

Examensarbete
15 högskolepoäng, grundnivå

Litteraturstudie: Tillämpningen av maskininlärning vid algoritmisk handel

Literature review: The application of machine
learning in algorithmic trading

Karl Paradis

Therése Larsson

Examen: Kandidatexamen 180 hp
Huvudområde: Datavetenskap
Program: Systemutvecklare
Datum för slutseminarium: 2019-05-31

Handledare: Agnes Tegen
Examinator: Magnus Johnsson

Abstract

We conduct a literature review in which we study and analyze publications in the area of machine learning in combination with algorithmic trading. In this study we investigate what types of data and which machine learning techniques that are shown to be applicable to systems used for algorithmic trading. For our literature review we use peer-reviewed publications from trustworthy databases.

The result shows that we find mainly three types of data that are relevant for algorithmic trading. These are financial data quotes, technical indicators and the types of data that is relevant for fundamental analysis. Financial data quotes often seem to be used as a basis for later processing into other types of data. The most common example of this is technical indicators that are frequently used as a source of data in systems for algorithmic trading.

We also find a number of machine learning techniques that have been demonstrated by previous publications to be applicable for algorithmic trading. Publications show that a machine learning technique called SVM (support vector machine) can be applied on technical indicators as well as for analysis of news headlines. We also find publications that demonstrate the application of two types of neural networks, classification and regression network. These are used in order to generate trade signals in an algorithmic trading system. In our study we also find an application of evolutionary machine learning which is used to approximate an optimal solution to the order execution problem.

Moreover, we also discuss a financial incentive that disadvantage academic openness and the publications of new discoveries in the relevant area of research. This financial incentive exists because advantageous results may be financially beneficial to withhold.

Keywords:

machine learning, algorithmic trading, literature review, technical indicators, order execution problem, support vector machine, financial data quotes, efficient-market hypothesis, neural networks, VaR.

Sammanfattning

Vi genomför en litteraturstudie där vi studerar och analyserar publikationer inom maskininlärning i kombination med algoritmisk handel. I denna studie undersöker vi vilka typer av data samt vilka maskininlärningstekniker som kunnat visas vara tillämpningsbara vid system för algoritmisk handel. Till vår litteraturstudie använder vi oss av publikationer som är peer-reviewed från trovärdiga databaser.

Resultatet visar att det huvudsakligen finns tre typer av data som är av betydelse för algoritmisk handel. Dessa är historisk prisdata, tekniska indikatorer samt den typ av data som ingår i fundamental analys. Historisk prisdata tycks ofta användas som bas för att sedan bearbetas om till andra typer av data. Det vanligaste exemplet på detta är tekniska indikatorer som ofta förekommer som datakälla i system för algoritmisk handel.

Vi finner även ett antal maskininlärningstekniker som av tidigare publikationer demonstreras vara tillämpningsbara för algoritmisk handel. Publikationer påvisar att en maskininlärningsteknik kallad SVM (support vector machine) kan tillämpas på tekniska indikatorer och även analys av nyhetsrubriker. Vi påträffar även publikationer som demonstrerar tillämpningen av två typer av neurala nätverk, klassifikationsnätverk samt regressionsnätverk. Dessa nyttjas för att generera trade signals i ett algoritmiskt handelssystem. I vår studie hittar vi också en tillämpning av evolutionär maskininlärning som används för att approximera en lösning på det optimala orderexekveringsproblemet.

Vi diskuterar även ett ekonomiskt incitament som missgynnar akademisk öppenhet och publikation av nya upptäckter inom området. Detta existerar på grund av att fördelaktiga resultat kan vara finansiellt gynnsamma att undanhålla.

Nyckelord:

maskininlärning, algoritmisk handel, litteraturstudie, tekniska indikatorer, orderexekveringsproblem, support vector machine, historisk prisdata, efficient-market hypotesen, neurala nätverk, VaR.

Innehållsförteckning

1 Introduktion	1
1.1 Bakgrund	1
1.1.1 Vad är maskininlärning?	1
1.1.2 Olika typer av maskininlärning	2
1.1.3 Hur maskininlärning fungerar	2
1.1.4 Träningsdata	3
1.1.5 Support vector machine	4
1.1.6 Neurala nätverk	5
1.1.7 Finansmarknaden	7
1.1.8 Arbitrage	7
1.1.9 Vad är algoritmisk handel?	8
1.2 Problemformulering	9
1.3 Syfte och frågeställning	9
1.4 Avgränsningar	9
2 Metod	11
2.1 Metodbeskrivning	11
2.1.1 Informationssökningsprocessen	11
2.1.2 Urval av referenser: inklusions- och exklusionskriterier	11
2.1.3 Exempel på sökningar och dess resultat	12
2.2 Metoddiskussion	14
3 Resultat	15
3.1 Relevanta typer av data vid algoritmisk handel	15
3.1.1 Historisk prisdata	15
3.1.2 Tekniska indikatorer	15
3.1.3 Andra typer av data	16
3.2 Maskininlärningstekniker lämpliga vid algoritmisk handel	17
3.2.1 SVM tillämpat på tekniska indikatorer och textbaserad analys	17
3.2.2 Classification and regression networks	18
3.2.3 Evolutionär maskininlärning tillämpat på det optimala orderexekveringsproblemet	20
4 Analys och diskussion	22
4.1 Diskussion om data i relation till efficient-market hypotesen	22
4.2 Undanhållande av forskningsframsteg	23
4.3 Optimal orderexekvering och arbitrage	23
5 Slutsatser	25
5.1 Framtida studier	25
Referenser	26

1 Introduktion

1.1 Bakgrund

Att systematiskt bearbeta enormt stora mängder data är en utmanande och tidskrävande uppgift för människor, maskininlärning har dock en förmåga att snabbt bearbeta och se mönster i väldigt stora datamängder. Algoritmisk handel, även kallat AT för algorithmic trading, beräknas utgöra ca 50-60% av all aktiehandel i USA och Europa [36]. Förutsatt att det finns tillräckligt mycket data att tillgå för en maskininlärningsteknik att träna upp sig på vore det intressant att undersöka vilka former av maskininlärning som kan appliceras på algoritmisk handel.

Det här inledande avsnittet ger en bakgrund till de tekniker och begrepp som är centrala för detta arbete. Fokus kommer ligga på att beskriva vad maskininlärning¹ och algoritmisk handel är samt hur dessa relaterar till varandra.

1.1.1 Vad är maskininlärning?

Maskininlärning utgör en viktig del i det som kallas för artificiell intelligens, som i sig är ett mycket stort område som kan delas upp i flera delområden. Precis som människor kan lära sig hantera och anpassa sig till nya situationer, så kan datorer göra detsamma med hjälp av maskininlärning. Detta kan uppnås genom att låta datorprogrammet utgå från en mängd data och analysera denna för att försöka förutsäga sannolika framtida scenarion, för att på så vis vägleda datorns beslutstagande i en given situation. Målet är att försöka identifiera mönster, vilket inte är olik hur vi människor lär oss, för att få ett datorprogram att utföra en viss given uppgift på ett effektivt och träffsäkert² sätt. Ett något förenklat sätt att definiera maskininlärning, enligt Marsland [4], vore att låta ett datorprogram bli bättre på att utföra en uppgift genom att öva. Ett datorprogram som agerar med ett syfte, tillsammans med egenskaper som innefattar att kunna anpassa sig till förändringar, fungera autonomt³ och uppfatta sin miljö, brukar kallas för en agent [2].

Maskininlärning, och även artificiell intelligens överlag, har fått mycket stor uppmärksamhet de senaste decennierna. Detta är inte bara begränsat till det datavetenskapliga området utan även många andra fält har uppvisat intresse för och använt maskininlärning [4] [5]. Maskininlärning har exempelvis använts inom den gentekniska forskningen för att skapa en bättre förståelse för det mänskliga genomet [6] och har även nyttjats för att modellera mänsklig aktivitet i olika sammanhang [7]. Maskininlärning har även använts för att förutse Turkiets koldioxidutsläpp utifrån faktorer som t.ex. tid, bränslekonsumtion och befolkningsmängd [15].

¹ eng. *machine learning*

² eng. *accuracy*

³ Autonoma datorprogram förväntas kunna agera självständigt och oberoende av mänsklig övervakning eller styrning.

1.1.2 Olika typer av maskininlärning

Som tidigare nämnt så innebär maskininlärning att låta ett datorprogram bli bättre på att lösa ett problem genom att öva. Men hur vet datorprogrammet att det faktiskt blir bättre på att utföra en uppgift? Om datorprogrammet behöver förbättra sin lösning, hur ska det veta hur detta uppnås? [4]. Svaren på dessa frågor genererar olika typer av maskininlärning som beskrivs nedan⁴.

Supervised learning: Detta är den vanligaste typen av maskininlärning. Algoritmen tränas på en mängd annoterad data, utifrån vilken den lär sig att generalisera, för att kunna ge korrekt respons för alla möjliga input-värden. Man kan se det som att algoritmen lär sig från exempel [4].

Unsupervised learning: Vid vissa tillfällen så kanske det nödvändigtvis inte finns någon annoterad data tillgänglig. I sådana fall är målet för algoritmen att helt enkelt hitta och gruppera input som liknar varandra [4].

Reinforcement learning: Denna typ är som en blandning mellan supervised och unsupervised learning. Algoritmen blir tillsagd när lösningen är felaktig, men får ingen vägledning om hur den felaktiga lösningen ska åtgärdas för att den ska bli korrekt. Algoritmen får helt enkelt pröva sig fram tills den når en korrekt lösning [4].

Evolutionary learning: Den här typen är inspirerad av den inlärningsprocess som kan observeras hos biologisk evolution där organismer anpassar sig till sin miljö i syftet att öka sin chans till överlevnad och att få avkomma [4]. I samband med evolution myntas talesättet "survival of the fittest" på sv. den mest välanpassade överlever. Marsland [4] menar att "fitness" korresponderar till hur bra en nuvarande lösning är. Den här typen av maskininlärning försöker då förbättra sin lösning genom att välja den bästa lösningen.

1.1.3 Hur maskininlärning fungerar

Algoritmer som tillämpar maskininlärning har förmågan att generalisera utifrån givna exempel, vilket är något som inte är genomförbart eller kostnadseffektivt för manuellt programmerade datorprogram. Ju mer data som blir tillgänglig, desto mer ambitiösa problem kan algoritmen lösa. Processen för maskininlärning brukar följa ett visst mönster, varav Marsland [4] är en av de som redogör för hur denna process går till. Han beskriver processen i sex steg⁵, som förklaras nedan.

Data collection and preparation: Först måste lämplig data samlas in, eller genereras, som datorprogrammet kommer att lära sig från. Mängden data måste övervägas. Det måste finnas tillräcklig träningsdata för att en generalisering ska kunna ske, samtidigt som för mycket data ökar beräkningskostnaderna [4].

