NTNU
Norwegian University of
Science and Technology

Research method in AI: Reproducibility of results

Nicklas Grimstad Nilsen

Master of Science in Informatics
Submission date: June 2018
Supervisor: Odd Erik Gundersen, IDI

Norwegian University of Science and Technology
Department of Computer Science

Abstract

In recent years, increasing attention has been given to an apparent issue within the scientific fields where the reproducibility of scientific results is questioned. This trend is commonly termed a crisis of reproducibility, or a reproducibility crisis. While medicine is a central scientific field here, it spans most other fields as well, including the computational sciences.

This thesis investigates the reproducibility of recent and central scientific publications within artifical intelligence, through a selection of the most cited papers the past few years. The reproducibility is investigated by attempting to reproduce the presented results, based on the description of the method presented in the paper.

Sammendrag

I de senere årene har det blitt gitt økt oppmerksomhet til et stadig synligere problem innen det vitenskapelige feltene, der det stilles spørsmål ved reproduserbarheten til vitenskapelige resultater. Denne trended henvises ofte til som en reproduserbarhetskrise. Medisin er et av de mer sentrale vitenskapsfeltene her, men problemet dekker de fleste felter, inkludert de komputasjonelle.

Denne oppgaven undersøker reproduserbarheten til senere og sentrale vitenskapelige publikasjoner innen feltet kunstig intelligens, gjennom et utvalg av de mest siterte publikasjonene de seneste årene. Reproduserbarheten undersøkes ved å forsøke å reprodusere resultatet som blir presentert, basert på beskrivelsen av metoden presentert i paperet.

Acknowledgment

I would like to thank my supervisor, Odd Erik Gundersen, and research collaborators Odd Cappelen and Martin Mølnå, for the time we have spent together over the past year.

N.N.

Contents

Acronyms

AAAI Association for the Advancement of Artificial Intelligence. 10, 13, 56, 58, 62

IEEE Institute of Electrical and Electronics Engineers. 25

IJCAI International Joint Conferences on Artificial Intelligence. 10, 13, 56, 58, 59

LNCS Lecture Notes in Computer Science. 25

NTNU Norwegian University of Science and Technology. 3, 5

Chapter 1

Introduction

1.1 Motivation

A core part of the scientific method is the cycle of posing a hypothesis, devising an experiment to test this hypothesis, and performing that experiment. The results (as well as prior knowledge) can then be used to refute or revise the hypothesis, (or perhaps strengthen it). The idea is then that the field of science through this methodology moves forward as a whole, building on past results. [Oat06, pp. 283-285]

Carefully documenting this process has several benefits for the community:

- Others in the community will be able to share methods, and perhaps improve on elements missing from what they themselves were doing

- Others can perform the same experiment to ascertain that they too reach the same result, or if not, through comparing documentations perhaps find weak points in the hypothesis or methodology.

- Higher transparency is achieved, which lowers the chances of (both accidental and intentional) fraudulence

The ability of others to perform the same experiment to compare the results is referred to by terms like reproducibility, repeatability, or replicability. Unfortunately there is a trend (a so-called reproducibility crisis) of researchers being able to reproduce scientific results, both independent researchers and the researchers who achieved the results in the first place. [Bak16]

With validation of experimental results being a central part of the scientific method, these problems are concerning.

1.2 Problem Outline

The purpose of this thesis is to investigate reproducibility of research in computer science. More specifically, the extent of various degrees of reproducibility (for an independent entity) of the results published in papers, relying on the papers' documentation of the process (methodology, experimental setup, dataset(s), source code(s), and so on).

The belief is that investigations into the reproducibility of published research might provide insights into whether there are aspects that can be identified as contributing to this, and if there are aspects that increase reproducibility.

1.3 Research Context

This research was conducted as part of my master's thesis at Norwegian University of Science and Technology (NTNU). The research task was formulated by Odd Erik Gundersen, my supervisor. The research, although this thesis being written independently, is a collaboration. with another group's master's thesis. My research was conducted alongside the research of this second group, with both overlapping areas of contribution as well as independent contributions. The details of this collaboration is elaborated on in chapter 4.

The research builds on prior research into contents of papers from conferences where the papers were scored based on whether they included or did not various elements. The results from this prior research indicate that the degree to which papers document the method is highly variable, for instance whether source code – or even pseudo code – is published, or whether the dataset is documented. The results additionally indicate that the documentation of results data in papers is rather lackluster.

Moving forward, it could be argued that the presented findings from the analysis (which pertain to the documentation level of the papers) might not be indicative of the difficulty in reproducing or replicating the results of the papers. A possible step forward, which is investigated in our research, is to actually attempt to reproduce the results presented in a selection of papers, using the available documentation. The hope is that this research may provide an insight into among other things whether the methods presented in the prior research can provide indications for the degree of difficulty in reproducing results

1.4 Hypothesis, Objectives, and Research Questions

This thesis covers two of the three overall topics in our research. The first topic covered concerns the method by which a set of papers to attempt reproducing is selected, and the second topic covered (which is also covered by the other group) concerns the results of attempting to reproduce the selection of papers. The selection methodology is documented in chapter 5, and the reproduction is documented in chapter 7.

The third overall topic, which is covered by the other group, concerns the methodology to follow when reproducing a paper, and the documentation framework for this process.

1.4.1 Paper Selection Process

HYP-sel 1 The search results from Scopus return sets of higher impact papers that cover a diverse set of subjects within the field of Artifical Intelligence

RQ-sel 1 Is Scopus a decent choice as a bibliographic database with respect to the content coverage and citation impact of the results?

The objective of these research questions on paper selection is to provide an idea of whether the selection process is feasible with respect to producing a list of papers, where that list has decent coverage over topics in the field.

1.4.2 Paper Reproducibility

The hypothesis **HYP-repr 1** as well as the research questions **RQ-repr 1**, **RQ-repr 2**, and **RQ-repr 3** have been adapted from [Kje17]

HYP-repr 1 The documentation of experiments in published AI papers is not sufficient to be considered feasibly reproducible for independent researchers.

HYP-repr 2 There is a correspondence between the documentation level of a paper and the reproducibility of the paper.

RQ-repr 1 What is the state of reproducibility of AI papers?

RQ-repr 2 What documentation is missing from AI papers to support reproducibility?

RQ-repr 3 Documentation practices have improved over time.

RQ-repr 4 Are there elements of the documentation that surface when using the framework of the collaborating group of researchers (Odd Cappelenand Martin Mølnå), that indicate reproducibility problems?

These research questions aim to identify whether there are difficulties in actually reproducing research results in artificial intelligence, and if there is, attempt to identify what makes this difficult.

1.5 Research Approach

The prior work on which the thesis builds examines papers from the top conferences in the AI-field. Instead of using the same selection process, this thesis attempts to motivate a selection process based on the citation counts of papers.

The thesis first investigates methods by which papers can be selected based on citation counts, as well as analyzing such a selection to substantiate that this is a reasonable method to select papers. Reproduction of the selected papers is then attempted and data of the process gathered by using a common framework and process, developed in a separate but collaborating thesis.

1.6 Research Contributions

This theis has the following contributions

- A methodology by which to select papers for use in our reproduction attempts.
- An evaluation of the reproducibility of a subset of the selected papers, and the main issues encountered in the process.

1.7 Limitations

- The research is limited to the duration of a master's thesis at NTNU, i.e two semesters.
- The presented sample of papers attempted is fairly small, so it is difficult to draw broader conclusions based on statistics.
- The sample of papers is not random, so generalizing to all papers (in AI) is not really feasible.
- We have access to some hardware, but this access is still limited. If e.g a cluster of GPUs is required, this is unavailable to us.

1.8 Structure

The paper is structured as follows: Chapter 1 introduces the overall topic of the research, the overall goals and contributions, as well as the context of the research.

The related research is then covered in chapter 2, as well as common understandings of terms like reproducibility and replicability in the scientific community.

Chapter 3 presents our understanding and use of terms like reproducibility, as well as presenting some of the prior work that our research is a continuation of.

The methodology followed in our research is presented in broad strokes in chapter 4, from how to select papers to the results from attempting to reproduce them. The problem of selecting papers is then presented in 5, and a methodology for selection is proposed. The proposed methodology is evaluated with respect to the research questions on selection, the results presented and analysed in chapter 6.

The reproduction process is covered in chapter 7, with the reresults presented in chapter 8. Chapters 9 discusses these results. 10 concludes the thesis.

1.9 Disclosure

In line with the idea that properly documenting research is necessary to increase its reproducibility, we will be striving for releasing the artefacts generated from the research, e.g method code, experiment code, and code for generating the figures and results presented in the paper.

The released source code and artefacts are made available on various GitHub repositories, gathered under a common Organization Account [1] for our research project – with “us” being the two collaborating groups of researchers. In this paper, more specific references to repositories will be provided as results generated from code in that repository are presented, where feasible.

[1] https://github.com/AIReproducibility2018

Chapter 2

Background

Problems with reproducing results is prevalent in the scientific fields. More than half of the researchers in various fields when asked in a questionnaire, responded that they had failed to reproduce an experiment, for some fields the proportion remaining above 50% even for a researchers own results [Bak16] [BI15].

There is little consensus on the terminology when it comes to reproducibility [GFI16]. Terms like reproducibility or replicability are often used without being explicitly defined, and other times used interchangably, it being difficult to say whether the choice is stylistic or due to subtle differences. In [BI15, p. 117], at first the phrase "...in proportion to their [available] resources, the cost of attempting to reproduce a study can be substantial ..." is used, and shortly after the phrase "What is difficult to quantify is the opportunity cost associated with studies that fail to replicate". [GFI16, p. 2], noting that the usage of the terms reproducibility and replicability have varying definitions, proposes new terminology centered around the word reproducibility (as opposed to giving new definitions for existing words): methods reproducibility, results reproducibility, and inferential reproducibility. Here, methods reproducibility covers "[The] ability to implement, as exactly as possible, the experimental and computational procedures, with the same data and tools, to obtain the same results", results reproducibility covers "[The] production of corroborating results in a new study, having followed the same experimental methods", and inferential reproducibility covers "[The] making of knowledge claims of similar strength from a study replication or reanalysis".

"Reproducible research" is a somewhat more agreed upon term, often being attributed to Claerbout and his colleagues in the 90s [GFI16] [VKV09] [LMS12] through their efforts in computational science to provide interactive runnable code on CD-ROMs accompanying their publications. The ideal is that a reader should be able to run the accompanying code and be able to produce the same figures as in the paper. A central point is that the paper is not the research itself, but that it rather advertises the research – the research being the environment (e.g collection of data and code) which produced the analysis and figures presented in the paper. [BD98]

[Pen11] calls replication the standard by which scientific claims are judged, and goes on to use the term reproducibility in the context of making the data and computer code used to analyze results publicly available, so that others can analyze the same data. They then say that this falls short of “full replication”, with the reasoning that the same data is used rather than independently collected data. This use of the terms is repeated in [LP15], reproducibility being recomputation while replication is independent researchers targeting the same research question reaching the same result.

In contrast, [LMS12] uses the term “replicable” to mean the ability to re-run the same code and achieve the same result, and “reproducible” to mean creating an independent implementation that verifies the published result.

[VK11] uses the term repetition for rerunning the exact same experiment with the same method on the same or similar system, where the result obtained is the same or very similar. Their definition of reproduction is similar to that of [LMS12], used in the context of independent confirmation – where independent researchers base themselves on the contents of the paper and linked resources (though the term replication is not covered).

[VKV09, p. 39] uses a definition of reproducible research, where “a research work is called reproducible if all information to the work, including, but not limited to, text, data and code, is made available, such that an independent researcher can reproduce the results”, and presents six degrees of reproducibility based on time and effort required of an independent researcher. They additionally perform a survey on 134 papers, using a short questionnaire, to assess the reproducibility practices in signal processing at the time. Their overall conclusion is that algorithms are well-documented, though they raise the concern that the review did not involve actually implementing the algorithms, and that some issues might not surface until during the process of attempting an implementation.

[CPW15] presents a study on 601 papers, registering whether the results presented in the papers are backed by code, and if so whether obtaining the code is possible, and if so whether they were able to build the code (the result of attempting to run the code is not a high priority in their study). Out of 402 papers with results backed by code, they were able to obtain the code for 226 papers (either through the paper, through a web search, or through emailing the authors). For 130 papers the code built successfully within 30 minutes, and for 64 of the papers it took longer (for 23 of the papers, the authors ensured the code would build with reasonable effort). This comes out to the code being obtainable in 56% of the cases, and out of the attempts at building obtained code, 32% were successful within 30 minutes (increasing to 56% with no hard time limit).

