N/A | Worldwide | Red |

| Authors | Article Title | Journal | Pub. Year | ML Cat. | Field | Grouping | # of ML Algo. | Best Algo. | Continent | Cluster |
|---|---|---|---|---|---|---|---|---|---|---|
| Zhang, YJ; Chu, G; Shen, DH | The role of investor attention in predicting stock prices: The long short-term memory networks perspective | FINANCE RESEARCH LETTERS | 2021 | Supervised | Fin | Markets & Predictions | 1 | N/A | Asia | Green |

Author's Statement:

I hereby expressly declare that, according to the article 8 of Law 1559/1986, this dissertation is solely the product of my personal work, does not infringe any intellectual property, personality and personal data rights of third parties, does not contain works/contributions from third parties for which the permission of the authors/beneficiaries is required, is not the product of partial or total plagiarism, and that the sources used are limited to the literature references alone and meet the rules of scientific citations.