Därefter måste träningsdatan förberedas på ett sådant sätt att den går att använda för datorprogrammet. Brus, inkomplett data som saknar viktiga attribut, inkonsekvent data och felaktig data måste kasseras för att erhålla ren och

⁴ Begreppen för de olika typer av maskininlärning benämns med de engelska termer som författaren använder, då vissa av termerna kan vara svåra att översätta till svenska på ett adekvat sätt

⁵ Processens steg benämns med de engelska termer som författaren använder, då vissa av termerna kan vara svåra att översätta till svenska på ett adekvat sätt

träffsäker träningsdata [4]. Datan måste även vara representerad på ett sådant sätt att datorprogrammet kan förstå och använda datan.

Feature selection: Nästa steg är att identifiera och samla in de egenskaper⁶ som är mest användbara för det givna problemet som maskininlärningsalgoritmen är avsedd att lösa. Det är nödvändigt att identifieringen och insamlingen inte konsumerar betydande tid eller kostnad. Utöver detta är det också av betydelse att identifieringen och insamlingen är robust på ett sådant sätt att brus och korruption av data undviks [4].

Algorithm choice: Härnäst måste en lämplig algoritm utses med hänsyn till träningsdatan. Urvalet av algoritmer är mycket stort och valet beror även på vad algoritmen är avsedd att uppnå [4].

Parameter and model selection: Härdanefters måste de relevanta parametrarna hos algoritmen väljas. Även en inlärningsmodell måste väljas, vars syfte är att avgöra när det är lämpligt för algoritmen att avsluta inlärningen [4].

Training: Härnäst används de tillgängliga beräkningsresurserna för att bygga upp en modell av träningsdatan, som används som inputdata, i syftet att kunna förutse output av ny indata [4]. Det är i det här steget som själva lärandet äger rum.

Evaluation: Det sista steget i processen går ut på att testa och därefter analysera och utvärdera hur träffsäkert systemet är för den data som ej inkluderats i träningsdatan. Många gånger är det behövt och fördelaktigt att utföra jämförelser med mänskliga experter inom det relevanta ämnesområdet. Ofta är det även nödvändigt att välja lämpliga mätvärden för jämförelsen som ska göras [4].

1.1.4 Träningsdata

Algoritmen som tillämpar maskininlärning måste som tidigare nämnts ha data som den lär sig av, det vill säga träningsdata. Koschnicke et al. [10] syftar även på betydelsen att datan är komplett och korrekt, då författarna menar att låg kvalitet på datan resulterar i felaktiga och svaga resultat från algoritmen som ska användas vid algoritmisk handel. Koschnicke et al. [10] menar att dessa felaktiga resultat i sin tur orsakar omfattande finansiella förluster.

Exakt vilken typ av träningsdata som behövs beror på sammanhanget. Om det handlar om s.k. *live trading* behövs s.k. *online data*, medan s.k. *backtesting* nyttjar sig av s.k. *offline data* [10]. Live trading innebär att handelsstrategin sker i realtid och backtesting går ut på att man simulerar en handelsstrategi genom att använda historisk data. Införskaffandet av offline och/eller live data kan ske på olika sätt. Koschnicke et al. [10] hävdar att det existerar många företag som säljer offline data som består av historisk data, av varierande kvalitet. Datan är i form av textfiler som måste analyseras och importeras innan de kan användas. Författarna menar även att live data kan erhållas från en rad olika webbsidor, och att dessa webbsidor även erbjuder möjligheten att införskaffa historisk data. Därav är det även viktigt att notera vem som är källan till den använda datan, då datan kan bero på utgivaren [10].

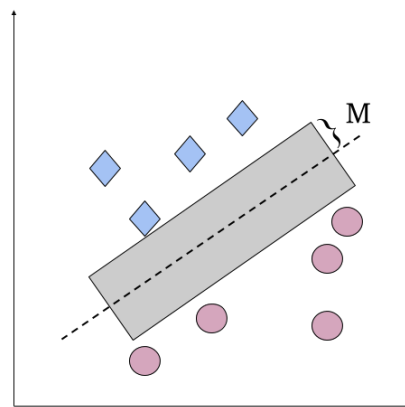
⁶ eng. *features*

Oavsett om offline eller live data används, så är datans tidsram en annan viktig aspekt, menar Koschnicke et al. [10]. Författarna förklarar att det är viktigt att veta under vilket tidsintervall som datan samlats in och att även ta reda på den korrekta tidszonen för datapunkterna. Noterandet av korrekt tidszon är särskilt betydelsefullt för data erhållen från webbsidor, då tiderna ofta är angivna utifrån en viss webbsidas lokala tidszon. Även Shynkevich et al. [9] beskriver tid som en viktig typ av data då tekniska analyser använder historiska tidsserier av data i syftet att försöka förutse framtida priser och trender inom den finansiella marknaden.

1.1.5 Support vector machine

Support vector machine (SVM) är en supervised maskininlärningsteknik som också räknas som en klassificeringsteknik som kartlägger datapunkter och kategoriserar dessa i en av två kategorier. SVM introducerades på tidigt 90-tal och är idag väldigt väletablerad då den är en mycket effektiv metod för att klassificera data på relativt stora dataset [4] [16]. SVM presterar dock inte så bra på extremt stora datamängder. Detta p.g.a. att beräkningarna som utförs på datamängden inte skalar bra med antalet träningsexemplar och således blir beräkningarna för kostsamma [4]. Det finns två termer som är viktiga att känna till för att förstå SVM: marginal och support vektor, vilket förklaras nedan.

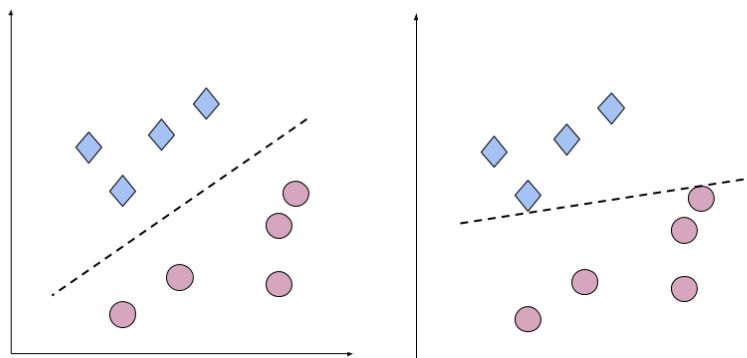
Figur 1 illustrerar hur en mängd datapunkter kan klassificeras. Diamanterna och cirklarna representerar datapunkter och den streckade linjen representerar klassificeringen av dessa datapunkter. Marsland [4] menar att det är optimalt om linjen kan dras linjärt för att skilja datapunkterna. De datapunkter som ligger närmast linjen i vardera klassificering kallas för support vektorer. Marginalen (betecknat M i figuren) är det avstånd som vi kan illustrera som den gråa zonen mellan separationslinjen och datapunkterna i figuren nedan. Marginalen tas fram genom att dra en så lång rätvinklig linje som möjligt från den streckade linjen tills att vi träffar en datapunkt.



Figur 1. Illustration av hur datapunkter (representeras som diamanter och cirklar) kan klassificeras (representeras som streckad linje) med hjälp av SVM.

Marsland [4] menar att SVM bör sträva efter att separationen mellan de två klassificeringarna ska vara så stor och tydlig som möjligt, dvs. att ha en godtyckligt stor marginal. Avsikten är att så få datapunkter som möjligt ska kunna tolkas som gränsfall mellan att tillhöra en av de två kategorierna. Om en datapunkt är ett

gränsfall kan den klassificeras felaktigt och algoritmen blir i sådana fall mindre träffsäker [4]. Figur 2 illustrerar två exempel på klassificering av identiska datapunkter. Den optimala klassificeringen återfinns i exemplet till vänster i figuren, där separationslinjen har ett godtyckligt avstånd från datapunkterna. I exemplet till höger befinner sig linjen mycket nära två datapunkter, då dessa skulle kunna tolkas som gränsfall av algoritmen. På grund av detta menar Marsland [4] att support vektorerna är de mest användbara och viktigaste datapunkterna för SVM. Efter att träningen av SVM-algoritmen är genomförd behövs endast support vektorerna för klassificeringen. Således kan alla datapunkter förutom support vektorerna kasseras efter träningen vilket gör SVM till en utrymmeseffektiv algoritm [4].



Figur 2. Illustration av olika optimala klassificeringar.

SVM användningsområden är stora och har sedan dess introducering använts mest inom mönsterigenkänning. Exempel på användningsområden där SVM har använts är anomalitetsdetektering⁷ [19], klassificering av gener [22], bioteknik [28], ansikts- och röstigenkänning samt textkategorisering [16]. En studie gjord av Lee och Tong [23] år 2016 nyttjade SVM för att reducera energiförbrukningen i flera olika hushåll genom att låta SVM förutse ett visst hushålls förväntade energikonsumtion i framtiden.

Trots att SVM är en väletablerad och effektiv maskininlärningsteknik så har den svagheter. En vanlig nackdel med SVM enligt Karamizadeh et al. [16] är dess brist på insyn i dess resultat. SVM största styrka är att dataträningen är relativt enkel.

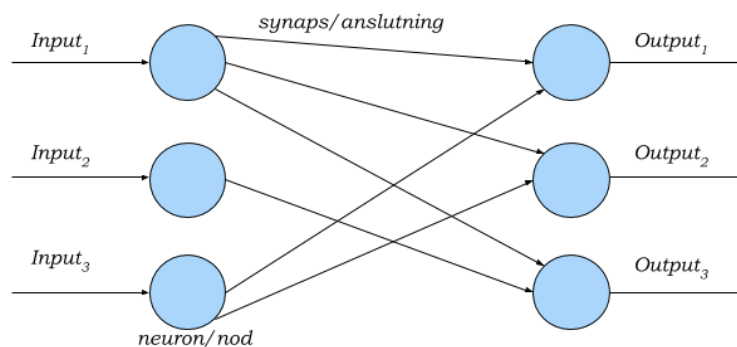
1.1.6 Neurala nätverk

Neurala nätverk (NN) är, precis som SVM, en supervised maskininlärningsteknik och den tar inspiration från hur hjärnan fungerar. För att förstå hur neurala nätverk fungerar så är det nödvändigt att ha grundläggande förståelse för hjärnans funktion.

Hjärnans mest fundamentala enhet är neuronerna/nervcellerna vars uppgift är att ta emot information från en annan neuron, bearbeta den och sedan skicka vidare den bearbetade informationen till andra neuroner. Hjärnan består således av ett massivt nätverk av neuroner som tack vare sin förmåga att bearbeta och skicka

⁷ Anomalitetsdetektering är en process som går ut på att identifiera avvikelser i ett dataset (anomalitet kan översättas till avvikelse).

vidare information ger oss förmågan att uppfatta vår omvärld och att lära oss. Anslutningen från en neuron till en annan, kallat synapser, etableras och förstärks när vi lär oss nya saker och kan förstärkas eller försvagas beroende på hur ofta den används (därför är repetition av ny kunskap så viktigt!). Anslutningen möjliggörs av neuronens tentakelliknande utskott kallat axon⁸. Buduma [32] menar att varje anslutnings styrka, m.a.o. vikt, hos en neuron summeras och summan transformeras till en ny signal, dvs. output, som skickas vidare till nästa neuron. Således är inputen till en mottagande neuron direkt beroende av outputen från den sändande neuronerna. Neurala nätverk försöker replikera hela denna process så att en dator kan lära sig på samma sätt som vår hjärna gör. För att representera en synaps använder sig neurala nätverk av begreppet anslutning⁹ och nod används för att representera en neuron. Figur 3 visualiserar ett exempel på hur ett mycket litet neuralt nätverk kan se ut.