Chapter 3

Documenting for reproducibility

This chapter presents terminology and work on documenting for reproducibility.

3.1 Defining reproducibility

In this and the following chapters, reproducibility will be understood as it is in [GK18]:

> **Definition**. Reproducibility in empirical artificial intelligence research is the ability of an independent research team to produce the same results using the same method based on the documentation made by the original research team.

Likewise, repeatability is understood as conducting the same experiment and obtaining the same result (as opposed to understanding it as reproducibility, which is understood to be broader than exactly repeating an experiment).

Independent research team is understood as a research team that only uses the documentation made available by the original research team in attempting to reproduce the results.

Additionally, [GK18] makes a distinction between the research method employed, and the AI method employed (a more abstract notion than the notion of the AI program that implements the AI method).

3.2 Evaluating reproducibility

In [Gun15], requirements for what is required of benchmarks in computing for them to be reproducible are presented. This covers what information needs to be disclosed – for instance what has been done (as specified by the software program running the experiment(s), any data, and the results), motivations, and detailed information about the environment and infrastructure used to run the experiment(s).

Based on the requirements, [Gun15] goes on to propose a metric that measures the replicability of a scientific body of work, by noting that the requirements can be broken down along three dimensions: **Experiment Procedures**, **Data and Results**, and **Documentation**. Each of these have several aspects to them, for instance under Experiment Procedures, aspects include the source code for the method and the source code for the experiment. Using these aspects, a metric function is proposed. In its simpler form as used, it counts the presence of an aspect in the published work as 1.0, and counts absence as 0.0. The values are then summed and divided by the total number of aspects, resulting in a reproducibility score in the interval $[0.0, 1.0]$.

[Kje17] adapts this survey, and applies it to a sample of 400 papers from two journals, visiting each twice in separate years. The journals covered are AAAI in the years 2014 and 2016, and IJCAI in the years 2013 and 2016. The survey is given more detailed evalation criteria, and an evaluation procedure to follow when going through a paper is devised.

[GK18] further develops the survey by in line with [GFI16] noting that there are degrees of reproducibility, and goes on to define three reproducibility levels – R1, R2, and R3. These levels are defined in order of increasing generality (and decreasing requirements of the available documentation), the definitions repeated here for reference:

- **R1: Experiment Reproducible** The results of an experiment are experiment reproducible when the execution of the same implementation of an AI method produces the same results when executed on the same data.

- **R2: Data Reproducible** The results of an experiment are data reproducible when an experiment is conducted that executes an alternative implementation of the AI method that produces the same results when executed on the same data.

- **R3: Method Reproducible** The results of an experiment are method reproducible when the execution of an alternative implementation of the AI method produces the same results when executed on different data.

Using the survey and the defined reproducibility levels, they present a metric for measuring the degree by which a paper fits each reproducibility level. The categories cover aspects presented in [Kje17] and [Gun15], relating to method, data, and experiment. Each reproducibility level depends on a given selection of these categories, R3 having the fewest dependencies (only relying on the method category) while R1 has the most documentation dependencies (relying on code, data, and the method). A metric for the degree to which a paper belongs in a given level is then calculated similarly to in [Gun15], except that the categories (and their aspects) that are not relevant for the given level are not used when calculating the degree the paper fits that level.

3.3 Moving forward

Our research involves attempting to reproduce a selection of papers. In that regard, it is beneficial to have a structured process to follow for consistency in both methodology and collection of data about the process. A collaborating group of researchers Odd Cappelen and Martin Mølnå, building on [GK18] and prior work, have devised such a framework for collecting data, as well as a methodology that is designed with the limitations of this research project in mind.

Our usage of this methodology and framework on a selection of papers is detailed in chapter 7, and the results of this process are detailed in chapter 8.

As a final note, while [GK18] presents categories R1, R2, and R3, our research groups came across papers which had code published, but not the dataset used. The original definition of R2 requires data to be available (as does R1), and this case does not truly fit with the definition of R3, where data is not available. We have therefore, as presented in the autumn project report of the other research group, decided to split the category R2 into two new categories:

- R2-D, Data Reproducible – An alternative implementation of the method being used on the same data (this is is identical to the original definition of R2).

- R2-M, Method Reproducible – The same implementation of the method being used on new data (this is a new definition, that covers the case described above).

In the rest of this thesis, unless R2-M is explicitly mentioned, mentions of R2 are taken to refer to the first category (R2-D).

Chapter 4

Research Method

This chapter presents the overall methodology followed in our research.

The overall goal of the research is to evaluate reproducibility in the field of Artificial Intelligence, by attempting to reproduce a selection of papers. A process to achieve this will have several steps, where each step may require particular methods.

With this goal in mind, it becomes clear that a selection of papers is necessary. A process for selecting papers (i.e generating data for the next steps of the research) is therefore devised, which is detailed in chapter 5. The process is then applied, resulting in the selection of papers presented and analysed in chapter 6.

In order to improve consistency, it is important to be methodical in the attempts at reproducing the selection of papers, as well as documenting each attempt and its results. To achieve this, a framework and methodology has been devised by the collaborating research group, as part of their autumn project for their master's thesis.. This was briefly described in section 3.3, and is revisited in chapter 7 where the methodology and survey is applied, and in chapter 8 where the data generated from the reproduction attempts is analysed.

With a selection of papers, and a methodology and framework to follow and use for recording data during the reproduction attempts, the next step is to follow through with the reproductions. The initial goal was to identify the reproducibility categories of the papers, and then attempt to reproduce the experimental papers in the R1 category (validating theoretical papers being considered out of scope for our research). That is, the category described in section 3.2 as the least general one, requiring the most documentation (more specifically, source code had to be available). A short while into the process, the experiment was modified so to incorporate reproduction of R2 papers as well. That is, as long as the data used is available, the paper is a candidate for reproduction, even if source code is unavailable. Papers classified as R3 were documented using the framework, but reproduction was not attempted.

Chapter 5

Paper selection methodology

This chapter presents the methodology followed for generating a selection of papers for use in the process of attempting to reproduce their presented results.

The chapter is divided into two parts: deciding where to select papers from, and deciding how to select the papers from that source.

5.1 Source to select papers from

The prior work on which the thesis builds examines papers from the top conferences in the AI-field. Using a similar selection process here is worth considering, though with reference to the main research goal, one wishes to be able to say something about the reproducibility of the field as a whole. In that sense, being limited to one or a few particular journals might not be too representative of the field. Another benefit of performing a different selection of papers is to make the dataset an independent variable, providing opportunities for validating the results of the prior work on a new dataset.

This could be achieved by selecting different conferences from the prior work (AAAI, IJCAI), though the concern that one journal might be too narrow still applies.

A different criteria is using a metric such as the citation count (or other citation impact measures) of a paper; using the citation count as a criteria for selection was proposed by our supervisor, Odd Erik Gundersen, considered by having in mind that highly cited papers by virtue of being highly cited, very likely have had a notable impact (be that a positive or negative impact).

Having the idea of using citation count as a selection criteria, the source to select papers from is constrained to sources that feasibly enable ranking papers by that measure. Several bibliographic databases provide such measures, so an investigation into the features provided by some of these to determine whether some have advantages over others is a natural next step.

5.2 Bibliographic databases

In the following, several bibliographic databases are considered, based on to what degree they fulfill certain (unformalized) criteria, e.g if some database turned out to have a usability feature that was advantageous enough to make a second database cumbersome to use in comparison.

Examples of criteria that surfaced during the investigation:

- The database provides citation counts for papers, or some other citation metric to rank papers.

  (this was a pre-defined requirement).

- The search system is convenient for restricting the returned papers to a particular field of interest, or restricting the search by other criteria.

  (i.e artificial intelligence).

- The database supports exporting the search result in a convenient way to a convenient format.

  (e.g exporting bibliographic information).

- Transparency in how the database and search operates.

  (e.g why a result was returned given a search should be reasonable).

Databases considered:

- Microsoft Academic [1]
- Google Scholar [2]
- Scopus [3]
- Web of Science [4]

[1] https://academic.microsoft.com/
[2] https://scholar.google.com/
[3] https://scopus.com/
[4] https://apps.webofknowledge.com/

Microsoft Academic is developed by Microsoft Research. When the user searches for a topic, e.g "Artificial Intelligence", it lists a set of subcategories in the side bar (for instance neural networks or evolutionary algorithms). A few test-searches to experiment with it revealed some obscurities in the way it handles searches, however. In this specific case, after having searched for "Artifical Intelligence", the engine returned a set of results. Appending part of the title of the first result to an attempted refined search, lead to this first result not being included in the results.

A bit of reading revealed that the search engine uses semantic search, which appears to attempt to interpret what the user wanted, rather than what the user actually typed. This has benefits in not limiting the accessibility of the database to only the users who have experience in formulating their queries, but unfortunately makes it difficult to somewhat explicitly know what is going on behind a given search.

Microsoft Academic otherwise enables sorting the result by citation count, and has a fairly intuitive way to browse results by field of study. There does not appear to be any easy way to batch-export the search results to a convenient format.

Google Scholar uses the Google Search interface to search for academic papers, listing results in a similar way to its usual search engine, with some included details about citation counts and sources. It otherwise appears to be very limiting in letting the user define the search or order and export results.

Scopus is a subscription-based bibliographic database owned by publisher Elsevier. [5] It is reasonable to raise questions about whether there is a conflict of interest, and whether the database favours papers published by Elsevier. These concerns become less prominent when it is revealed that the site is run by an independent international board, which decides what journals and papers to include. The content coverage policy is openly available. [Els17]

The actual search is far more flexible than that of Google Scholar and Microsoft Academic, permitting searches in particular datafields (keywords, title, abstract, and others), filtering by year, filtering by document type (conference paper, book chapter, review, etc), by conference, and various others. It also permits sorting by citation count, and lets the user export the metadata of the (in practice) first 2000 search results.

Web of Science is a subscription-based bibliographic database run by Clarivate Analytics.[6] It appears rather similar to Scopus in terms of features, though its interface seemed a bit more difficult to use than Scopus (although that might be due to having become acustomed to Scopus' interface beforehand).

Out of these, Scopus and Web of Science are the most advantageous for the purposes of sampling papers, with Google Scholar and Microsoft Academic being limited by their

[5] https://elsevier.com/

[6] https://clarivate.com/

interface, searching, or export-capabilities. Out of Scopus and Web of Science, it is difficult to say whether one has advantages over the other, though Scopus was chosen for our purposes largely due to there not being any clear advantages to Web of Science, and due to Scopus being considered earlier, already proving to have satisfactory functionality.

It is worth noting that due to there not being one true, consistent way to count citations, it is infeasible to compare citation counts across bibliographic databases. Because of this, choosing one database and sticking to it makes sense. Taking into account that the latter two databases in particular appear to have desirable features and content coverage, there does not appear to be much necessity in generating data using multiple databases, either way.

5.2.1 Performing a selection

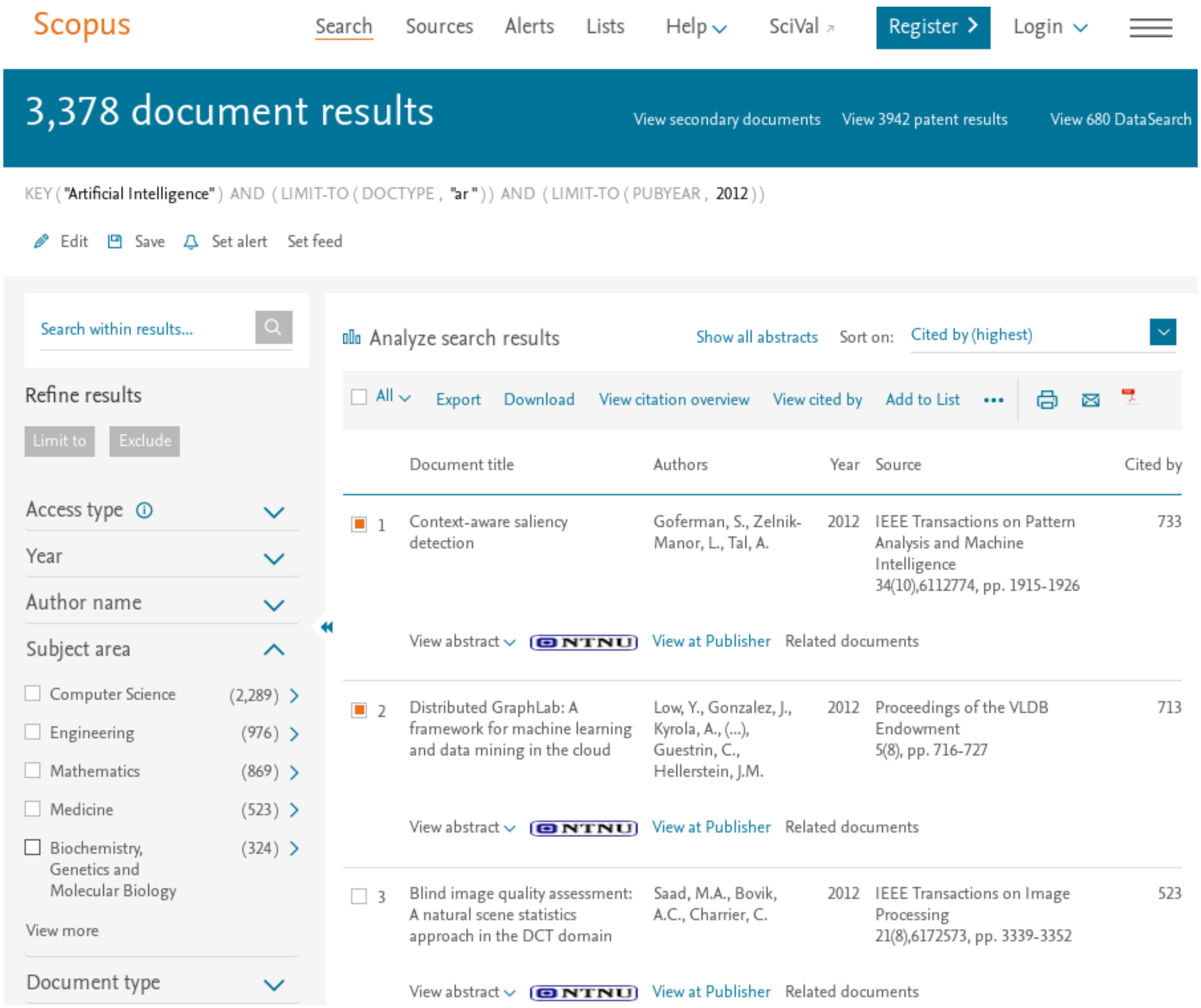
Having chosen a bibliographic database, this section proposes and motivates a procedure for using the database to generate a dataset (i.e a list of papers to use for the reproduction attempts).

Scopus has papers from a wide variety of scientific fields (and not just computer science or artificial intelligence). A natural first step is therefore to devise a method for filtering out papers from fields and on topics that are not of interest.

The interface for searching the bibliographic database (Scopus) permits constructing a query from field matches and mismatches, combined with logical connectives. An example of this is shown in figure 5.1, where a search has been performed for publications with document type of article, published in the year 2012, and with “Artificial Intelligence” in any of the keyword fields (either author keywords or index keywords).

Below, a search query to be used is proposed. The process of selecting it cannot be described as more than by trial and error, though the result of performing the search along with an analysis is presented and analysed in chapter 6. The goal of this analysis is an evaluation of the performance of using the search, with respect to the distribution of attributes of the papers (like source journals or index keywords), and whether this distribution can be used to say whether the paper selection appears to be representative of the field of artifical intelligence, e.g by whether the subfields covered by the keywords is wide, or whether they are very narrow (for instance, should almost all the results happen to be about neural networks).

The various parts of the query in figure 5.2 have specific motivations. The actual search term, “Artificial Intelligence”, has the underlying assumption that the bibliographic database (or the source journals) accurately adds the keyword to papers that either come from source journals with almost exclusively AI papers, or papers that have other author keywords from the field (for example Support Vector Machines); after all, it is unreasonable to expect that all authours of papers would themseles use the Artificial Intelligence keyword. Chapter 6 attempts to validate this assumption, through investigation of keyword attributes.



3,378 document results

KEY ("Artificial Intelligence") AND (LIMIT-TO (DOCTYPE , "ar")) AND (LIMIT-TO (PUBYEAR , 2012))

Subject area	
□ Computer Science	(2,289)
□ Engineering	(976)
□ Mathematics	(869)
□ Medicine	(523)
□ Biochemistry, Genetics and Molecular Biology	(324)

	Document title	Authors	Year	Source	Cited by
☒ 1	Context-aware saliency detection	Goferman, S., Zelnik-Manor, L., Tal, A.	2012	IEEE Transactions on Pattern Analysis and Machine Intelligence 34(10),6112774, pp. 1915-1926	733
☒ 2	Distributed GraphLab: A framework for machine learning and data mining in the cloud	Low, Y., Gonzalez, J., Kyrola, A., (...), Guestrin, C., Hellerstein, J.M.	2012	Proceedings of the VLDB Endowment 5(8), pp. 716-727	713
□ 3	Blind image quality assessment: A natural scene statistics approach in the DCT domain	Saad, M.A., Bovik, A.C., Charrier, C.	2012	IEEE Transactions on Image Processing 21(8),6172573, pp. 3339-3352	523

Figure 5.1: Example search for 2012 articles with "Artificial Intelligence" in the keywords

```
KEY ( "Artificial Intelligence" )
AND NOT  ( TITLE-ABS-KEY ( "survey" )  OR  TITLE-ABS-KEY ( "review" ) )
AND  ( LIMIT-TO ( DOCTYPE ,  "cp " )  OR  LIMIT-TO ( DOCTYPE ,  "ar " )  OR
       LIMIT-TO ( DOCTYPE ,  "ch " )  OR  LIMIT-TO ( DOCTYPE ,  "ip " ) )
AND  ( LIMIT-TO ( PUBYEAR ,  2012 ) )
```

Figure 5.2: Search query used for the year 2012. Other than changing the PUBYEAR, the other years are equivalent.

The decision to search only in the keyword-fields (as defined by Scopus), as opposed to inlcuding e.g the abstract-field as well, is motivated in part by a desire to have the presence of each paper in the results be attributable to an (to a larger degree) explicit reason, and in part to filter out some papers that mention AI in passing without otherwise being related to the field.

The document type codes are used by Scopus to indicate book chapter, article, article in press, and conference paper. The motivation for choosing only these is that the remaining document types are for documents like survey papers (although it appears some papers that perhaps should have been classified as that, are classified as e.g article, but these form a very small portion, that perhaps from Scopus' point of view have an ambigous document type), which rarely contain experimental results. Survey papers are not considered relevant to our research goals, which concern reproducing experimental results, and for that reason such document types have been excluded as far as that is achievable.

The motivation behind excluding papers that contain the words “survey” or “review” in the abstract, title, or keywords is largely intended to filter out leftover papers from the previous document type step, where these papers are less likely to contain experimental results, and therefore had they been present would be taking up time being investigated, only to be discarded manually for the same reason. There is some probability that papers with experimental results that inadvertantly mention these terms are excluded by this decision. This tradeoff between at one end of the scale excluding nothing for additional manual work, and at the other end of the scale excluding almost every imaginable term mostly used by survey, review, and tutorial papers, is deemed a reasonable tradeoff.

The reasoning for limiting the search to a specific year is twofold. Firstly because the research questions are about the current state of the field, so using all years would be less appropriate. Secondly, Scopus appears to have a limit on the number of search results that can be exported at once, thus searching for the full range of interest runs the risk of one year with a lot of highly cited papers pushing other years out of the list. By performing the search for each year separately, as long as each year actually has enough papers published, the same number of most cited papers from each year is guaranteed to be present in the final exported sets.

The actual search process essentially involves:

- Use the search query to get a list of results.
- Order the results by citation count
- Export the first (e.g) 2000 results to a convenient format

Chapter 6

Paper selection results and analysis

This chapter evaluates the selection of papers made by following the procedure devised in chapter 5. In particular, the research question **RQ-sel 1** is investigated.

6.1 The selected sample

The exported selections of papers for each year, being the 2000 most cited papers of each year (based on the citation count measurements of Scopus), is made available at one of our GitHub repositories. [1]

For illustrative purposes, a select few columns for one of the exported years (2016) is shown for the first few papers in table 6.1

6.2 Attributes of sampled papers

Keeping in mind that the overall research goal is to attempt reproducing a selection of papers, as well as the research limitations (i.e, time), it is highly unreasonable to expect more than 100 papers (and even that) to be investigated for each year. For this reason, it may be valuable to perform the following analyses on a limited set of papers per year, in addition to performing the analyses on all the 2000 papers per year. Given that the selection criteria is by citation count, it makes sense to select the first few papers from the 2000 for this purpose. The number of papers to use for the shorter set has been set to 60 per year, resulting in $60 * 5 = 300$ papers in total for the shorter set – compared to the $2000 * 5 = 10000$ papers for the full set.

[1] `https://github.com/AIReproducibility2018/UTILS_exported_papers`

Table 6.1: First few papers for year 2016, sorted by citation count (limited to a subset of the columns).

	Cited by	Title	Authors	Source title
1	571	Mastering the game of Go w...	Silver D., Huang A., Maddi...	Nature
2	152	Deep Convolutional Neural ...	Shin H.-C., Roth H.R., Gao...	IEEE Transactions on Medic...
3	98	MLlib: Machine learning in...	Meng X., Bradley J., Yavuz...	Journal of Machine Learnin...
4	97	XGBoost: A scalable tree b...	Chen T., Guestrin C.	Proceedings of the ACM SIG...
5	89	Social big data: Recent ac...	Bello-Orgaz G., Jung J.J.,...	Information Fusion
6	77	Deep neural networks: A pr...	Jia F., Lei Y., Lin J., Zh...	Mechanical Systems and Sig...
7	71	ISuc-PseOpt: Identifying l...	Jia J., Liu Z., Xiao X., L...	Analytical Biochemistry
8	69	Classification with Noisy ...	Liu T., Tao D.	IEEE Transactions on Patte...
9	52	Deep convolutional and LST...	Ordóñez F.J., Roggen D.	Sensors (Switzerland)
10	50	Deep learning for remote s...	Zhang L., Zhang L., Du B.	IEEE Geoscience and Remote...
11	49	Generalized Correntropy fo...	Chen B., Xing L., Zhao H.,...	IEEE Transactions on Signa...
12	45	Learning Rotation-Invarian...	Cheng G., Zhou P., Han J.	IEEE Transactions on Geosc...
13	45	Transition-Aware Human Act...	Reyes-Ortiz J.-L., Oneto L...	Neurocomputing
14	43	Multi-Directional Multi-Le...	Ding C., Choi J., Tao D., ...	IEEE Transactions on Patte...
15	42	”Why should i trust you?” ...	Ribeiro M.T., Singh S., Gu...	Proceedings of the ACM SIG...

Generated_by: Appendix B: F)

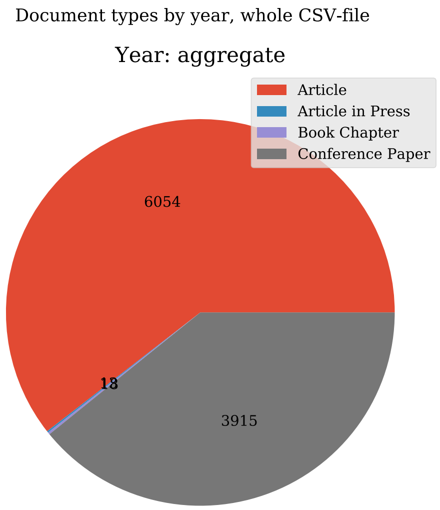


Figure 6.1: Proportions for each document type of all exported papers for each year

Generated_by: Appendix B: F)

6.2.1 Document Types

While it might be less relevant for investigating the hypothesis that the sample of papers (both the 60 and the 2000) span a wide range of topics, figures 6.1 and 6.2 show the proportion of each document type that the exported papers are classified as by Scopus. It is mostly of interest for the purpose of seeing what kind of document types (among the ones exported) are the more common in the field. Based on these results, articles and conference papers are by far the most common in artificial intelligence. It also makes sense that article in press is rare, as the years in question are 2012 through and including 2016, while the export was made late 2017, when most papers from the previous years should have been published already.