Figur 3. Visualisering av hur ett litet neuralt nätverk kan se ut¹⁰.

Hjärnans olika delar är uppbyggda av flera olika lager av neuroner, hur många beror på delen av hjärnan. Exempelvis så består största delen av hjärnbarken¹¹ av sex lager [41]. Precis som hjärnan så har även neurala nätverk olika skikt av noder, och enligt Marsland [4] blir ett neuralt nätverk mer kraftfullt om det består av flera olika lager av noder.

Det finns många olika typer av neurala nätverk, t.ex. Deep Neural Networks, Recurrent Neural Network, Feed-Forward Neural Network, Convolutional Neural Networks etc.

Neurala nätverk har många möjliga tillämpningsområden och flera olika studier använder neurala nätverk för att skapa nya tekniker. I en studie designar Han och Oh [33] ett navigationssystem för en robot baserat på neurala nätverk i syftet att säkert och effektivt kunna ta sig till en position. Roboten har inte tillgång till någon tidigare kartläggning av miljön som den måste navigera sig igenom. Författarna testar framgångsrikt sitt system i form av både datorsimulationer och experiment i verkligheten. Neurala nätverk har även använts till bl.a. olika typer av signalbehandling, t.ex. brusreducering [37], prognostisering av energikonsumtion

⁸ Utskott som ansluter till andra neuroner kallas för axon, "lösa" utskott kallas dendrit.

⁹ eng. connection.

¹⁰ Den långa anslutningen representerar egentligen neuronens axon, synapsen befinner sig precis i slutändan av axonen där den ansluter till en annan neuron. Det är i synapsen som signalsubstansen frigörs från en neuron till en annan. För att underlätta förståelsen för konceptet så är endast synapsen/anslutningen utmärkt i figuren.

¹¹ eng. cerebral cortex: den del av hjärnan som har huvudansvar för vår mänskliga intelligens.

[38], optimering av vattenkonsumtion [39] och inom medicin för att försöka förutse aktiviteten hos receptorer för olika varianter av HIV-1-viruset [40].

1.1.7 Finansmarknaden

Handel med olika typer av tillgångar som exempelvis aktier eller valutor bedrivs oftast på digitala marknadsplatser. Där bestäms det mest aktuella marknadspriset för de tillgångar som finns tillgängliga för handel. På dessa marknadsplatser kan köpare och säljare placera ordrar. En order är ett anbud eller förfrågan om att antingen köpa eller sälja en viss kvantitet av en tillgång, även kallat orderns volym. Alla ordrar på en viss tillgång samlas i dess orderbok på marknadsplatsen. Genom orderboken kan spekulanter få en överblick av alla ouppfyllda köp- och sälj-order på just den tillgången. När kriterierna för pris och volym hos en köp-order och en sälj-order överlappar så matchas och verkställs orderarna i fråga. Det genomförs då en transaktion där de involverade parterna utbyter den kvantitet av tillgången till det pris som bestämdes av orderarna. En order kan bli fullständigt eller delvis uppfylld och tas automatiskt bort från orderboken först när hela dess kvantitet har blivit uppfylld. En order kan även stängas av skaparen innan den hinner bli fullständigt uppfylld och tas då också bort från orderboken [1].

Det används huvudsakligen två olika typer av ordrar för att köpa eller sälja en tillgång: market orders och limit orders. En market order genomförs omedelbart vid marknadens nuvarande pris. En limit order genomförs endast om ett förspecificerat pris godtas av marknaden. Vid en market order har man ingen kontroll över det slutgiltiga priset. Medan vid en limit order har man viss kontroll över priset men ingen garanti att ordern kommer att genomföras [1].

Det finns begränsningar för hur stor volym av en tillgång som går att köpa eller sälja på en marknadsplats till ett visst pris vid ett givet tillfälle. Detta kallas för likviditet. En marknads likviditet bestäms av den sammanlagda volymen av de tillgängliga köp- och sälj-order som finns för olika priser i orderboken. En orderbok är ofta väldigt föränderlig på så vis att ett stort antal ordrar öppnas och stängs varje sekund. Det är framförallt något som kallas för *high frequency trading* (HFT) som står för skapandet av majoriteten av alla ordrar. HFT syftar till datorprogram som automatisk placerar ordrar väldigt snabbt. System för HFT kan placera tusentals ordrar per sekund och fatta blixtnabba beslut om vilka strategier som ska användas. Detta medför att både priset och likviditeten kan förändras väldigt frekvent på tillgångar där sådan handel förekommer [1] [34]. Hur instabil eller hur kraftiga prissvängningarna är hos en tillgång är benämns av den ekonomiska termen volatilitet [48].

1.1.8 Arbitrage

Priset på en tillgång vid ett givet tillfälle kan skilja sig åt från en marknad till en annan. Detta innebär att det existerar en obalans i priset mellan dessa marknader. Denna obalans kan utnyttjas genom att köpa andelar av en tillgång på en marknad och sedan sälja andelar av samma tillgång på en annan marknad fast till ett något högre pris. Denna typ av handel mellan olika marknader kallas för arbitrage [1].

1.1.9 Vad är algoritmisk handel?

Algoritmisk handel¹², förkortat till algo-trading eller AT, innebär att man med hjälp av datorprogram automatiserar ett eller flera steg i investeringsprocessen. Exempelvis för att fatta beslut om att köpa eller sälja aktier eller andra tillgångar [1]. System för algoritmisk handel består oftast av följande komponenter: *pretrade analysis*, *trade signal generation* och *trade execution*. Dessa komponenter utför vardera sina uppgifter i följd efter varandra [1].

Pretrade analysis

Vid pretrade analysis analyseras finansiell data, nyheter eller andra typer av data med målet att förutsäga framtida prisförändringar eller volatilitet. Tekniker för att göra detta kan delas in i tre huvudsakliga grupper: fundamental analys, teknisk analys och kvantitativ analys [1].

Fundamental analys försöker fastslå en tillgångs verkliga värde eller framtida värde genom att undersöka relevant information såsom räntor, BNP eller arbetslöshetssiffror. Detta baserat på antagandet att priset på en tillgång inte alltid representerar det verkliga värdet [1].

Teknisk analys baseras på *efficient-market hypotesen* (EMH) som menar att priset alltid reflekterar all relevant information och därav representerar det aktuella marknadspriset alltid en tillgångs verkliga värde. EMH är omtvistad då den i princip innebär att det är omöjligt att slå marknaden när det kommer till att bestämma en tillgångs verkliga värde. Skeptiker till EMH hänvisar till att investerare i praktiken har lyckats slå marknaden genom att identifiera irrationella priser på tillgångar [49]. Teknisk analys syftar till att analysera historiska prisförändringar för att identifiera mönster som antas upprepa sig och på så vis förutspå framtida prisrörelser eller volatilitet [1].

Kvantitativ analys betraktar priset som slumpmässigt och använder matematiska modeller för att beskriva denna slumpmässighet och sedan dra nytta av den [1].

Trade signal generation

Vissa system för algoritmisk handel automatiserar endast pretrade analysis-fasen och låter sedan människor fatta beslut om det är läge att köpa eller sälja en tillgång. Medan i ett fullständigt automatiserat handelssystem är det komponenten för trade signal generation, som baserat på analysen i föregående komponent, gör en riskbedömning och därefter ger en köp- eller sälj-signal om ett visst pris och kvantitet [1].

Trade execution

Trade execution är komponenten som fattar beslut om hur ordrar ska se ut. Komponentens fattar även beslut om en stor order ska delas upp till flera mindre delordrar samt på vilken eller vilka marknader som ordrar ska läggas på. Ett problem kallat *det optimala orderexekverings-problemet* (*trade execution problem*) är

¹² eng. *algorithmic trading*, förkortat till *algo-trading*

ett problem som kan uppstå när kunder eller system för algo-trading lägger stora ordrar på en marknad. Problemet grundar sig i att själva ordern är så stor att den kan driva priset på marknaden i ofördelaktig riktning för kunden. Detta kallas för *market impact* och en strategi för att undvika detta är att dela upp ordern i mindre delar. Nackdelen med den strategin är att man riskerar något som kallas för *opportunity cost*: när marknadspriset börja röra sig ofördelaktigt innan alla delordrar har hunnit genomföras. Syftet med komponenten för trade execution är att hitta en optimal balans mellan market impact och opportunity cost [1] [34].

1.2 Problemformulering

För att kunna utforska och vidareutveckla möjligheterna som finns med att tillämpa maskininlärning vid algo-trading så är det nödvändigt att börja med att identifiera en rad olika faktorer. Först och främst måste man ta reda på vilken typ av data som är relevant för algo-trading och hur den datan i så fall kan tillämpas vid maskininlärning. Eftersom data är centralt för hur väl maskininlärning fungerar inleds studien med en frågeställning gällande just data. Givetvis är det även viktigt att ta reda på vilka typer av maskininlärningstekniker som går att tillämpa vid algo-trading. Det kan även vara av betydelse att identifiera de egenskaper hos en maskininlärningsteknik som gör den tillämpbar vid algo-trading.

Det skulle även vara intressant att ta reda på hur effektiv en maskininlärningsteknik är i sammanhanget algo-trading. Då vore det nödvändigt att jämföra olika maskininlärningstekniker som kan tillämpas vid algo-trading med varandra. Även här är det en god idé att ta reda på vilka egenskaper som är relevanta för detta ändamål. En rimlig jämförelse vore att urskilja respektive maskininlärningstekniks styrkor och svagheter, samt vilka begränsningar som förekommer.

1.3 Syfte och frågeställning

Denna studie avser att genom diskussion och resonemang i form av en litteraturstudie undersöka hur maskininlärning kan tillämpas vid algoritmisk handel. Följande är forskningsfrågorna som denna studie avser att besvara:

1. Vilka typer av data är relevanta vid algoritmisk handel?
2. Vilka maskininlärningstekniker kan tillämpas på algoritmisk handel?

1.4 Avgränsningar

Faktorer som begränsar detta arbetets omfång innefattar antalet timmar tilldelade för kursen. På grund av det väljer vi att inte implementera några av algoritmerna för praktiska jämförelser eller för att samla empirisk data. Vidare kommer vi inte heller i detta arbete att designa någon egen algoritm för testning och utvärdering.