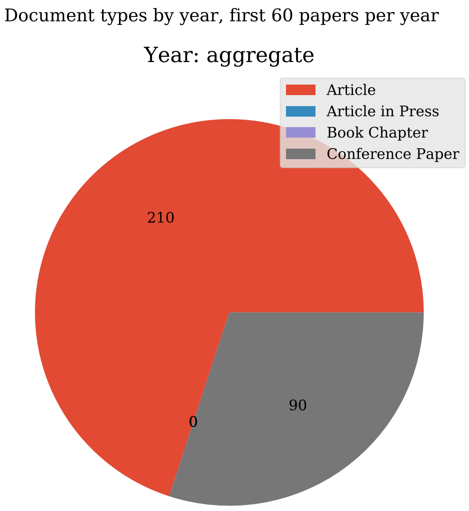


Figure 6.2: Proportions for each document type of first 60 exported papers for each year

Generated_by: Appendix B: F)

6.2.2 Citation Counts

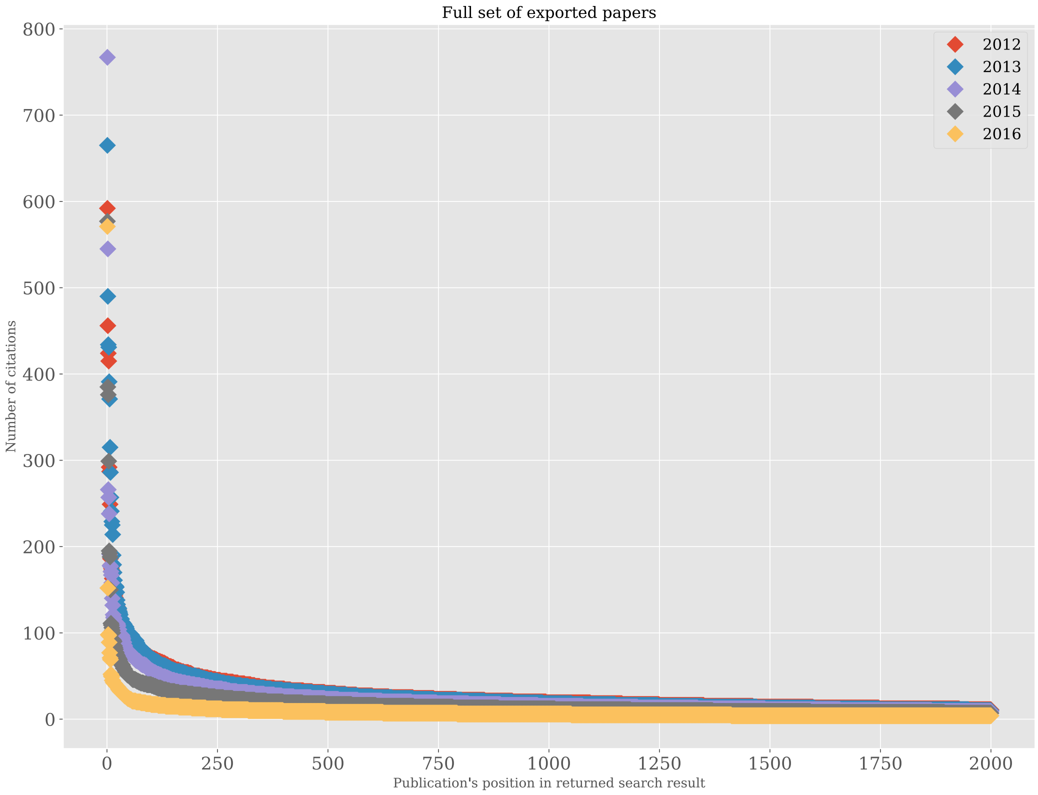


Figure 6.3: Citation counts as a function of position in exported list, for all papers. Tail of data is shown in table 6.2

Table 6.2: Last few papers and their citation counts, full set

	2012	2013	2014	2015	2016
1996	11	10	9	7	4
1997	11	10	9	7	4
1998	11	10	9	7	4
1999	11	10	9	7	4
2000	11	10	9	7	4

Generated_by: Appendix B: F)

Figures 6.3 and 6.4 indicate that the larger portion of the first 60 papers for each year do have more than just a few citations each, alleviating some of the concern that after the first couple

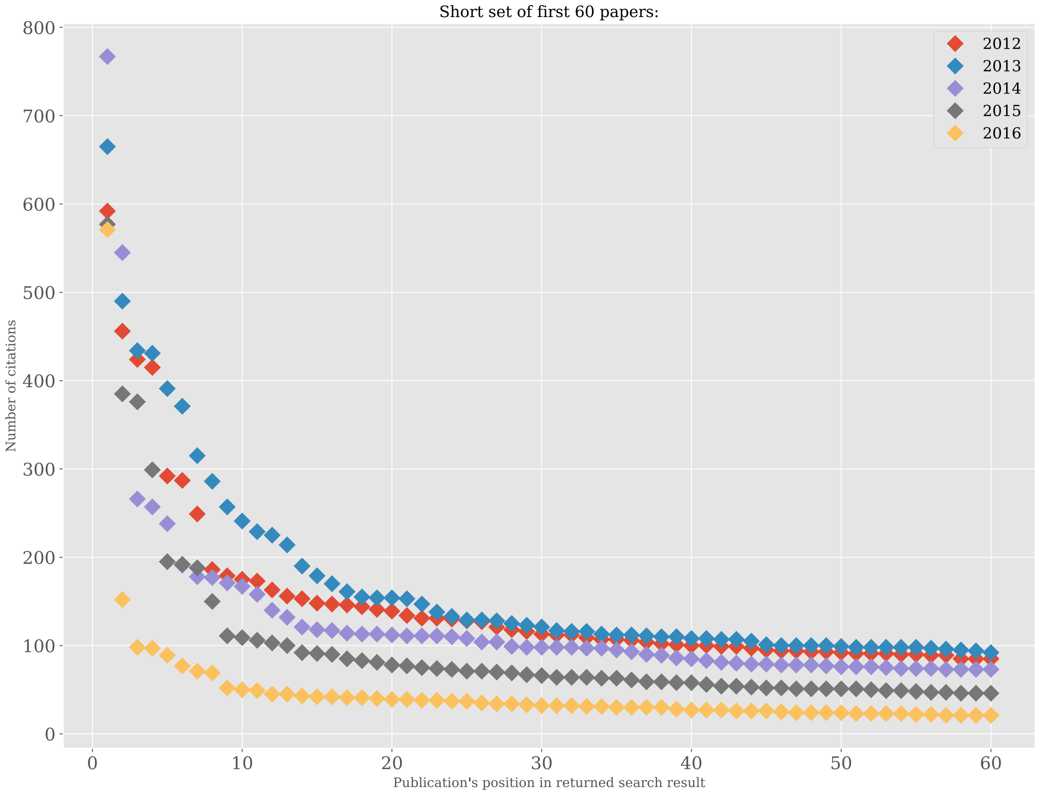


Figure 6.4: Citation counts as a function of position in exported list, for the first few papers. Tail of data is shown in table 6.3

Table 6.3: Last few papers and their citation counts, short set

	2012	2013	2014	2015	2016
56	89	97	74	47	22
57	89	96	73	47	21
58	85	95	73	46	21
59	85	94	73	46	21
60	85	92	73	46	21

Generated_by: Appendix B: F)

of papers all citation counts show only one a handful of citations (e.g 2-3 citations). Had this not been the case, this sample of 60 papers would fail to fulfill some of the expectations – namely, one of the intentions of making a selection based on citation count being to obtain papers that in some way are important to the field/having high impact.

There is nonetheless a downwards trend in the number of citations as the position of the paper in the list increases, which by the 60th paper while not reaching zero, does reach fairly low compared to the highest citation counts.

Because it is difficult to read the exact values for the later papers (where the variance starts to even out), The last few papers for each of the full sets and shortened sets are shown in tables 6.2 and 6.3, respectively. Here, the last exported paper from the most recent year (2016) has 4 citations, the last exported paper from 2012 having 11 citations. For the shorter set of 60 papers, the last paper with the least citations is from 2016, with 21 citations.

Compared to the papers with the most citations (more than 500 for each year), these are rather low values (though it should be noted that the search result from Scopus contained a lot more papers than just the 2000 that were exported, many having 0 registered citations).

Keeping in mind that the year 2016 as of writing is fairly recent (not to mention that there is a time difference between 2012 and 2016), it is reasonable that papers from 2016 will tend to have gathered fewer citations than older papers like the ones from 2012; the last paper in the shorter set from year 2012 has 85 citations (compared to the 21 of the paper from 2016), which is a far step up from the citation counts of 4-10 at the bottom of the full set – it appears somewhat unreasonable to call the last few papers in the shorter set of 60 papers obscure (in the sense that they are unknown).

6.2.3 Journals and Other Sources

Looking into which journals the exported papers come from, the figures 6.5 and 6.6 (with statistics presented in tables 6.4 and 6.5) indicate that even though the various IEEE transactions and Lecture Notes in Computer Science (LNCS) are by far more common, other journals are still represented (there being around half as many journals as there are papers, for the short set – the number of journals to number of papers is closer to $\frac{1}{4}$ for the full set of 2000 papers). Some conferences also only have papers in the sample for one or a few of the years, as indicated by the bar charts.

It is reasonable to expect larger journals to have a larger presence, though it is nonetheless apparent that more than just a few journals are represented in the sample.

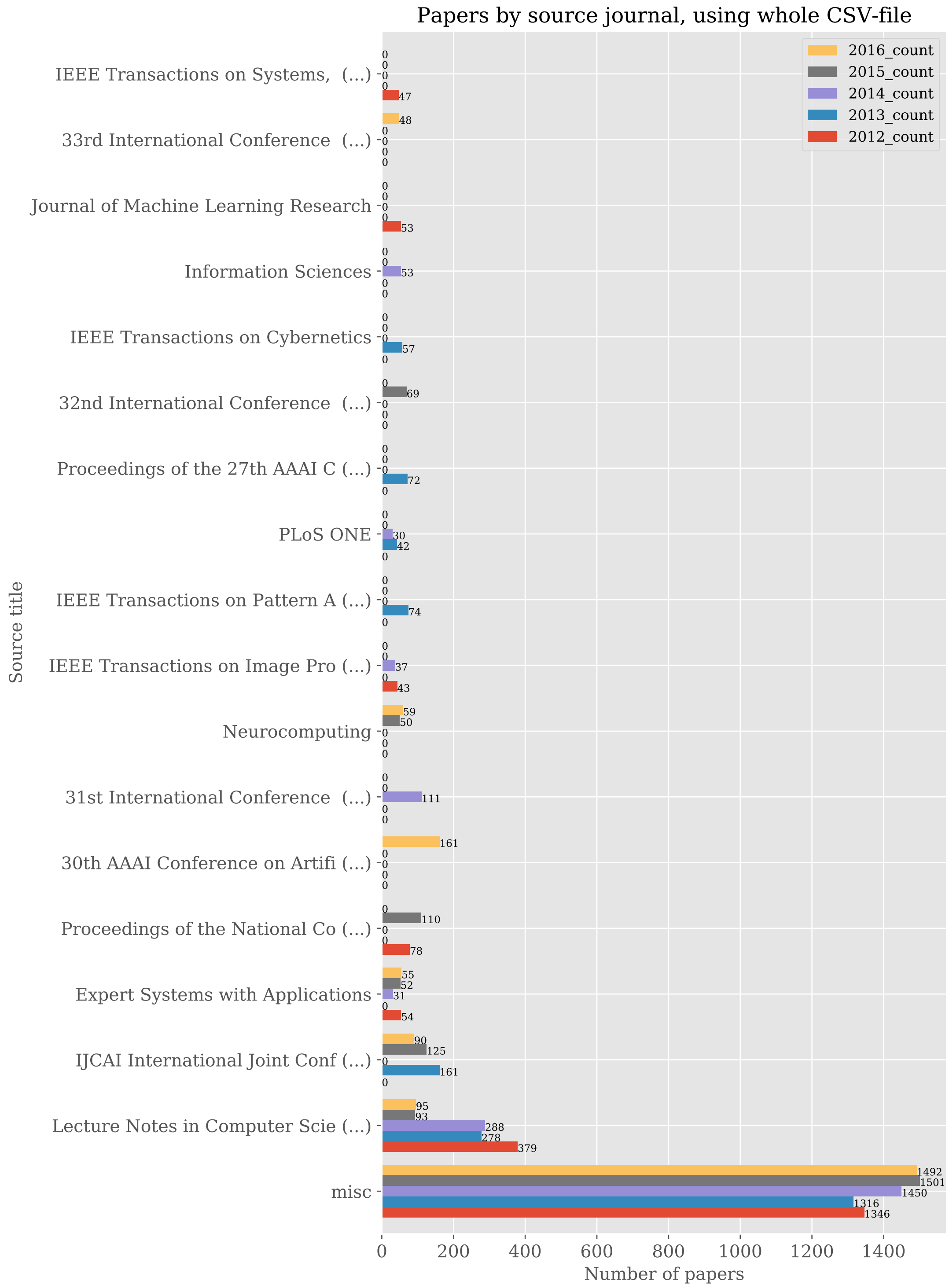


Figure 6.5: Proportion of papers per journal for the full set of papers, sorted by aggregate count for all years for each journal. Only the journals with the most papers are listed separately, the rest have been grouped into misc.