Vår utbildningsnivå och områdeskunskap är också en begränsande faktor. Specifikt ingår inte matematisk analys i det kunskapsområde som vi har som

bakgrund i vår utbildning. På grund av detta kommer ingen djupgående matematisk analys att utföras på de artiklar som innehåller avancerad matematik.

I detta arbete kommer det inte heller att utföras någon djupgående analys av något specifikt algoritmiskt handelssystem. Detta då en sådan undersökning sträcker sig utanför omfånget av våra frågeställningar som är inriktade på olika typer av data och maskininlärningstekniker i relation till algoritmisk handel.

För vår litteraturstudie avgränsar vi oss till att huvudsakligen använda material som är peer-reviewed. Detta för att säkerställa att de informationskällor som vi använder är granskat för att garantera faktamässigt korrekt och pålitligt innehåll.

2 Metod

2.1 Metodbeskrivning

För att det ska vara möjligt att diskutera de gällande frågeställningarna måste det finnas ett vetenskapligt underlag med information som utgör grunden till analysen. Målet är, innan någon diskussion kan börja, att hitta och sammanställa tidigare akademiska verk som berör maskininlärning i samband med algoritmisk handel.

2.1.1 Informationssökningsprocessen

Olika sökplattformar nyttjas i detta arbete i syftet att hitta relevant akademisk forskning. Huvudsakligen använder vi oss utav databaserna ACM Digital Library (ACM) samt IEEE Xplore Digital Library (IEEE) och i viss mån Google Scholar. Både ACM och IEEE är väletablerade inom det datavetenskapliga området och är av denna anledning de primära källorna som vi använder för att hitta tillförlitligt material. Google Scholar används mer i andra hand ifall vi anser att det inte går att hitta tillräcklig information via de primära källorna. Detta på grund av att Google Scholar vid många tillfällen kan ge en stor mängd sökträffar med lägre relevans än vid sökningar i ACM eller IEEE. Om en sökning ger låg andel relevanta sökträffar per sida så kan det vara lämpligt att pröva sökorden i en annan ordning, böja orden eller att använda andra sökord helt och hållet. Olika söktermer och nyckelord används tillsammans med varandra, separerade från varandra och i olika kombinationer för att generera olika resultat. Avsikten är även att uppnå en så stor mångfald bland sökträffarna som möjligt, samtidigt som resultaten naturligtvis ska vara relevanta i förhållande till forskningsfrågorna.

Vid sökningarna används engelska begrepp och sökord istället för de svenska som används i detta arbete. Detta på grund av att det finns en mycket större mängd och mångfald akademiska verk som är skrivna på engelska än på svenska. De två mest centrala nyckelbegreppen som används vid sökningarna är "machine learning" och "algorithmic trading". Även synonymen "algo-trading" är ett centralt nyckelord.

Hur urvalet sker när vi väl hittat en mängd källor beskrivs i avsnittet nedan.

2.1.2 Urval av referenser: inklusions- och exklusionskriterier

För att på ett tidseffektivt sätt undersöka varje enskilt akademiskt verks relevans så börjar vi med att inledningsvis läsa en artikel översiktligt för att bilda oss en uppfattning kring hur relevant och/eller intressant en viss artikel är. Om en artikel bedöms som lovande så läser vi resten av texten mer noggrant och ingående. Än bättre är det om artikeln kan bidra med innehåll som vi ännu inte haft tillgång till. Om artikelns innehåll är användbart för att besvara våra frågeställningar, så sparas den i Zotero¹³. Om vi anser att artikeln vid översiktlig läsning inte innehåller

¹³ Zotero är ett verktyg för att samla in och organisera digitala informationsresurser, t.ex. akademiska forskningsartiklar.

relevant information, eller information som är för redundant i förhållande till det vi redan hittat, så avstår vi från att använda artikeln.

Textens typ har även betydelse för urvalet av referenser. Innehållet ska vara så objektivt och opartiskt som möjligt för att vara trovärdigt. Dagstidningar, bloggar, Wikipedia etc. har lägst pålitlighet då innehållet i denna typ av källor ofta är åsiktsbaserat, partiskt och ogranskat. Sådana källor har vi avstått helt från att använda. Källorna som används i detta arbete ska huvudsakligen vara peer-reviewed. Material som erhålls via Google Scholar är inte garanterat granskade, medan sannolikheten är större att material från databaser som t.ex. ACM och IEEE är det. Anledningen till att vi huvudsakligen fokuserar på att använda peer-reviewed material är p.g.a. att innehållet i dessa är granskat vilket genererar så hög sannolikhet som möjligt att innehållet är adekvat och korrekt. Självklart ska innehållet i vårt arbete vara baserat på korrekt information. Något vi valt att inte ta hänsyn till vid urval av källor är antalet gånger en artikel blivit citerad. Detta eftersom antalet relevanta artiklar är förhållandevis få. Ett exempel på antalet sökträffar visas i tabellen i avsnitt 2.1.3. Det är även möjligt att de allra senaste publikationerna bidrar med värdefull information även om de ännu inte har hunnit bli citerade i andra publikationer.

För artiklarnas pålitlighet är det även av stor betydelse att de refererar sitt innehåll och att detta är gjort på ett korrekt och tydligt sätt med en väletablerad referensstil som t.ex. IEEE, Harvard, Vancouver etc. Om en akademisk artikel är adekvat och tillförlitlig med god referering till sitt innehåll, så har den möjlighet att tillhandahålla många fler pålitliga och relevanta källor. Givetvis granskas även då varje enskild källas trovärdighet innan de sparas och används.

Studier och arbeten som syftar till att lösa problem m.h.a. artificiell intelligens har allt mer skiftat från att lägga huvudfokus på valet av algoritm till att istället fokusera på själva datan som appliceras i algoritmen. Det här skiftet i fokus, som skedde 2001, gör att algoritmer som tillämpar artificiell intelligens ses i ett nytt ljus som kan förändra hur man tänker och arbetar när man utvecklar denna typ av algoritmer. Av den här anledningen kommer vi att arbeta med referenser som huvudsakligen är daterade från år 2001 och framåt.

2.1.3 Exempel på sökningar och dess resultat

Som tidigare nämnt är de mest centrala nyckelbegreppen “machine learning” samt “algorithmic trading” som använts vid sökningar. Även andra termer och sökord har använts och då även i olika kombinationer för att generera sökträffar. Tabellen nedan presenterar några exempel på de källor som valts under arbetets gång. Syftet är att illustrera den sökprocess vi använt för att hitta artiklar och ge några exempel på vad dessa har resulterat i. Tabellen visar inte samtliga sökningar som gjorts då detta går utanför detta arbetes omfattning.

För att härleda hur en viss referens hittats så presenterar tabellen vilka söktermer som använts, på vilken plattform sökningen har utförts i, antal träffar som sökningen genererade samt den valda källans namn bland alla de träffar som sökningen resulterade i. Tillsammans med källans namn redovisas även det referensnummer som källan har i referenslistan i detta arbete.

Sökterm	Sök-plattform	Antal träffar	Källans titel [nummer i referenslistan]
“machine learning”	ACM	94817	Lowering the barrier to applying machine learning [5]
((algorithmic trading) AND (machine learning))	IEEE	14	Quality and consistency assurance of quote data for algorithmic trading strategies [10]
((algorithmic trading) AND (machine learning))	IEEE	14	Forecasting stock price directional movements using technical indicators: Investigating window size effects on one-step-ahead forecasting [9]
“important data in algorithmic trading”	Google Scholar	61600	Algorithmic Trading [1]
((algorithmic trading) AND value at risk)	IEEE	9	Optimal order execution for algorithmic trading: A CVaR approach [12]
((algorithmic trading) AND value at risk)	IEEE	9	Robust Optimization of Order Execution [21]
((algorithmic trading) AND value at risk)	IEEE	9	Robust order execution under box uncertainty sets [20]
((machine learning) AND usage)	IEEE	1502	Forecasting CO2 Emission with Machine Learning Methods [15]
((definition) AND RSI) OR MACD)	IEEE	65	Technical analysis of three stock oscillators testing MACD, RSI and KDJ rules in SH & SZ stock markets [29]
((definition) AND RSI) OR MACD)	IEEE	65	Discussion on the MACD and MA expert systems of securities software [30]
((algorithmic trading) AND (machine learning))	IEEE	14	Neural Network based Trading Signal Generation in Cypto-Currency Markets [34]

Figur 4. Tabell med exempel på söktermer och dess resultat från olika plattformar.

2.2 Metoddiskussion

Den generella forskningsmetod som vi har valt att använda är huvudsakligen en litteraturstudie då vi i detta arbete avser att undersöka den möjliga tillämpningen av maskininlärning vid algoritmisk handel. Först och främst måste vi ta reda på om det överhuvudtaget är genomförbart med dagens teknik. Därefter måste vi ta reda på *hur* detta i så fall skulle vara möjligt och vad som eventuellt redan har gjorts inom området. En betydelsefull faktor är att undersöka vilka faktorer som är relevanta vid maskininlärningens tillämpning vid algoritmisk handel.

För att kunna föra en diskussion kring ämnet så måste arbetet börja med att undersöka det så kallade state-of-the-art, dvs. att undersöka det senaste steget i utvecklingen av det aktuella ämnet. Detta innefattar att underrätta oss om de senaste och mest aktuella idéer och funktioner inom området. Syftet är då att tillhandahålla en grund för diskussionen vars mål är att besvara forskningsfrågorna. Oates [8] menar att det är just detta som är litteraturstudiens syfte, vilket passar detta arbetes struktur väl. Att ta del av andra studier och arbeten inom ett specifikt fält kan även vara en ingångsväg till en stor mängd betydande information. Ett verk kan t.ex. referera till flera andra betydelsefulla studier inom samma område, vilket ger en lättillgänglig åtkomst till en stor mångfald värdefulla informationskällor.

Ett annat tillvägagångssätt att utföra denna studie hade kunnat vara att designa en egen algoritm utifrån våra forskningsfrågor för att studera tillämpningen av maskininlärning vid algo-trading. Detta tillvägagångssätt har övervägts men avstods ifrån då det hade krävt mer tid än vad som finns till förfogande, som tidigare nämnt i avsnitt 1.4. Hade det funnits mer tid att tillgå hade det kunnat vara intressant att utvidga studien med praktiska experiment av olika slag. Den övergripande metodiken som då hade kunnat användas hade varit s.k. design and creation som Oates [8] beskriver den.

En annan möjlighet för att besvara frågeställning 1 hade kunnat vara att använda sig av ett öppet ramverk för algoritmisk handel för att testa det med flera olika typer av data. Något att ha i åtanke vid ett sådant tillvägagångssätt är att valet av ramverk kan ha stor påverkan på resultatet, därför bör flera ramverk testas och deras resultat jämföras.