Generated_by: Appendix B: F)

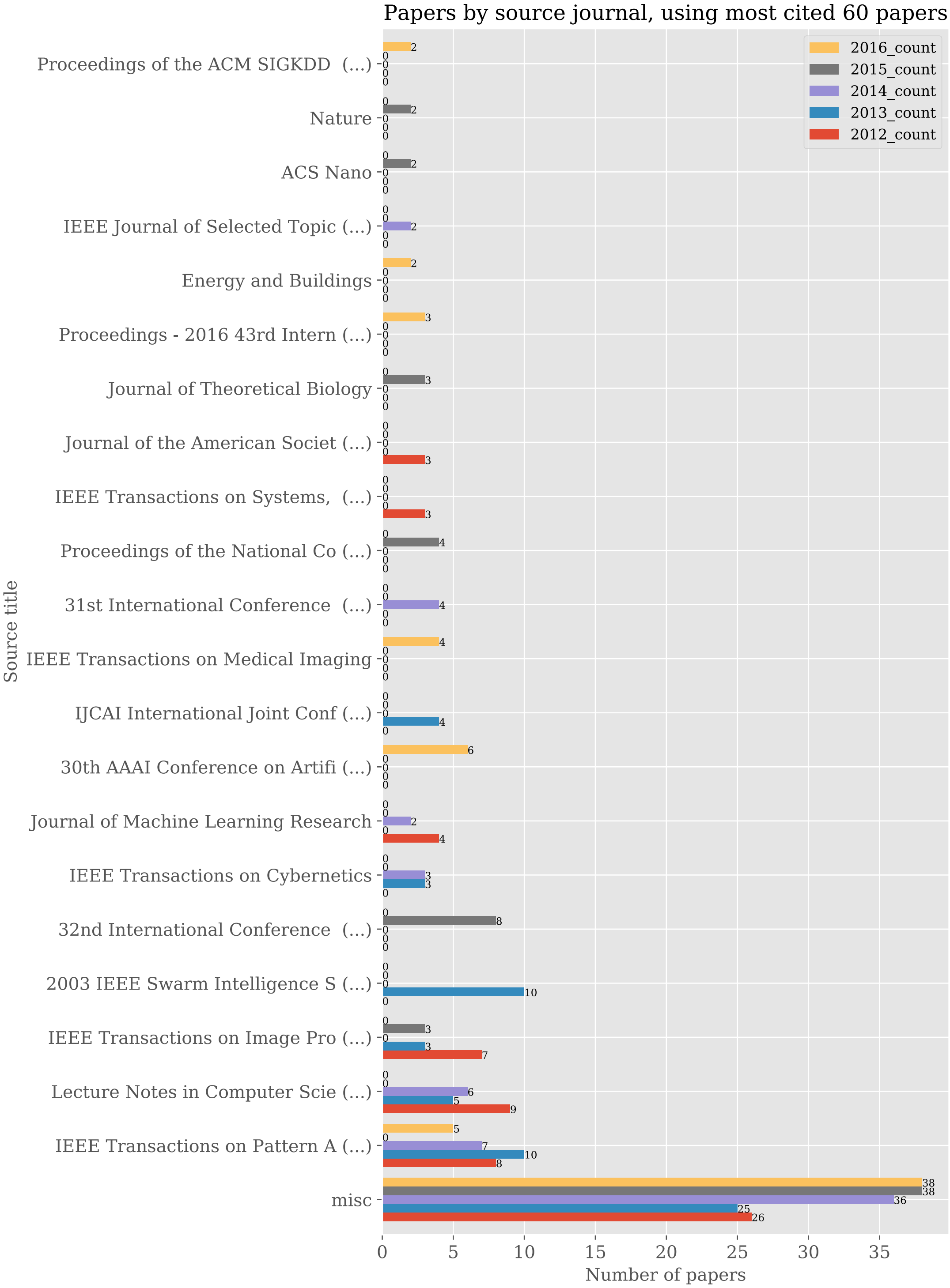


Figure 6.6: Proportion of papers per journal for the 60 first papers, sorted by aggregate count for all years for each journal. Only the journals with the most papers are listed separately, the rest have been grouped into misc.

Generated_by: Appendix B: F)

Table 6.4: Statistics for the proportion of papers for each journal. Count is the number of journals. Set of all papers.

	2012	2013	2014	2015	2016
count	541.000000	538.000000	621.000000	660.000000	666.000000
mean	3.696858	3.717472	3.220612	3.030303	3.003003
std	17.493823	15.122778	13.056501	8.775690	9.109525
min	1.000000	1.000000	1.000000	1.000000	1.000000
25%	1.000000	1.000000	1.000000	1.000000	1.000000
50%	1.000000	1.000000	1.000000	1.000000	1.000000
75%	2.000000	2.000000	2.000000	2.000000	2.000000
max	379.000000	278.000000	288.000000	125.000000	161.000000

Table 6.5: Statistics for the proportion of papers for each journal. Count is the number of journals. Set of first 60 papers.

	2012	2013	2014	2015	2016
count	28.000000	30.000000	38.000000	41.000000	42.000000
mean	2.142857	2.000000	1.578947	1.463415	1.428571
std	2.223016	2.392517	1.348304	1.246947	1.107466
min	1.000000	1.000000	1.000000	1.000000	1.000000
25%	1.000000	1.000000	1.000000	1.000000	1.000000
50%	1.000000	1.000000	1.000000	1.000000	1.000000
75%	2.000000	1.000000	1.750000	1.000000	1.000000
max	9.000000	10.000000	7.000000	8.000000	6.000000

Generated_by: Appendix B: F)

6.2.4 Keywords

One of the larger concerns that remain is whether the exported papers cover a wide portion of the field of artificial intelligence, or if they cover a very narrow area (for instance, almost purely neural networks). The research concerns artificial intelligence, and not just neural networks, after all (even if neural networks happen to be a large subfield). In the following, an ever shorter set of 15 of the top cited papers per year is used in addition to the full 2000 and the short 60. This is to investigate the state of the highest cited papers in the sample, in particular considering the steep drop-off in citation counts from section 6.2.2.

The first thing to note is that the occurences of the keyword "Artificial Intelligence" are not always 100% (even when all keywords have been converted to lowercase, as in the tables). One reasonable explanation for this is that perhaps the Index Keywords did not contain the term, while the Author Keywords did. This is not exclusively true, however, even in a merged table (treating both keyword fields as one field), the occurence is not 100% (see table 6.12).

The actual reason for this (or at least a reason that appears consistent) is that Scopus appears to use substring matching rather than exact matching, so a paper with the keyword "Artificial Intelligence Research" that does not have the keyword "Artificial Intelligence", will still be returned for a search for the latter, as was performed in our paper selection.

A second thing to note is that for the very short selection of papers (the first 15), see table 6.11 of author keywords, no one exact keyword appears for more than a handful of the 15 papers for each year. The few potential exceptions to this are the keywords "machine learning" (which could be considered rather broad), which occurs for 9 out of the 75, papers spread fairly evenly over the selection of years – and particle swarm optimization, which occurs 1 time in 2012 and 6 times in 2013 (which for 2013 is 40% of the 15 papers (though in return, the keyword does not appear again for the rest of the years)).

A concern with the above is that the keywords have not been normalized, that is to say "particle swarm optimization" and "particle swarm optimization (pso)" are considered separate keywords. An example of this is present in table 6.11, on lines 43, 44, and 45. Here, the three keywords "support vector machine[s,, (svm)]" are all represented, through at least two papers (as one occurence is from 2016, the others from 2014). This is not just a quirk of author keywords alone; table 6.8 for lines 59 and 60 (here too about support vector machines) show the same phenomenon.

A consequence of this is that the prevalence of a keyword may be undercounted by only looking at keywords in isolation, and consequently, it is not very reasonable to conclude that a diverse area is covered solely on the basis that each individual keyword only is used by a small proportion of the selection. It is nonetheless possible to conclude that none the individual exact keywords are overrepresented.

When extended to the 2000 samples (table 6.9), some trends in keywords start to emerge.

Table 6.6: Index keywords, all 2000 papers, top few most common keywords

	Keyword	2012	2013	2014	2015	2016	mean	total
1	artificial intelligence	1973.0	1975.0	1974.0	1990.0	1989.0	1980.2	9901.0
2	learning systems	99.0	138.0	518.0	967.0	1017.0	547.8	2739.0
3	algorithms	625.0	601.0	532.0	503.0	346.0	521.4	2607.0
4	article	578.0	582.0	381.0	211.0	172.0	384.8	1924.0
5	humans	362.0	372.0	357.0	199.0	119.0	281.8	1409.0
6	human	317.0	336.0	346.0	216.0	155.0	274.0	1370.0
7	algorithm	342.0	343.0	343.0	200.0	141.0	273.8	1369.0
8	optimization	146.0	145.0	148.0	231.0	309.0	195.8	979.0
9	learning algorithms	76.0	78.0	145.0	276.0	273.0	169.6	848.0
10	priority journal	174.0	173.0	154.0	116.0	130.0	149.4	747.0
11	pattern recognition, automated	245.0	229.0	176.0	74.0	17.0	148.2	741.0
12	decision support systems	175.0	149.0	129.0	136.0	135.0	144.8	724.0
13	classification (of information)	57.0	68.0	133.0	223.0	205.0	137.2	686.0
14	procedures	22.0	127.0	242.0	172.0	82.0	129.0	645.0
15	automated pattern recognition	186.0	185.0	164.0	80.0	20.0	127.0	635.0
16	neural networks	80.0	62.0	78.0	158.0	236.0	122.8	614.0
17	methodology	271.0	260.0	69.0	3.0	7.0	122.0	610.0
18	machine learning	75.0	105.0	119.0	142.0	150.0	118.2	591.0
19	data mining	59.0	93.0	95.0	172.0	160.0	115.8	579.0
20	support vector machines	87.0	62.0	151.0	146.0	127.0	114.6	573.0
21	forecasting	71.0	62.0	93.0	157.0	181.0	112.8	564.0
22	sensitivity and specificity	145.0	162.0	146.0	60.0	35.0	109.6	548.0
23	reproducibility of results	172.0	153.0	138.0	55.0	17.0	107.0	535.0
24	computer simulation	170.0	148.0	113.0	39.0	25.0	99.0	495.0
25	software engineering	130.0	116.0	81.0	41.0	53.0	84.2	421.0
26	female	93.0	98.0	120.0	66.0	43.0	84.0	420.0
27	computer assisted diagnosis	102.0	128.0	114.0	54.0	22.0	84.0	420.0
28	image interpretation, computer-assisted	114.0	151.0	104.0	36.0	14.0	83.8	419.0
29	controlled study	96.0	84.0	101.0	74.0	62.0	83.4	417.0
30	decision making	63.0	72.0	70.0	88.0	114.0	81.4	407.0
31	feature extraction	46.0	44.0	78.0	109.0	124.0	80.2	401.0
32	reproducibility	103.0	101.0	123.0	55.0	17.0	79.8	399.0
33	male	91.0	90.0	114.0	65.0	39.0	79.8	399.0
34	computer science	24.0	147.0	174.0	12.0	33.0	78.0	390.0
35	artificial neural network	94.0	69.0	74.0	69.0	77.0	76.6	383.0
36	support vector machine	66.0	57.0	96.0	84.0	72.0	75.0	375.0
37	semantics	57.0	71.0	56.0	86.0	92.0	72.4	362.0
38	image processing	82.0	56.0	72.0	78.0	64.0	70.4	352.0
39	pattern recognition	29.0	42.0	85.0	110.0	82.0	69.6	348.0
40	decision trees	37.0	42.0	53.0	124.0	90.0	69.2	346.0
41	computers	23.0	36.0	190.0	51.0	34.0	66.8	334.0
42	image enhancement	87.0	125.0	90.0	24.0	6.0	66.4	332.0
43	computer vision	68.0	39.0	66.0	78.0	79.0	66.0	330.0
44	particle swarm optimization (pso)	72.0	64.0	57.0	60.0	66.0	63.8	319.0
45	genetic algorithms	56.0	44.0	50.0	75.0	86.0	62.2	311.0
46	decision support system	98.0	70.0	49.0	46.0	47.0	62.0	310.0
47	evolutionary algorithms	39.0	64.0	55.0	68.0	81.0	61.4	307.0
48	regression analysis	36.0	48.0	71.0	65.0	81.0	60.2	301.0
49	adult	67.0	66.0	74.0	51.0	31.0	57.8	289.0
50	swarm intelligence	74.0	56.0	57.0	53.0	47.0	57.4	287.0
51	prediction	56.0	61.0	71.0	47.0	42.0	55.4	277.0
52	ant colony optimization	8.0	69.0	59.0	68.0	70.0	54.8	274.0
53	physiology	55.0	47.0	82.0	60.0	26.0	54.0	270.0
54	signal processing	36.0	36.0	72.0	59.0	48.0	50.2	251.0
55	machine learning techniques	8.0	16.0	55.0	81.0	90.0	50.0	250.0
56	social networking (online)	18.0	27.0	57.0	92.0	52.0	49.2	246.0
57	accuracy	69.0	56.0	53.0	42.0	25.0	49.0	245.0
58	iterative methods	22.0	38.0	48.0	68.0	62.0	47.6	238.0
59	neural networks (computer)	75.0	56.0	36.0	40.0	29.0	47.2	236.0
60	clustering algorithms	39.0	36.0	29.0	69.0	63.0	47.2	236.0
61	classification	53.0	43.0	58.0	51.0	31.0	47.2	236.0
62	animals	53.0	68.0	64.0	22.0	21.0	45.6	228.0
63	complex networks	14.0	24.0	34.0	65.0	90.0	45.4	227.0
64	benchmarking	38.0	32.0	36.0	52.0	68.0	45.2	226.0
65	statistical model	65.0	54.0	54.0	30.0	22.0	45.0	225.0
66	image segmentation	48.0	49.0	40.0	38.0	47.0	44.4	222.0
67	automation	47.0	43.0	38.0	53.0	39.0	44.0	220.0
68	diagnosis	28.0	24.0	34.0	69.0	62.0	43.4	217.0
69	software	61.0	54.0	50.0	21.0	21.0	41.4	207.0
70	image analysis	46.0	47.0	36.0	45.0	32.0	41.2	206.0
71	stochastic systems	17.0	15.0	41.0	63.0	68.0	40.8	204.0
72	state of the art	20.0	16.0	22.0	68.0	66.0	38.4	192.0
73	problem solving	20.0	25.0	35.0	58.0	54.0	38.4	192.0
74	fuzzy logic	50.0	49.0	30.0	35.0	28.0	38.4	192.0
75	aged	42.0	43.0	45.0	39.0	21.0	38.0	190.0
76	magnetic resonance imaging	46.0	44.0	54.0	28.0	15.0	37.4	187.0
77	brain	50.0	44.0	45.0	29.0	17.0	37.0	185.0
78	middle aged	44.0	43.0	44.0	30.0	19.0	36.0	180.0
79	information retrieval	41.0	44.0	39.0	33.0	23.0	36.0	180.0
80	ant colony optimization (aco)	58.0	48.0	21.0	27.0	26.0	36.0	180.0

Generated_by: Appendix B: F)

Table 6.7: Index keywords, first 60 papers, top few most common keywords

	Keyword	2012	2013	2014	2015	2016	mean	total
1	artificial intelligence	59.0	58.0	60.0	60.0	60.0	59.4	297.0
2	algorithms	29.0	25.0	12.0	16.0	16.0	19.6	98.0
3	article	28.0	21.0	16.0	12.0	10.0	17.4	87.0
4	learning systems	4.0	7.0	12.0	28.0	34.0	17.0	85.0
5	algorithm	22.0	17.0	11.0	12.0	13.0	15.0	75.0
6	human	19.0	13.0	14.0	14.0	14.0	14.8	74.0
7	humans	20.0	13.0	14.0	13.0	13.0	14.6	73.0
8	pattern recognition, automated	23.0	16.0	8.0	2.0	2.0	10.2	51.0
9	automated pattern recognition	19.0	15.0	8.0	2.0	3.0	9.4	47.0
10	procedures	3.0	6.0	9.0	16.0	10.0	8.8	44.0
11	methodology	23.0	15.0	5.0	0.0	0.0	8.6	43.0
12	neural networks	2.0	5.0	3.0	9.0	18.0	7.4	37.0
13	computer vision	5.0	6.0	9.0	7.0	7.0	6.8	34.0
14	image interpretation, computer-assisted	12.0	9.0	4.0	2.0	5.0	6.4	32.0
15	image processing	12.0	4.0	5.0	5.0	4.0	6.0	30.0
16	computer assisted diagnosis	9.0	8.0	5.0	2.0	6.0	6.0	30.0
17	reproducibility of results	12.0	6.0	5.0	3.0	3.0	5.8	29.0
18	optimization	6.0	9.0	4.0	5.0	5.0	5.8	29.0
19	priority journal	6.0	4.0	6.0	6.0	5.0	5.4	27.0
20	sensitivity and specificity	10.0	5.0	5.0	4.0	1.0	5.0	25.0
21	machine learning	2.0	2.0	5.0	5.0	11.0	5.0	25.0
22	learning algorithms	1.0	3.0	6.0	7.0	7.0	4.8	24.0
23	image processing, computer-assisted	10.0	4.0	3.0	5.0	2.0	4.8	24.0
24	image enhancement	11.0	6.0	4.0	1.0	1.0	4.6	23.0
25	reproducibility	8.0	4.0	4.0	3.0	3.0	4.4	22.0
26	particle swarm optimization (pso)	3.0	14.0	1.0	2.0	2.0	4.4	22.0
27	classification (of information)	0.0	5.0	8.0	6.0	3.0	4.4	22.0
28	support vector machines	2.0	1.0	5.0	6.0	5.0	3.8	19.0
29	signal processing	4.0	4.0	3.0	2.0	6.0	3.8	19.0
30	artificial neural network	4.0	1.0	2.0	3.0	9.0	3.8	19.0
31	data mining	0.0	4.0	4.0	4.0	5.0	3.4	17.0
32	convolutional neural network	1.0	0.0	1.0	4.0	11.0	3.4	17.0
33	computer simulation	8.0	4.0	3.0	1.0	1.0	3.4	17.0
34	support vector machine	2.0	2.0	1.0	6.0	5.0	3.2	16.0
35	physiology	3.0	3.0	3.0	5.0	2.0	3.2	16.0
36	pattern recognition	1.0	3.0	3.0	5.0	4.0	3.2	16.0
37	image segmentation	4.0	4.0	4.0	2.0	2.0	3.2	16.0
38	feature extraction	1.0	2.0	4.0	3.0	6.0	3.2	16.0
39	databases, factual	5.0	3.0	2.0	2.0	4.0	3.2	16.0
40	animals	4.0	4.0	5.0	2.0	1.0	3.2	16.0
41	factual database	4.0	3.0	2.0	2.0	4.0	3.0	15.0
42	subtraction technique	9.0	2.0	2.0	0.0	1.0	2.8	14.0
43	state of the art	1.0	1.0	0.0	7.0	5.0	2.8	14.0
44	semantics	1.0	1.0	4.0	5.0	3.0	2.8	14.0
45	models, theoretical	5.0	8.0	1.0	0.0	0.0	2.8	14.0
46	image analysis	4.0	0.0	3.0	3.0	4.0	2.8	14.0
47	evolutionary algorithms	2.0	9.0	0.0	2.0	1.0	2.8	14.0
48	biometry	2.0	3.0	4.0	3.0	2.0	2.8	14.0
49	neural networks (computer)	3.0	1.0	1.0	2.0	6.0	2.6	13.0
50	female	2.0	2.0	4.0	2.0	3.0	2.6	13.0
51	face	1.0	5.0	3.0	3.0	1.0	2.6	13.0
52	software engineering	8.0	3.0	0.0	0.0	1.0	2.4	12.0
53	nonhuman	3.0	0.0	5.0	4.0	0.0	2.4	12.0
54	magnetic resonance imaging	2.0	4.0	1.0	2.0	3.0	2.4	12.0
55	image classification	0.0	2.0	2.0	5.0	3.0	2.4	12.0
56	genetic algorithms	3.0	4.0	1.0	1.0	3.0	2.4	12.0
57	deep neural networks	0.0	0.0	2.0	6.0	4.0	2.4	12.0
58	deep learning	0.0	0.0	2.0	3.0	7.0	2.4	12.0
59	convolution	0.0	0.0	2.0	2.0	8.0	2.4	12.0
60	three dimensional imaging	6.0	0.0	2.0	1.0	2.0	2.2	11.0
61	theoretical model	5.0	6.0	0.0	0.0	0.0	2.2	11.0
62	nuclear magnetic resonance imaging	2.0	4.0	1.0	2.0	2.0	2.2	11.0
63	network architecture	1.0	0.0	2.0	3.0	5.0	2.2	11.0
64	male	3.0	2.0	3.0	1.0	2.0	2.2	11.0
65	image subtraction	6.0	2.0	2.0	0.0	1.0	2.2	11.0
66	controlled study	3.0	1.0	4.0	2.0	1.0	2.2	11.0
67	computers	2.0	1.0	6.0	2.0	0.0	2.2	11.0
68	computer science	0.0	5.0	6.0	0.0	0.0	2.2	11.0
69	big data	0.0	0.0	4.0	2.0	5.0	2.2	11.0
70	benchmarking	1.0	2.0	1.0	4.0	3.0	2.2	11.0
71	animal	3.0	2.0	3.0	2.0	1.0	2.2	11.0
72	state-of-the-art methods	1.0	0.0	3.0	3.0	3.0	2.0	10.0
73	signal processing, computer-assisted	3.0	4.0	2.0	0.0	1.0	2.0	10.0
74	problem solving	3.0	3.0	1.0	2.0	1.0	2.0	10.0
75	imaging, three-dimensional	7.0	0.0	1.0	0.0	2.0	2.0	10.0
76	diagnosis	1.0	1.0	1.0	4.0	3.0	2.0	10.0
77	video recording	4.0	2.0	0.0	2.0	1.0	1.8	9.0
78	swarm intelligence	2.0	2.0	2.0	2.0	1.0	1.8	9.0
79	regression analysis	3.0	1.0	2.0	1.0	2.0	1.8	9.0
80	optimization problems	3.0	1.0	3.0	2.0	0.0	1.8	9.0

Generated_by: Appendix B: F)