Ett annat sätt att utvidga studien hade kunnat vara att djupare analysera de algoritmer som redan finns, som t.ex. att fördjupa sig i matematiken bakom dem. För att få en nyanserad och generaliserbar analys skulle det vara nödvändigt att djupare analysera en relativt stor mängd algoritmer, och dessutom jämföra dessa med varandra. Men återigen skulle det vara mer tidsödande än vad detta arbete tillåter. Dessvärre skulle detta tillvägagångssätt även kräva högre matematiska färdigheter än vad vår utbildningsbakgrund tillhandahåller, som också nämns i avsnitt 1.4.

3 Resultat

3.1 Relevanta typer av data vid algoritmisk handel

Detta avsnitt behandlar den första forskningsfrågan som lyder:

(1) *Vilka typer av data är relevanta vid algoritmisk handel?*

Syftet är att ta reda på vilken typ av data som är relevant och viktig att ta hänsyn till vid algoritmisk handel. Det är en mycket viktig aspekt att ta reda på innan en övervägning sker kring val av maskininlärningsteknik att tillämpa vid algoritmisk handel. Den algoritm som väljs eller utvecklas måste besitta förmågan att ta hänsyn till egenskaper och data som är relevant vid algoritmisk handel. Detta är en förutsättning för att algoritmen ska kunna utföra handeln på ett effektivt sätt.

3.1.1 Historisk prisdata

Den data som en maskininlärningsteknik lär sig av, det vill säga träningsdata, kan vara av olika typer beroende på vilken källa datan ursprungligen kommer från. En viktig typ av data som är användbar är historisk prisdata, den hämtas oftast från en marknadsplats som författarna av [35] förklarar. Historisk prisdata är en samling datapunkter som representerar hur olika egenskaper av priset på en tillgång har förändrats över tid. Samtliga ordrar som genomfördes under en viss tidsperiod sammanfattas i en datapunkt. Varje datapunkt kan lagra flera egenskaper såsom volym, högsta pris, lägsta pris, öppningspris och stängningspris för tidsperioden. Att varje datapunkt representerar en viss tidsperiod innebär till exempel att varje datapunkt representerar en sekund, en minut, en dag eller längre tid av handel, det är ett mått på datans granularitet [17]. Mer granulär data innebär att den har högre detaljnivå och således är mer komplett. Vid den allra högsta detaljnivån av historisk prisdata skulle varje datapunkt representera varje enskild order som genomförts på marknaden. Vissa marknader tillhandahåller publikt tillgänglig historisk prisdata som vem som helst är fri att använda medan andra marknader tar betalt för den [10] [35].

Historisk prisdata tillåter beräkningen av stokastiska modeller¹⁴ som ingår i gruppen för kvantitativ analys. *Arithmetic random walk model* är ett populärt exempel på en sådan modell som beskriver dynamiken av framtida prisutveckling. Modellen bygger på att marknadspriset slumpmässigt rör sig upp eller ner med en viss sannolikhet som är proportionell mot tillgångens förväntade volatilitet. En optimal strategi för att dra nytta av dessa prisrörelser kan uppnås genom att lösa ett stokastiskt dynamiskt optimeringsproblem [1].

3.1.2 Tekniska indikatorer

Historisk prisdata tillåter även beräkningen av tekniska indikatorer. Tekniska indikatorer skiljer sig från historisk prisdata som datakälla på så vis att de

¹⁴ Stokastisk modell: En matematisk modell som beskriver ett återkommande fenomen utan att ta hänsyn till någon bakomliggande orsak [51].

indikerar andra egenskaper hos priset. Egenskaper såsom trender, momentum, volatilitet och volym, vilket ger en annan bild av prisets utveckling än vad rå historisk prisdata gör. Tekniska indikatorer är verktyg som är till för att underlätta för en investerare att bättre förstå och förutsäga prisutvecklingen. På så vis kan spekulanter fatta mer välgrundade beslut om hur hen ska handla.

Olika typer av tekniska indikatorer har betydelse för algoritmisk handel på så vis att de kan användas för att utföra kvantitativa analyser på finansiella marknader [29] [30]. Shynkevich et al. [9] förklarar att tekniska indikatorer beräknas utifrån en sammansättning av volym, öppningspris, stängningspris, högsta pris och lägsta pris för en viss tidsperiod. Olika tekniska indikatorer har utvecklats oberoende som hjälpmedel av ett flertal investerare. De kan bland annat identifiera om en tillgång är såld över sitt egentliga värde (översåld¹⁵), köpt över sitt egentliga värde (överköpt¹⁶) eller hur stark alternativt svag en pristrend är. Exempel på kategorier av tekniska indikatorer innefattar de som illustrerar volatilitet, rörliga medelvärden och s.k. oscillatorer [9]. Exempel på oscillatorer är RSI (relative strength index), Stochastic Oscillator (KDJ) och MACD (moving average convergence/divergence) [29]. Chong och Ng [31] påvisade i sin studie RSI och MACD:s effektivitet för Londons aktiemarknad under en längre tidsperiod. RSI går ut på att indikera hur stark en pågående pristrend är i en viss riktning genom att beräkna förändringshastigheten hos prisrörelserna [29]. KDJ används till att identifiera nivåerna av översålda och överköpta tillgångar [29]. MACD är en väl utbredd och använd teknisk indikator [9]. MACD tas fram genom att beräkna skillnaden mellan en långsam EMA (exponential moving average) och en snabb EMA. Med långsam EMA menas ett rörligt medelvärde som beräknas utifrån en längre tidsperiod och snabb EMA innebär att ett rörligt medelvärde beräknas utifrån en kortare tidsperiod [30]. När värdet på MACD går från ett negativt till ett positivt tal så indikerar det en köpsignal. Vice versa gäller för säljsignalen som indikeras av att MACD:s värde passerar från ett positivt tal till ett negativt [29].

Olika kombinationer av tekniska indikatorer kan användas. Wu och Diao [29] menar att de i sin studie visar att en kombination av de tre tekniska indikatorerna MACD, RSI och KDJ kan ha en relativt högre sannolikhet att förutse förändringar på finansiella marknader under kort sikt. Atsalakis och Valavanis [11] hävdar att tekniska indikatorer nyttjas som input i approximativt 20% av de tekniker som utför börsmarknadsprognoser.

3.1.3 Andra typer av data

Det finns alternativa typer av data som kan användas för att uppskatta en tillgångs verkliga värde eller framtida prisutveckling i algoritmiska handelssystem. Relevant information som såsom det övergripande läget för ett lands ekonomi ingår i gruppen för fundamental analys. Fundamental analys strider mot efficient-market hypotesen eftersom man utgår från antagandet att det aktuella marknadspriset för en tillgång kan vara irrationellt och därför inte alltid motsvarar tillgångens verkliga värde [1]. Exempel på data som är relevant för fundamental analys innefattar

¹⁵ eng. *oversold*

¹⁶ eng. *overbought*

räntor, BNP, arbetslöshetssiffror samt politiska beslut. Relationen mellan sådana data och en tillgångs framtida prisutveckling har kunnat indikeras med hjälp av statistiska modeller och maskininlärning [1]. När det värde som förutspås av dessa data skiljer sig från det aktuella marknadsvärdet på en tillgång så har en potentiell framtida priskorrektion förutspåtts [1].

Andra relevanta typer av data för algoritmisk handel inkluderar sociala medier, forum och nyhetssidor. Författarna av [17] demonstrerade i sin studie ett system som klassificerar finansiella nyhetsartiklar med hjälp av maskininlärning. Ett annat exempel på en studie som nyttjar denna typ av data för sin algoritm är Smailović et al. [46] studie. I sin studie analyserar de text i flöden från Twitter. Hagenau et al. [47] är ytterligare ett exempel på en studie som använder data i form av text från finansiella nyheter till sin maskininlärningsteknik.

3.2 Maskininlärningstekniker lämpliga vid algoritmisk handel

Detta avsnitt behandlar andra forskningsfrågan. Här beskrivs hur andra studier tillämpat maskininlärningstekniker på olika typer av data relevanta för algoritmisk handel. Frågeställningen lyder:

(2) Vilka maskininlärningstekniker kan tillämpas vid algoritmisk handel?

För att kunna utvärdera möjligheterna kring att nyttja maskininlärning vid algoritmisk handel är det nödvändigt att ta reda på vilka möjligheter som redan finns. Då är det behövt att ta reda på vilka maskininlärningstekniker som används vid algoritmisk handel, verklig såväl som simulerad handel. Utifrån detta kan man även resonera kring framtida möjligheter.

Målet för trade signal generation komponenten i ett algoritmiskt handelssystem är att avgöra om det är mest lönsamt att avvakta, köpa eller sälja innehavet i en tillgång [1]. Dessa val kan generellt betraktas som de klasser som en maskininlärningsalgoritm har som uppgift att ge som output baserat på olika indata. Av detta skäl har maskininlärningstekniker för klassificering lämpat sig väl. Författare har beskrivit, utvecklat och testat system där sådan klassificering har uppnåtts med hjälp av maskininlärningstekniker. Dessa tekniker har tränats på olika typer av data för att identifiera vilken kategori som tidigare osedd data av samma typ tillhör.

3.2.1 SVM tillämpat på tekniska indikatorer och textbaserad analys

Som tidigare nämnt är SVM (support vector machine) en maskininlärningsteknik vars uppgift är att kartlägga och kategorisera datapunkter. SVM-baserade tekniker är väletablerade och har framgångsrikt använts i stor utsträckning i syftet att försöka förutse bl.a. priser och framtida trender inom finansiella prognostekniker [9] [42] [44]. Shynkevich et al. [9] använder huvudsakligen SVM som teknik i sin studie för att finna det optimala tidsintervall som ska användas för att beräkna resultatet av olika tekniska indikatorer. Författarna av [17] har också använt sig av maskininlärningstekniken SVM, fast de tillämpade den istället på data genererad från bland annat nyhetsrubriker på Internet för att fatta beslut om köp- eller sälj-signaler på finansiella marknader. Författarna av [17] utvecklade flera modeller för ett sådant system. Författarna tränade tre olika typer av algoritmer på att

klassificera data från nyhetsrubriker. Samtliga maskininlärningstekniker använde SVM och en av modellerna gav särskilt goda förutsägelseresultat med en träffsäkerhet på 57,1% och med 2,06% avkastning i simulerad handel. Kim [43] utför en studie som också visar lovande resultat med att använda SVM för att förutse framtida aktiekursindex. Författaren använder 12 olika tekniska indikatorer som inputvariabler till algoritmen. I hans experiment överträffar SVM-tekniken två andra metoder, BPN (back-propagation neural network) och CBR (case-based reasoning), som testas på samma data. Kim anser dock att SVM-tekniken kan bli än mer effektiv genom att förbättra relevanta parametrar. Kercheval och Zhang [45] nyttjade SVM i sin studie för att klassificera kortsiktiga prisförändringar från ytterst finkornig historisk prisdata av limit orders.