Table 6.8: Index keywords, first 15 papers, top few most common keywords

	Keyword	2012	2013	2014	2015	2016	mean	total
1	artificial intelligence	15.0	14.0	15.0	15.0	15.0	14.8	74.0
2	algorithms	11.0	8.0	3.0	4.0	5.0	6.2	31.0
3	article	11.0	5.0	4.0	5.0	4.0	5.8	29.0
4	learning systems	2.0	0.0	3.0	7.0	10.0	4.4	22.0
5	human	6.0	2.0	4.0	5.0	5.0	4.4	22.0
6	humans	6.0	2.0	4.0	5.0	4.0	4.2	21.0
7	algorithm	8.0	3.0	3.0	3.0	4.0	4.2	21.0
8	pattern recognition, automated	9.0	4.0	4.0	1.0	0.0	3.6	18.0
9	automated pattern recognition	6.0	4.0	4.0	1.0	1.0	3.2	16.0
10	priority journal	4.0	1.0	1.0	4.0	3.0	2.6	13.0
11	methodology	8.0	4.0	1.0	0.0	0.0	2.6	13.0
12	procedures	1.0	1.0	2.0	5.0	3.0	2.4	12.0
13	reproducibility of results	6.0	1.0	1.0	1.0	1.0	2.0	10.0
14	neural networks	1.0	2.0	0.0	3.0	4.0	2.0	10.0
15	image interpretation, computer-assisted	4.0	3.0	1.0	1.0	1.0	2.0	10.0
16	computer vision	1.0	1.0	3.0	2.0	3.0	2.0	10.0
17	particle swarm optimization (pso)	3.0	6.0	0.0	0.0	0.0	1.8	9.0
18	learning algorithms	1.0	0.0	2.0	3.0	3.0	1.8	9.0
19	sensitivity and specificity	5.0	0.0	1.0	2.0	0.0	1.6	8.0
20	computer assisted diagnosis	2.0	3.0	1.0	1.0	1.0	1.6	8.0
21	reproducibility	4.0	0.0	1.0	1.0	1.0	1.4	7.0
22	pattern recognition	1.0	1.0	2.0	1.0	2.0	1.4	7.0
23	optimization	3.0	3.0	0.0	0.0	1.0	1.4	7.0
24	machine learning	1.0	0.0	0.0	1.0	5.0	1.4	7.0
25	image enhancement	4.0	1.0	1.0	1.0	0.0	1.4	7.0
26	evolutionary algorithms	2.0	5.0	0.0	0.0	0.0	1.4	7.0
27	artificial neural network	1.0	0.0	1.0	1.0	4.0	1.4	7.0
28	signal processing	0.0	2.0	1.0	0.0	3.0	1.2	6.0
29	image processing	4.0	0.0	1.0	0.0	1.0	1.2	6.0
30	databases, factual	3.0	0.0	1.0	0.0	2.0	1.2	6.0
31	data mining	0.0	2.0	0.0	0.0	4.0	1.2	6.0
32	animals	3.0	1.0	0.0	1.0	1.0	1.2	6.0
33	subtraction technique	4.0	0.0	1.0	0.0	0.0	1.0	5.0
34	state of the art	0.0	0.0	0.0	4.0	1.0	1.0	5.0
35	regression analysis	2.0	1.0	1.0	0.0	1.0	1.0	5.0
36	prediction	1.0	1.0	1.0	1.0	1.0	1.0	5.0
37	network architecture	0.0	0.0	2.0	1.0	2.0	1.0	5.0
38	image processing, computer-assisted	4.0	0.0	0.0	0.0	1.0	1.0	5.0
39	image analysis	2.0	0.0	0.0	1.0	2.0	1.0	5.0
40	factual database	2.0	0.0	1.0	0.0	2.0	1.0	5.0
41	digital storage	0.0	1.0	0.0	3.0	1.0	1.0	5.0
42	convolutional neural network	1.0	0.0	0.0	1.0	3.0	1.0	5.0
43	convolution	0.0	0.0	1.0	1.0	3.0	1.0	5.0
44	three dimensional imaging	4.0	0.0	0.0	0.0	0.0	0.8	4.0
45	testing	0.0	4.0	0.0	0.0	0.0	0.8	4.0
46	semantics	0.0	0.0	2.0	1.0	1.0	0.8	4.0
47	remote sensing	0.0	0.0	2.0	0.0	2.0	0.8	4.0
48	models, theoretical	2.0	2.0	0.0	0.0	0.0	0.8	4.0
49	imaging, three-dimensional	4.0	0.0	0.0	0.0	0.0	0.8	4.0
50	image classification	0.0	1.0	0.0	1.0	2.0	0.8	4.0
51	feature extraction	0.0	0.0	1.0	0.0	3.0	0.8	4.0
52	deep learning	0.0	0.0	1.0	1.0	2.0	0.8	4.0
53	data handling	0.0	0.0	0.0	1.0	3.0	0.8	4.0
54	computer science	0.0	2.0	2.0	0.0	0.0	0.8	4.0
55	complex networks	0.0	0.0	1.0	1.0	2.0	0.8	4.0
56	clustering algorithms	0.0	2.0	2.0	0.0	0.0	0.8	4.0
57	animal	2.0	0.0	0.0	1.0	1.0	0.8	4.0
58	theoretical model	2.0	1.0	0.0	0.0	0.0	0.6	3.0
59	support vector machines	0.0	0.0	2.0	0.0	1.0	0.6	3.0
60	support vector machine	1.0	0.0	0.0	0.0	2.0	0.6	3.0
61	stochastic systems	0.0	3.0	0.0	0.0	0.0	0.6	3.0
62	statistics	1.0	2.0	0.0	0.0	0.0	0.6	3.0
63	state-of-the-art performance	0.0	0.0	1.0	1.0	1.0	0.6	3.0
64	state-of-the-art methods	0.0	0.0	2.0	1.0	0.0	0.6	3.0
65	signal processing, computer-assisted	0.0	2.0	0.0	0.0	1.0	0.6	3.0
66	sensor	0.0	0.0	2.0	0.0	1.0	0.6	3.0
67	random processes	0.0	3.0	0.0	0.0	0.0	0.6	3.0
68	protein	0.0	0.0	0.0	2.0	1.0	0.6	3.0
69	problem solving	1.0	2.0	0.0	0.0	0.0	0.6	3.0
70	principal component analysis	1.0	0.0	1.0	0.0	1.0	0.6	3.0
71	particle swarm	1.0	2.0	0.0	0.0	0.0	0.6	3.0
72	object recognition	1.0	0.0	1.0	0.0	1.0	0.6	3.0
73	nonhuman	1.0	0.0	1.0	1.0	0.0	0.6	3.0
74	neurons	0.0	1.0	2.0	0.0	0.0	0.6	3.0
75	neurology	1.0	0.0	0.0	2.0	0.0	0.6	3.0
76	neural networks (computer)	0.0	0.0	0.0	1.0	2.0	0.6	3.0
77	movement (physiology)	1.0	0.0	2.0	0.0	0.0	0.6	3.0
78	movement	1.0	0.0	2.0	0.0	0.0	0.6	3.0
79	mobile phone	0.0	0.0	1.0	1.0	1.0	0.6	3.0
80	knowledge acquisition	0.0	0.0	2.0	1.0	0.0	0.6	3.0

Generated_by: Appendix B: F)

Table 6.9: Author keywords, all 2000 papers, top few most common keywords

	Keyword	2012	2013	2014	2015	2016	mean	total
1	machine learning	51.0	69.0	193.0	272.0	307.0	178.4	892.0
2	artificial intelligence	65.0	47.0	78.0	44.0	55.0	57.8	289.0
3	ant colony optimization	57.0	67.0	38.0	40.0	34.0	47.2	236.0
4	classification	34.0	35.0	52.0	57.0	54.0	46.4	232.0
5	swarm intelligence	53.0	41.0	46.0	43.0	41.0	44.8	224.0
6	feature selection	20.0	21.0	41.0	52.0	45.0	35.8	179.0
7	particle swarm optimization	38.0	40.0	28.0	22.0	26.0	30.8	154.0
8	support vector machine	13.0	14.0	41.0	40.0	32.0	28.0	140.0
9	decision support system	35.0	29.0	27.0	17.0	20.0	25.6	128.0
10	neural networks	17.0	19.0	22.0	29.0	37.0	24.8	124.0
11	data mining	12.0	19.0	28.0	29.0	27.0	23.0	115.0
12	decision support systems	36.0	19.0	21.0	15.0	15.0	21.2	106.0
13	optimization	25.0	18.0	20.0	18.0	23.0	20.8	104.0
14	active learning	8.0	5.0	26.0	38.0	24.0	20.2	101.0
15	genetic algorithm	24.0	15.0	14.0	20.0	20.0	18.6	93.0
16	deep learning	0.0	1.0	11.0	30.0	43.0	17.0	85.0
17	support vector machines	14.0	15.0	22.0	15.0	16.0	16.4	82.0
18	clustering	16.0	10.0	18.0	15.0	12.0	14.2	71.0
19	computational intelligence	12.0	10.0	12.0	19.0	15.0	13.6	68.0
20	big data	1.0	6.0	12.0	22.0	22.0	12.6	63.0
21	artificial neural networks	9.0	11.0	6.0	13.0	24.0	12.6	63.0
22	pattern recognition	5.0	9.0	20.0	19.0	6.0	11.8	59.0
23	random forest	4.0	6.0	12.0	22.0	14.0	11.6	58.0
24	computer vision	13.0	10.0	12.0	13.0	10.0	11.6	58.0
25	feature extraction	9.0	11.0	13.0	11.0	13.0	11.4	57.0
26	cloud computing	14.0	18.0	6.0	11.0	8.0	11.4	57.0
27	neural network	8.0	7.0	13.0	5.0	23.0	11.2	56.0
28	fuzzy logic	9.0	11.0	13.0	12.0	8.0	10.6	53.0
29	artificial neural network	10.0	4.0	8.0	18.0	13.0	10.6	53.0
30	natural language processing	4.0	5.0	9.0	18.0	13.0	9.8	49.0
31	evolutionary computation	4.0	20.0	11.0	5.0	7.0	9.4	47.0
32	artificial bee colony	6.0	8.0	10.0	12.0	9.0	9.0	45.0
33	genetic algorithms	9.0	8.0	7.0	7.0	12.0	8.6	43.0
34	activity recognition	7.0	8.0	7.0	16.0	5.0	8.6	43.0
35	svm	8.0	5.0	10.0	11.0	8.0	8.4	42.0
36	segmentation	7.0	12.0	11.0	5.0	7.0	8.4	42.0
37	prediction	5.0	5.0	8.0	12.0	11.0	8.2	41.0
38	support vector machine (svm)	5.0	3.0	10.0	15.0	5.0	7.6	38.0
39	genetic programming	4.0	11.0	7.0	9.0	7.0	7.6	38.0
40	multi-objective optimization	6.0	11.0	6.0	8.0	6.0	7.4	37.0
41	image segmentation	5.0	14.0	5.0	5.0	8.0	7.4	37.0
42	extreme learning machine	1.0	4.0	7.0	13.0	12.0	7.4	37.0
43	evolutionary algorithms	6.0	8.0	10.0	5.0	8.0	7.4	37.0
44	dimensionality reduction	2.0	6.0	13.0	6.0	9.0	7.2	36.0
45	ambient intelligence	5.0	10.0	12.0	3.0	6.0	7.2	36.0
46	sentiment analysis	2.0	6.0	5.0	10.0	12.0	7.0	35.0
47	remote sensing	6.0	5.0	4.0	12.0	6.0	6.6	33.0
48	uncertainty	9.0	4.0	7.0	6.0	6.0	6.4	32.0
49	dictionary learning	7.0	2.0	11.0	6.0	6.0	6.4	32.0
50	simulation	13.0	8.0	4.0	2.0	4.0	6.2	31.0
51	artificial bee colony algorithm	7.0	7.0	6.0	6.0	5.0	6.2	31.0
52	ant colony optimization (aco)	8.0	6.0	4.0	7.0	6.0	6.2	31.0
53	anomaly detection	2.0	4.0	6.0	9.0	10.0	6.2	31.0
54	sparse representation	6.0	7.0	10.0	3.0	4.0	6.0	30.0
55	semi-supervised learning	6.0	3.0	8.0	7.0	6.0	6.0	30.0
56	ensemble learning	5.0	2.0	4.0	10.0	9.0	6.0	30.0
57	support vector regression	1.0	2.0	8.0	7.0	11.0	5.8	29.0
58	image processing	6.0	5.0	4.0	8.0	6.0	5.8	29.0
59	differential evolution	6.0	8.0	8.0	2.0	5.0	5.8	29.0
60	wireless sensor networks	6.0	3.0	3.0	7.0	9.0	5.6	28.0
61	unsupervised learning	6.0	3.0	7.0	6.0	6.0	5.6	28.0
62	bayesian networks	5.0	6.0	8.0	5.0	4.0	5.6	28.0
63	supervised learning	3.0	4.0	2.0	10.0	8.0	5.4	27.0
64	breast cancer	4.0	4.0	5.0	6.0	8.0	5.4	27.0
65	text mining	3.0	4.0	4.0	8.0	7.0	5.2	26.0
66	random forests	0.0	2.0	7.0	9.0	8.0	5.2	26.0
67	online learning	4.0	8.0	7.0	4.0	3.0	5.2	26.0
68	metaheuristics	2.0	7.0	4.0	5.0	8.0	5.2	26.0
69	face recognition	3.0	10.0	8.0	3.0	2.0	5.2	26.0
70	decision support	6.0	9.0	3.0	6.0	2.0	5.2	26.0
71	decision making	10.0	5.0	3.0	3.0	5.0	5.2	26.0
72	alzheimer's disease	3.0	6.0	7.0	9.0	1.0	5.2	26.0
73	logistic regression	5.0	4.0	6.0	4.0	6.0	5.0	25.0
74	image classification	5.0	6.0	1.0	5.0	7.0	4.8	24.0
75	fault diagnosis	2.0	2.0	4.0	8.0	8.0	4.8	24.0
76	twitter	1.0	2.0	6.0	6.0	8.0	4.6	23.0
77	scheduling	5.0	8.0	3.0	4.0	3.0	4.6	23.0
78	global optimization	11.0	3.0	3.0	3.0	3.0	4.6	23.0
79	gis	6.0	2.0	5.0	5.0	5.0	4.6	23.0
80	boosting	3.0	7.0	3.0	5.0	5.0	4.6	23.0