Det finns flera studier som har tillämpat SVM för textbaserad analys som går ut på att beskriva och tolka egenskaper i en text. Dessa egenskaper kan innefatta bl.a. innehåll och struktur. Li et al. [50] hävdar att finansiella nyhetsartiklar påverkar en tillgångs avkastning. Författarna påpekar även att textinnehållets känslomässiga attityd har stor betydelse för prisrörelsen men att detta sällan berörs. För att undersöka saken implementerar författarna ett generiskt ramverk för att förutse prisrörelser och utvärderar ramverket m.h.a. sex olika modeller med olika analyseringstekniker. Modellernas förutsägelseträffsäkerhet utvärderas och jämförs empiriskt för att mäta deras prestanda vid olika marknadsklassificeringsnivåer. Experiment utförs genom att använda fem år gamla nyhetsartiklar och börspriser från Hong Kong. Resultaten visar bl.a. att modellerna som använder attitydspolaritet inte kan ge användbara förutsägelser och att modellerna som nyttjar attitydsanalys överträffar s.k. *bag-of-words*¹⁷ modellen, i både valideringsuppsättning och oberoende testuppsättning, för enskilda tillgångar. Smailović et al. [46] utför en studie där de analyserar hur känslor uttryckta i text från Twitter-flöden kan indikera framtida prisförändringar hos en tillgång. I sin studie kategoriserar författarna texten i Twitter-flöde baserat på tre olika sinnesstämningar: positiva, negativa eller neutralt laddade ord. I studiens analys utför författarna experiment som antyder att positiva förändringar i attityd kan användas som indikator på förändringar i en tillgångs stängningspris. En annan studie som också nyttjar SVM för textbaserad analys är Hagenau et al. [47]. De undersöker om det finns möjlighet att förbättra prisförutsägelser som är baserade på textinformation från finansiella nyheter. De lyckas demonstrera att detta är möjligt genom robust urval av egenskaper kombinerat med komplexa egenskapstyper som förbättrar träffsäkerheten för SVM:s klassificering. Författarna visar även att deras utvecklade tillvägagångssätt är lönsamt för handel i praktiken.

3.2.2 Classification and regression networks

SVM och neurala nätverk räknas som de två mest effektiva metoder för att förutse prisrörelser. Neurala nätverk har använts i bred utsträckning i olika studier i syftet att försöka förutse prisrörelser hos tillgångar [42]. En av dessa studier är författarna av artikeln [35] som har utvecklat och testat två olika typer av neurala nätverk. Dessa använder data från ett antal tekniska indikatorer för att förutsäga

¹⁷ Bag-of-words är en metod för att extrahera egenskaper från en text genom att skapa ett vokabulär utifrån de ord som förekommer i en text. Texten betraktas som "bag".

en kryptovalutas framtida prISRörelser. För att träna och testa de neurala nätverken delade författarna in sin data i 80% för träning och 20% för testning. Det första nätverket var ett regressionsnätverk som tränades på att koppla värden från tekniska indikatorer till värden på tillgången. Detta resulterade inte i tillräckligt pålitliga resultat för att användas för verklig handel då felfrekvensen var för stor. I ett försök att förbättra resultaten utvecklade författarna två klassificeringsnätverk för tre respektive fem klasser. I nätverket med tre klasser representerades klasserna av "Buy", "Wait" och "Sell" vilket nådde en medelträffsäkerhet på 65,3%. Klassificeringsnätet med fem klasser kompletterades med ytterligare två klasser representerade av "Strong Buy" och "Strong Sell" för att få mer finkorniga resultat. Denna metod uppnådde en lägre medelträffsäkerhet på 43,5% vilket var förväntat av författarna. Författarna drar slutsatsen att deras resultat är lovande men fortfarande inte tillräckliga för att användas för verklig handel. Detta eftersom minsta felaktiga prISRörelse kan leda till stora monetära förluster [35]. Kara et al. [53] utför en studie där de bl.a. utvecklar en neural nätverks-modell för att försöka förutse prISRörelser för Istanbul's aktiebörs. Tio tekniska indikatorer väljs ut som input för modellen och vid experiment ger modellen en lovande genomsnittlig förutsägelseprestanda på 75,54%. Gencay och Gibson [55] samt Zhang [54] är ytterligare två exempel på studier som använder neurala nätverk för att göra prognoser för prISRörelser på den finansiella marknaden. Gencay och Gibson [55] lyckas visa att s.k. *feedforward neural network* har en framstående effekt för nämnd förutsägelseförmåga för S&P500¹⁸. Geva och Zahavi [26] bekräftar i sin studie att icke-linjära neurala nätverks-modeller har en betydande effekt på hanteringen av textbaserad data vilket innebär att denna typ av neurala nätverks-modeller kan nyttjas för att göra prognoser för marknaden.

Även hybridsystem som delvis består av neurala nätverk visar lovande resultat. Kwon och Moon [56] kombinerar s.k. recurrent neural network och en generisk algoritm för att förut säga bl.a. Nasdaq¹⁹ samt aktiekurser för 36 andra företag. I sin studie får de mer gynnsamma resultat än med traditionella finansiella metoder. Armano et al. [57] kombinerar en generisk algoritm för att hantera tekniska indikatorer och neurala nätverk för att hantera andra data för att på så vis få fram en förutsägelsemodell. Författarnas modell bedömdes som framgångsrik då den upprepade gånger fick resultat som överträffade den s.k. *buy and hold*²⁰ strategin.

Enligt Kim [43] är s.k. *back-propagation neural network* (BPNN) den mest populära NN-modellen och har använts i stor utsträckning för att förutse prISRörelser. Detta då Kim menar att BPNN har en överlägsen inlärningsförmåga samt en omfattande tillämpningsmöjlighet för många affärsproblem. Dock menar också Kim att BPNN har sina svagheter då den har svårt för att selektera ett stort antal kontrollparametrar som inkluderar relevanta inputvariabler, dold lagerstorlek samt inlärningsfrekvens. Dock så anser [43] [52] [53] att neurala nätverk överlag

¹⁸ S&P500 är ett amerikanskt aktiemarknadsindex som är baserat på marknadsvärdena av 500 stora företag.

¹⁹ Nasdaq är en amerikansk börs och är den näst största börsen i världen.

²⁰ Buy and hold är en investeringsstrategi som går ut på att en köpare köper tillgångar och behåller dessa under lång tid oavsett hur marknaden än fluktuerar.

ger oförutsägbara och inkonsekventa resultat när datan innehåller mycket brus. Även Upadhyay et al. [42] påpekar neurala nätverks problem med att hantera brus.

3.2.3 Evolutionär maskininlärning tillämpat på det optimala orderexekveringsproblemet

En viktig faktor som är av betydelse i komponenten för trade execution i ett algoritmiskt handelssystem är det som kallas för *order execution*²¹ [12] [21] [20]. Hur och när en lagd order faktiskt genomförs kan variera p.g.a. tidsfördröjningar vilket kan påverka kostnaden för en genomförd transaktion och även det slutliga priset som köparen betalar för en tillgång. Orderexekvering går ut på att utveckla en strategi som går ut på att försöka hitta en balans mellan förväntad exekveringskostnad och kalkylerad risk [12] [21] [20]. Exekveringskostnaden utgörs av skillnaden mellan ankomstpriset och det pris som erhålls för den utförda exekveringsordern²² [12]. Risk definieras i det här sammanhanget som variansen av exekveringskostnaderna [12].

Omfattande undersökningar har genomförts i strävan att analysera det optimala orderexekveringsproblemet [12]. Författarna av artikeln [34] använde en genetisk algoritm för att approximera en optimal strategi för att hitta en avvägning mellan market impact och opportunity cost. Genetiska algoritmer är en del av evolutionär maskininlärning som går ut på att man låter en population av lösningar utvecklas iterativt där sämre lösningar sällas bort för varje iteration. Vid sista iterationen återstår endast den lösning som presterat bäst. För att utveckla och testa sin strategi använde författarna sig utav en simulerad aktiemarknad. Den resulterande strategin från den genetiska algoritmen presterade betydligt bättre än den naiva referensstrategi som man jämförde med [34].

Bertsimas och Lo [13] föreslår en lösning som går ut på att utföra handel som sker i konstant takt, då avsikten är att minimera det förväntade värdet av exekveringskostnaden under en bestämd tidsperiod. Almgren och Chriss [14] utökar Bertsimas och Lo's [13] lösningsförslag genom att inkludera variansen av exekveringskostnaden. Denna varians används som ett mått på riskerna med exekveringen för att kunna göra en avvägning mellan förväntad exekveringskostnad och variansen.

Variansen av exekveringskostnaden, dvs. exekveringskostnadens spridning, används inte sällan för att mäta risken med en exekvering. Dock anser McNeil et al. [24] att variansen inte är ett lämpligt sätt att uppskatta denna risk när det handlar om finansiella avkastningar från icke-normala, negativt snedställda och s.k. leptokurtosiska distributioner²³.

Somliga anser att s.k. value at risk (VaR) kan vara ett bättre sätt att mäta risken med en exekvering [12] [20] [21]. Det finns många olika sätt att mäta VaR och även förvirring kring definitionen av VaR [18]. Peng och Li [18] definierar VaR som det mest använda måttet av risk i modern riskhanteringspraxis. VaR är ett sätt att mäta hur stor en potentiell förlust kan bli i en ekonomisk investering. Feng

²¹ sv. *orderexekvering*

²² Inom den finansiella litteraturen är detta även känt under namnet eng. *implementation shortfall* [12]

²³ eng. *leptocurtic distribution* är en term inom den statistiska världen som är en fördelning (distribution) som är mer kraftigt toppad än en normal fördelning [27].

m.fl [12] utforskar möjligheterna med VaR i sin studie. Förutom att angripa det optimala orderexekveringsproblemet är avsikten även att möjliggöra för mer generella avkastningsfördelningar. De använder sig av en specifik typ av VaR som kallas för eng. *conditional VaR*²⁴ (CVaR) som endast fokuserar på den del av avkastningen som är relaterad till oönskade utfall [12]. Författarna menar att tillämpningen av CVaR på exekveringskostnaden som riskmått i orderexekveringsproblemet möjliggör att mäta risken för olika avkastningsfördelningar mer utförligt. Feng's et al. [12] hävdar att deras studie genererar resultat som visar att deras CVaR-teknik reducerar exekveringskostnaden och att den då kan förbättra orderexekveringsstrategin. Ytterligare studier utforskar möjligheterna med CVaR för samma ändamål [21] [20].

Det råder även delade meningar kring om VaR eller CVaR är mest effektiv, då Sarykalin et al. [25] menar att valet helt enkelt beror på sammanhanget. CVaR är överlägsen VaR när det kommer till de matematiska egenskaperna som tekniken besitter. Därför menar författarna att CVaR enkelt kan användas inom optimering och statistik.

²⁴ Även känt under namnen: Expected Shortfall, Expected Tail Loss, Tail Conditional Expectation, and Tail VaR [12].

4 Analys och diskussion

4.1 Diskussion om data i relation till efficient-market hypotesen

Som resultat för den första frågeställningen fann vi huvudsakligen tre relevanta typer av data som visade sig vara tillämpningsbara vid algoritmisk handel: historisk prisdata, tekniska indikatorer samt andra datatyper relevanta för fundamental analys såsom räntor, BNP och arbetslöshetssiffror. En iakttagelse från detta resultat är att historisk prisdata ofta används som bas för att beräkna andra typer av data eller indikatorer. Modeller såsom arithmetic random walk model samt beräkningen av alla de olika tekniska indikatorer som finns bygger på någon form av bearbetning av historisk prisdata. Inget av de system för algoritmisk handel som undersökts i denna studie använde sig utav historisk prisdata direkt. Historisk prisdata används däremot ofta indirekt genom framförallt tekniska indikatorer.

Den något omtvistade efficient-market hypotesen (EMH) säger att marknadspriset reflekterar all relevant information och därför representerar marknadspriset alltid en tillgångs verkliga värde. Eftersom EMH antyder att framtida kurspåverkande information inte kan förutspås av historisk data innebär det att alla typer av data som beräknas utifrån historisk prisdata, exempelvis tekniska indikatorer, indirekt också besitter samma negativa egenskap. Däremot genom att anta att tidigare händelser är periodiskt återkommande och att marknadspriset alltid reflekterar all relevant information så borde dessa återkommande mönster också reflekteras i historisk prisdata. Tekniska indikatorer bygger på just det antagandet och försöker identifiera och utnyttja återkommande trender och mönster i prISRörelser snarare än att undersöka de bakomliggande faktorer som orsakade dem. Därav kan man dra slutsatsen att alla typer av data som beräknas utifrån historisk prisdata endast kan identifiera återkommande mönster och kan inte förutsäga nya och tidigare osedda händelser eller information som är kurspåverkande.

Datatyperna kan därav delas in i två kategorier. Kategorin med de datatyper som bygger på att EMH stämmer och att marknadspriset alltid reflekterar all relevant information. Och den andra kategorin med de typer av data som bygger på att marknadspriset på en tillgång kan vara irrationellt och inte alltid representerar det verkliga värdet, alltså data relevant för fundamental analys. Data relevant för fundamental analys kan innehålla ny och tidigare osedd kurspåverkande information, exempelvis information från nyheter eller sociala medier.

Detta leder till frågan om de båda kategorierna av data kombinerat i ett och samma handelssystem skulle kunna ge mer korrekta prisförutsägelser. I ett sådant system skulle fundamentala data ge en förankring till vad en tillgångs verkliga värde uppskattas till samt tillåta systemet reagera på ny kurspåverkande information och händelser. Medan de datatyper som beräknas utifrån historisk prisdata, som exempelvis tekniska indikatorer, skulle ge en förutsägelse utifrån återkommande trender och mönster. Baserat på data från de båda kategorierna

skulle ett handelssystem kunna sammanställa resultatet och på så vis fatta mer välgrundade beslut och möjligtvis bli mer lönsamt.

Det observerades dock relativt få exempel på algoritmiska handelssystem som använde sig utav datatyper för fundamental analys, med undantag för textbaserad analys av nyhetsrubriker och sociala medier. Detta skulle kunna bero på att det möjligtvis är svårare att få tag i stora mängder av sådana data och samtidigt säkerställa en genomgående hög kvalitet av den. Det skulle även kunna bero på att fundamentala data är svårare att tillämpa för maskininlärning i jämförelse med andra datatyper.

4.2 Undanhållande av forskningsframsteg

Vi har noterat att det kan finnas ett ekonomiskt incitament att inte publicera de studier som uppnått finansiellt gynnsamma resultat. Detta eftersom författarna har mer att vinna på att hålla sina upptäckter hemliga och använda sin kunskap för att implementera ett system för egen ekonomisk vinst. Ifall en lyckad studie hade publicerat sina fynd och andra sedan bygger liknande system skulle det kunna leda till att systemet inte längre blir lönsamt då det i sådana fall skulle tävla mot andra system som bygger på samma principer. På grund av detta incitament kan man spekulera i att många lyckade resultat möjligtvis hemlighålls istället för att publiceras. Då går det att diskutera hur verklighetstrogen state-of-the-art egentligen är inom området och det skulle även strida mot den moraliska konsensus som Vetenskapsrådet förespråkar. Vetenskapsrådet [3] menar att nya kunskaper som erhållits genom forskning inte ska få hemlighållas eller döljas och att vetenskapliga framsteg ska ses som resultatet av samarbetet mellan olika generationer av forskare.

4.3 Optimal orderexekvering och arbitrage

Många problem inom finans innebär ofta hög komplexitet då ett stort antal dimensioner och parametrar måste tas hänsyn till. Två exempel på sådana problem är val av tillgångar i en portfölj eller optimal orderexekvering. Anledningen till att evolutionära tekniker lämpar sig väl för att lösa dessa typer av problem inom finans beror på att många av problemen ofta är optimeringsproblem. Optimeringsproblem är alltså en grundläggande aspekt av finans. Evolutionära tekniker har redan tillämpats inom många områden av finans och har visats fungera väl för att lösa vissa typer av finansiella problem. Cui et al. [34] approximerar en optimal orderexekveringsstrategi m.h.a. en genetisk algoritm som är en heuristisk metod avsedd för att lösa just optimeringsproblem. Författarna belyser även att en annan evolutionär teknik kallad *particle swarm optimization* (PSO) också kan tillämpas på problemet [34].

Utmaningar med optimal orderexekvering är ett något som förekommer när stora ordrar ska placeras på en marknad. Marknadspriset kan förändras i en potentiellt negativ riktning för kunden innan en stor order har hunnit genomföras. Kundens strävan är att optimera parametrar såsom hur en order ska delas upp för att den ska genomföras så snabbt som möjligt [1] [34]. Kriterierna för optimal

orderexekvering är identiska och av speciellt stor betydelse vid den typ av handel som kallas för arbitrage [1]. En utmaning vid arbitrage är att det är väldigt tidskänsligt då marknadspriser kan förändras väldigt snabbt, därför strävar man efter omedelbar exekvering av ordrar vid ett exakt pris. Skillnader i pris mellan marknader är oftast små. Därför krävs det handel med stora belopp för att arbitrage ska vara lönsamt, detta medför även att riskerna ökar då mer kapital står på spel.

Sammantaget är utmaningarna med arbitrage identiska med de utmaningar som tekniker för optimal orderexekvering ämnar lösa. I resultatet för vår andra frågeställning fann vi en tillämpning av evolutionär maskininlärning för att approximera en optimal lösning på just detta problem. En sådan lösning skulle möjligtvis även kunna tillämpas för handelssystem som utför arbitrage.

5 Slutsatser

Vi har undersökt vilka typer av data som är relevanta vid algoritmisk handel samt vilka maskininlärningstekniker som är möjliga att tillämpa. För att besvara dessa frågor har vi genomfört en litteraturstudie där vi studerar och analyserar tidigare publikationer inom relevanta ämnesområden. Dessa publikationer påvisade huvudsakligen tre relevanta typer av data samt ett flertal maskininlärningstekniker som kunde tillämpas vid algoritmisk handel. De relevanta typer av data som identifierades var historisk prisdata, tekniska indikatorer samt den typ av data som ingår i fundamental analys. En iakttagelse som gjordes var att historisk prisdata ofta används som grund för att sedan bearbetas om till andra typer av data. Det vanligaste exemplet på detta är tekniska indikatorer. Vi fann även ett antal maskininlärningstekniker som av tidigare publikationer visades vara tillämpningsbara vid algoritmisk handel. Publikationer påvisade att en maskininlärningsteknik kallad SVM (support vector machine) kunde tillämpas på tekniska indikatorer samt analys av nyhetsrubriker. Vi fann även publikationer som visade tillämpningen av två typer av neurala nätverk, klassifikationsnätverk samt regressionsnätverk, för att generera trade signals. Vi fann också en tillämpning av evolutionär maskininlärning som användes för att approximera en lösning på det optimala orderexekveringsproblemet.

5.1 Framtida studier

Vidare undersökningar som vore intressant är att försöka återskapa ett handelssystem med maskininlärning som beskrivits i en tidigare studie. Resultaten från den ursprungliga undersökningen och den nya implementationen skulle sedan kunna jämföras för att antingen stärka eller motsäga de resultat som tidigare redovisats.

Även ett nytt system för algoritmisk handel skulle kunna designas, implementeras och utvärderas. För att på ett akademiskt vis utföra en sådan undersökning hade en metodik liknande design and creation kunnat användas [8]. En sådan studie bör inledas med en analys av vilka algoritmer för maskininlärning som finns och vilka av dessa som redan beprövats med framgång samt vilka som haft mindre lyckad framgång. Vilka egenskaper som är relevanta när olika maskininlärningstekniker jämförs är en annan fråga att ta hänsyn till. När det kommer till att välja vilken maskininlärningsteknik som ska användas kan det vara lämpligt att undersöka vilka fördelar och nackdelar har dessa tekniker gentemot varandra samt deras begränsningar. Det bör även noga övervägas vilken typ av data samt vilken kombination av data man väljer för att träna sin algoritm på för att generera köp- och sälj-signaler.

Referenser

- [1] G. Nuti, M. Mirghaemi, P. Treleaven, och C. Yingsaeree, "Algorithmic Trading", i *Computer*, vol. 44, nr 11, s. 61–69, nov. 2011.
- [2] S. J. Russel och P. Norvig. "Acting humanly: The Turing Test approach", i *Artificial Intelligence: A Modern Approach*, 3rd ed. New Jersey, USA: Pearson Education, 2016, s. 2-5.
- [3] Vetenskapsrådet, *God forsknings sed*. Stockholm: Vetenskapsrådet, 2017.
- [4] S. Marsland, *Machine Learning: An Algorithmic Perspective*, 2:a uppl. Chapman and Hall/CRC, 2014.
- [5] K. Patel, "Lowering the Barrier to Applying Machine Learning", i *CHI '10 Extended Abstracts on Human Factors in Computing Systems*, New York, NY, USA, 2010, s. 2907–2910.
- [6] H. Wang, E. Segal, A. Ben-Hur, D. Koller, och D. L. Brutlag, "Identifying Protein-protein Interaction Sites on a Genome-wide Scale", i *Proceedings of the 17th International Conference on Neural Information Processing Systems*, Cambridge, MA, USA, 2004, s. 1465–1472.
- [7] J. Lester, T. Choudhury, N. Kern, G. Borriello, och B. Hannaford, "A Hybrid Discriminative/Generative Approach for Modeling Human Activities", i *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, San Francisco, CA, USA, 2005, s. 766–772.
- [8] B. J. Oates, "Researching Information Systems and Computing", London: Sage Publications, 2005, s. 71-87.
- [9] Y. Shynkevich, T. M. McGinnity, S. Coleman, Y. Li, och A. Belatreche, "Forecasting stock price directional movements using technical indicators: Investigating window size effects on one-step-ahead forecasting", i *2014 IEEE Conference on Computational Intelligence for Financial Engineering Economics (CIFEr)*, 2014, s. 341–348.
- [10] S. Koschnicke, V. Grossmann, C. Starke, och M. Schimmler, "Quality and consistency assurance of quote data for algorithmic trading strategies", i *2014 IEEE Conference on Computational Intelligence for Financial Engineering Economics (CIFEr)*, 2014, s. 255–261.

- [11] G. S. Atsalakis och K. P. Valavanis, "Surveying Stock Market Forecasting Techniques - Part II: Soft Computing Methods", i *Expert Syst. Appl.*, vol. 36, nr 3, s. 5932–5941, apr. 2009.
- [12] Y. Feng, F. Rubio, och D. P. Palomar, "Optimal order execution for algorithmic trading: A CVaR approach", i *2012 IEEE 13th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, 2012, s. 480–484.
- [13] D. Bertsimas och A. W. Lo, "Optimal control of execution costs", i *Journal of Financial Markets*, vol. 1, nr 1, s. 1–50, apr. 1998.
- [14] R. Almgren och N. Chriss, "Optimal execution of portfolio transactions", i *Risk*, vol. 3, s. 5–40, 2001.
- [15] E. GARİP och A. B. OKTAY, "Forecasting CO2 Emission with Machine Learning Methods", i *2018 International Conference on Artificial Intelligence and Data Processing (IDAP)*, 2018, s. 1–4.
- [16] S. Karamizadeh, S. M. Abdullah, M. Halimi, J. Shayan, och M. j Rajabi, "Advantage and drawback of support vector machine functionality", i *2014 International Conference on Computer, Communications, and Control Technology (I4CT)*, 2014, s. 63–65.
- [17] R. P. Schumaker och H. Chen, "Textual analysis of stock market prediction using breaking financial news: The AZFin text system", i *ACM Transactions on Information Systems*, vol. 27, nr 2, s. 1–19, feb. 2009.
- [18] J. Peng och S. Li, "Analysis of Different Versions of the Credibilistic Value at Risk", i *2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery*, 2009, vol. 6, s. 96–100.
- [19] S. Shakya och S. Sigdel, "An approach to develop a hybrid algorithm based on support vector machine and Naive Bayes for anomaly detection", i *2017 International Conference on Computing, Communication and Automation (ICCCA)*, 2017, s. 323–327.
- [20] Y. Feng, D. P. Palomar, och F. Rubio, "Robust order execution under box uncertainty sets", i *2013 Asilomar Conference on Signals, Systems and Computers*, 2013, s. 44–48.
- [21] Y. Feng, D. P. Palomar, och F. Rubio, "Robust Optimization of Order Execution", i *IEEE Transactions on Signal Processing*, vol. 63, nr 4, s. 907–920, feb. 2015.

- [22] J. Ma, M. N. Nguyen, och J. C. Rajapakse, "Gene Classification Using Codon Usage and Support Vector Machines", i *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 6, nr 1, s. 134–143, jan. 2009.
- [23] J. S. Lee och J. D. Tong, "Applications of support vector machines to standby power reduction", i *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2016, s. 002738–002743.
- [24] A.J. McNeil, R. Frey, och P. Embrechts, "Quantitative risk management: Concepts, techniques and tools", New Jersey, USA: Princeton University, 2005.
- [25] S. Sarykalin, G. Serraino, och S. Uryasev, "Value-at-Risk vs. Conditional Value-at-Risk in Risk Management and Optimization", i *State-of-the-Art Decision-Making Tools in the Information-Intensive Age*, Z.-L. Chen, S. Raghavan, P. Gray, och H. J. Greenberg, Red. INFORMS, 2008, s. 270–294.
- [26] T. Geva och J. Zahavi, "Empirical Evaluation of an Automated Intraday Stock Recommendation System Incorporating Both Market Data and Textual News", i *Decis. Support Syst.*, vol. 57, s. 212–223, jan. 2014.
- [27] J. Law, "Leptokurtic distribution", i *A Dictionary of Finance and Banking*, Oxford University Press, 2018.
- [28] P. Ranganarayanan, N. Thanigesan, V. Ananth, V. K. Jayaraman, och V. Ramakrishnan, "Identification of Glucose-Binding Pockets in Human Serum Albumin Using Support Vector Machine and Molecular Dynamics Simulations", i *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 13, nr 1, s. 148–157, jan. 2016.
- [29] M. Wu och X. Diao, "Technical analysis of three stock oscillators testing MACD, RSI and KDJ rules in SH SZ stock markets", i *2015 4th International Conference on Computer Science and Network Technology (ICCSNT)*, 2015, vol. 01, s. 320–323.
- [30] H. Huang och P. Wang, "Discussion on the MACD and MA Expert Systems of Securities Software", i *2013 International Conference on Computer Sciences and Applications*, 2013, s. 59–62.
- [31] T. T.-L. Chong och W.-K. Ng, "Technical analysis and the London stock exchange: testing the MACD and RSI rules using the FT30", i *Applied Economics Letters*, vol. 15, nr 14, s. 1111–1114, nov. 2008.
- [32] N. Buduma, *Fundamentals of Deep Learning: Designing Next-Generation Machine Intelligence Algorithms*, 1:a uppl. O'Reilly, 2017.
- [33] S. J. Han och S. Y. Oh, "Evolutionary algorithm based neural network controller with selective sensor usage for autonomous mobile robot navigation", i

IJCNN'01. International Joint Conference on Neural Networks. Proceedings (Cat. No.01CH37222), 2001, vol. 3, s. 2194–2199 vol.3.

[34] W. Cui, A. Brabazon, och M. O'Neill, "Efficient Trade Execution Using a Genetic Algorithm in an Order Book Based Artificial Stock Market", i *Proceedings of the 11th Annual Conference Companion on Genetic and Evolutionary Computation Conference: Late Breaking Papers*, New York, NY, USA, 2009, s. 2023–2028.

[35] A. Özdemir och S. Bogosyan, "Neural Network based Trading Signal Generation in Cypto-Currency Markets", i *2018 International Conference on Artificial Intelligence and Data Processing (IDAP)*, 2018, s. 1–4.

[36] T. Hendershott, C. M. Jones, och A. J. Menkveld, "Does Algorithmic Trading Improve Liquidity?", i *The Journal of Finance*, vol. 66, nr 1, s. 1–33, 2011.

[37] N. Vlahović och G. Kvaščev, "Noise reduction by using autoassociative neural networks", i *2016 13th Symposium on Neural Networks and Applications (NEUREL)*, 2016, s. 1–5.

[38] T. Hossen, A. S. Nair, R. A. Chinnathambi, och P. Ranganathan, "Residential Load Forecasting Using Deep Neural Networks (DNN)", i *2018 North American Power Symposium (NAPS)*, 2018, s. 1–5.

[39] J. R. dela Cruz, R. G. Baldovino, A. A. Bandala, och E. P. Dadios, "Water usage optimization of Smart Farm Automated Irrigation System using artificial neural network", i *2017 5th International Conference on Information and Communication Technology (ICoIC7)*, 2017, s. 1–5.

[40] S. L. Lamers, M. Salemi, M. S. McGrath, och G. B. Fogel, "Prediction of R5, X4, and R5X4 HIV-1 Coreceptor Usage with Evolved Neural Networks", i *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 5, nr 2, s. 291–300, apr. 2008.

[41] L. R. Squire, D. Berg, F. Bloom, S. du Lac, och A. Ghosh, *Fundamental neuroscience*, 3rd ed. Amsterdam; Boston: Elsevier / Academic Press, 2008.

[42] V. P. Upadhyay, S. Panwar, R. Merugu, och R. Panchariya, "Forecasting Stock Market Movements Using Various Kernel Functions in Support Vector Machine", i *Proceedings of the International Conference on Advances in Information Communication Technology & Computing*, New York, NY, USA, 2016, s. 107:1–107:5.

- [43] K. Kim, "Financial time series forecasting using support vector machines", *Neurocomputing*, vol. 55, nr 1–2, s. 307–319, sep. 2003.
- [44] S.-H. Huang, S.-H. Lai, och S.-H. Tai, "A Learning-based Contrarian Trading Strategy via a Dual-classifier Model", i *ACM Trans. Intell. Syst. Technol.*, vol. 2, nr 3, s. 20:1–20:20, maj 2011.
- [45] A. N. Kercheval och Y. Zhang, "Modeling high-frequency limit order book dynamics with support vector machines", i *Quantitative Finance*, vol. 15, nr 8, s. 1315-1329, juni 2015.
- [46] J. Smailović, M. Grčar, N. Lavrač, och M. Žnidaršič, "Stream-based active learning for sentiment analysis in the financial domain", i *Information Sciences*, vol. 285, s. 181-203, nov. 2014.
- [47] M. Hagenau, M. Liebmann, och D. Neumann, "Automated news reading: Stock price prediction based on financial news using context-capturing features", i *Decision Support Systems*, vol. 55, nr 3, s. 685-697, juni 2013.
- [48] NE Nationalencyklopedin AB, "volatilitet". [Online]. Tillgänglig vid: <https://www.ne.se/uppslagsverk/encyklopedi/lång/volatilitet>. [Åtkomstdatum: 11-juni-2019].
- [49] M. H. Pesaran, "Predictability of Asset Returns and the Efficient Market Hypothesis," A. Ullah och D.E. Giles, eds., *Handbook of Empirical Economics and Finance*, Taylor & Francis, 2010, s. 281-312.
- [50] X. Li, H. Xie, L. Chen, J. Wang, och X. Deng, "News impact on stock price return via sentiment analysis", i *Knowledge-Based Systems*, vol. 69, s. 14–23, okt. 2014.
- [51] NE Nationalencyklopedin AB, "stokastisk modell". [Online]. Tillgänglig vid: <https://www.ne.se/uppslagsverk/encyklopedi/lång/stokastisk-modell>. [Åtkomstdatum: 12-juni-2019].

- [52] K. Kim och I. Han, "Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index", i *Expert Systems with Applications*, vol. 19, nr 2, s. 125–132, aug. 2000.
- [53] Y. Kara, M. Acar Boyacioglu, och Ö. K. Baykan, "Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange", i *Expert Systems with Applications*, vol. 38, nr 5, s. 5311–5319, maj 2011.
- [54] G. P. Zhang, "A Neural Network Ensemble Method with Jittered Training Data for Time Series Forecasting", i *Information Sciences*, vol. 177, nr 23, s. 5329–5346, dec. 2007.
- [55] R. Gencay och R. Gibson, "Model Risk for European-Style Stock Index Options", i *IEEE Transactions on Neural Networks*, vol. 18, nr 1, s. 193–202, jan. 2007.
- [56] Y. Kwon och B. Moon, "A Hybrid Neurogenetic Approach for Stock Forecasting", i *IEEE Transactions on Neural Networks*, vol. 18, nr 3, s. 851–864, maj 2007.
- [57] G. Armano, M. Marchesi, och A. Murru, "A Hybrid Genetic-neural Architecture for Stock Indexes Forecasting", i *Information Sciences*, vol. 170, nr 1, s. 3–33, feb. 2005.