Generated_by: Appendix B: F)

Table 6.10: Author keywords, first 60 papers, top few most common keywords

	Keyword	2012	2013	2014	2015	2016	mean	total
1	machine learning	1.0	3.0	1.0	7.0	12.0	4.8	24.0
2	particle swarm optimization	1.0	11.0	1.0	2.0	1.0	3.2	16.0
3	deep learning	0.0	0.0	2.0	3.0	7.0	2.4	12.0
4	swarm intelligence	2.0	0.0	2.0	2.0	1.0	1.4	7.0
5	support vector machines	0.0	2.0	1.0	0.0	3.0	1.2	6.0
6	neural networks	1.0	3.0	1.0	0.0	1.0	1.2	6.0
7	neural network	0.0	0.0	1.0	1.0	4.0	1.2	6.0
8	feature selection	1.0	1.0	3.0	0.0	1.0	1.2	6.0
9	artificial neural networks	0.0	3.0	0.0	0.0	3.0	1.2	6.0
10	testing	0.0	5.0	0.0	0.0	0.0	1.0	5.0
11	sparse representation	3.0	1.0	1.0	0.0	0.0	1.0	5.0
12	segmentation	2.0	1.0	1.0	0.0	1.0	1.0	5.0
13	face recognition	0.0	2.0	1.0	1.0	1.0	1.0	5.0
14	evolutionary computation	0.0	5.0	0.0	0.0	0.0	1.0	5.0
15	support vector machine (svm)	1.0	0.0	1.0	2.0	0.0	0.8	4.0
16	stochastic processes	0.0	4.0	0.0	0.0	0.0	0.8	4.0
17	object detection	2.0	0.0	0.0	0.0	2.0	0.8	4.0
18	data mining	0.0	2.0	0.0	1.0	1.0	0.8	4.0
19	crowdsourcing	1.0	1.0	1.0	0.0	1.0	0.8	4.0
20	convolutional neural networks	1.0	0.0	0.0	1.0	2.0	0.8	4.0
21	classification	0.0	0.0	2.0	1.0	1.0	0.8	4.0
22	biometrics	0.0	1.0	1.0	2.0	0.0	0.8	4.0
23	artificial intelligence	1.0	1.0	2.0	0.0	0.0	0.8	4.0
24	artificial bee colony	1.0	1.0	1.0	1.0	0.0	0.8	4.0
25	alzheimer’s disease	0.0	1.0	1.0	1.0	1.0	0.8	4.0
26	support vector machine	0.0	0.0	1.0	2.0	0.0	0.6	3.0
27	pattern recognition	0.0	1.0	2.0	0.0	0.0	0.6	3.0
28	optimization	1.0	0.0	1.0	0.0	1.0	0.6	3.0
29	object recognition	1.0	1.0	1.0	0.0	0.0	0.6	3.0
30	neurons	0.0	1.0	0.0	2.0	0.0	0.6	3.0
31	mild cognitive impairment	0.0	1.0	1.0	0.0	1.0	0.6	3.0
32	image retrieval	1.0	0.0	1.0	1.0	0.0	0.6	3.0
33	genetic algorithms	0.0	2.0	1.0	0.0	0.0	0.6	3.0
34	gaussian processes	2.0	0.0	0.0	1.0	0.0	0.6	3.0
35	extreme learning machine (elm)	1.0	0.0	1.0	1.0	0.0	0.6	3.0
36	equations	0.0	3.0	0.0	0.0	0.0	0.6	3.0
37	dictionary learning	3.0	0.0	0.0	0.0	0.0	0.6	3.0
38	cloud computing	1.0	2.0	0.0	0.0	0.0	0.6	3.0
39	biomedical engineering	0.0	3.0	0.0	0.0	0.0	0.6	3.0
40	big data	0.0	2.0	0.0	0.0	1.0	0.6	3.0
41	ant colony optimization	0.0	1.0	0.0	1.0	1.0	0.6	3.0
42	unsupervised learning	0.0	0.0	2.0	0.0	0.0	0.4	2.0
43	transfer learning	0.0	0.0	1.0	1.0	0.0	0.4	2.0
44	topology	0.0	2.0	0.0	0.0	0.0	0.4	2.0
45	system dynamics	2.0	0.0	0.0	0.0	0.0	0.4	2.0
46	subspace learning	0.0	1.0	0.0	0.0	1.0	0.4	2.0
47	spectral clustering	0.0	1.0	0.0	1.0	0.0	0.4	2.0
48	social networks	0.0	1.0	0.0	0.0	1.0	0.4	2.0
49	sentiment analysis	0.0	2.0	0.0	0.0	0.0	0.4	2.0
50	sensor fusion	0.0	0.0	1.0	0.0	1.0	0.4	2.0
51	self-similarity	2.0	0.0	0.0	0.0	0.0	0.4	2.0
52	scene text detection	0.0	1.0	1.0	0.0	0.0	0.4	2.0
53	problem-solving	0.0	2.0	0.0	0.0	0.0	0.4	2.0
54	power engineering and energy	0.0	2.0	0.0	0.0	0.0	0.4	2.0
55	person re-identification	0.0	0.0	2.0	0.0	0.0	0.4	2.0
56	particle swarm optimization (pso)	1.0	1.0	0.0	0.0	0.0	0.4	2.0
57	optimization methods	0.0	2.0	0.0	0.0	0.0	0.4	2.0
58	opinion mining	0.0	1.0	1.0	0.0	0.0	0.4	2.0
59	object segmentation	1.0	0.0	1.0	0.0	0.0	0.4	2.0
60	non-local means	2.0	0.0	0.0	0.0	0.0	0.4	2.0
61	natural language processing	0.0	0.0	1.0	1.0	0.0	0.4	2.0
62	multilevel thresholding	0.0	1.0	1.0	0.0	0.0	0.4	2.0
63	multi-objective optimization	0.0	1.0	0.0	0.0	1.0	0.4	2.0
64	motion segmentation	0.0	1.0	1.0	0.0	0.0	0.4	2.0
65	microgrid	0.0	1.0	0.0	0.0	1.0	0.4	2.0
66	memristor	0.0	0.0	0.0	0.0	2.0	0.4	2.0
67	magnetic resonance imaging	0.0	1.0	0.0	1.0	0.0	0.4	2.0
68	large-scale learning	1.0	1.0	0.0	0.0	0.0	0.4	2.0
69	kapur’s entropy	0.0	1.0	1.0	0.0	0.0	0.4	2.0
70	intelligent fault diagnosis	0.0	0.0	0.0	0.0	2.0	0.4	2.0
71	image segmentation	0.0	1.0	1.0	0.0	0.0	0.4	2.0
72	image representation	0.0	0.0	1.0	1.0	0.0	0.4	2.0
73	image classification	0.0	2.0	0.0	0.0	0.0	0.4	2.0
74	hyperspectral data classification	0.0	0.0	1.0	1.0	0.0	0.4	2.0
75	genetic mutations	0.0	2.0	0.0	0.0	0.0	0.4	2.0
76	fusion	1.0	0.0	0.0	1.0	0.0	0.4	2.0
77	flexible electronics	0.0	0.0	0.0	1.0	1.0	0.4	2.0
78	feedforward neural networks	1.0	1.0	0.0	0.0	0.0	0.4	2.0
79	feature learning	0.0	0.0	0.0	1.0	1.0	0.4	2.0
80	extreme learning machine	0.0	0.0	1.0	1.0	0.0	0.4	2.0

Generated_by: Appendix B: F)

Table 6.11: Author keywords, first 15 papers, top few most common keywords

	Keyword	2012	2013	2014	2015	2016	mean	total
1	machine learning	1.0	1.0	0.0	3.0	4.0	1.8	9.0
2	particle swarm optimization	1.0	6.0	0.0	0.0	0.0	1.4	7.0
3	testing	0.0	4.0	0.0	0.0	0.0	0.8	4.0
4	evolutionary computation	0.0	4.0	0.0	0.0	0.0	0.8	4.0
5	deep learning	0.0	0.0	1.0	1.0	2.0	0.8	4.0
6	stochastic processes	0.0	3.0	0.0	0.0	0.0	0.6	3.0
7	big data	0.0	2.0	0.0	0.0	1.0	0.6	3.0
8	swarm intelligence	2.0	0.0	0.0	0.0	0.0	0.4	2.0
9	sparse representation	1.0	1.0	0.0	0.0	0.0	0.4	2.0
10	social networks	0.0	1.0	0.0	0.0	1.0	0.4	2.0
11	sensor fusion	0.0	0.0	1.0	0.0	1.0	0.4	2.0
12	self-similarity	2.0	0.0	0.0	0.0	0.0	0.4	2.0
13	problem-solving	0.0	2.0	0.0	0.0	0.0	0.4	2.0
14	pattern recognition	0.0	1.0	1.0	0.0	0.0	0.4	2.0
15	object detection	1.0	0.0	0.0	0.0	1.0	0.4	2.0
16	non-local means	2.0	0.0	0.0	0.0	0.0	0.4	2.0
17	neural networks	0.0	1.0	0.0	0.0	1.0	0.4	2.0
18	neural network	0.0	0.0	0.0	1.0	1.0	0.4	2.0
19	face recognition	0.0	0.0	1.0	0.0	1.0	0.4	2.0
20	extreme learning machine	0.0	0.0	1.0	1.0	0.0	0.4	2.0
21	data mining	0.0	1.0	0.0	0.0	1.0	0.4	2.0
22	convolutional neural networks	1.0	0.0	0.0	1.0	0.0	0.4	2.0
23	clustering	0.0	1.0	1.0	0.0	0.0	0.4	2.0
24	wearable sensors	0.0	0.0	0.0	0.0	1.0	0.2	1.0
25	wearable devices	0.0	0.0	1.0	0.0	0.0	0.2	1.0
26	wavelets transforms	0.0	0.0	0.0	1.0	0.0	0.2	1.0
27	voting system	0.0	0.0	0.0	1.0	0.0	0.2	1.0
28	visual saliency	1.0	0.0	0.0	0.0	0.0	0.2	1.0
29	vessel segmentation	0.0	0.0	0.0	1.0	0.0	0.2	1.0
30	venus	0.0	1.0	0.0	0.0	0.0	0.2	1.0
31	vaccination	0.0	0.0	0.0	1.0	0.0	0.2	1.0
32	unsupervised learning	0.0	0.0	1.0	0.0	0.0	0.2	1.0
33	unobtrusive sensing	0.0	0.0	1.0	0.0	0.0	0.2	1.0
34	transitions	0.0	0.0	0.0	0.0	1.0	0.2	1.0
35	trainable filters	0.0	0.0	0.0	1.0	0.0	0.2	1.0
36	traffic sign recognition	1.0	0.0	0.0	0.0	0.0	0.2	1.0
37	trace norm	0.0	1.0	0.0	0.0	0.0	0.2	1.0
38	topology adaptation	1.0	0.0	0.0	0.0	0.0	0.2	1.0
39	topology	0.0	1.0	0.0	0.0	0.0	0.2	1.0
40	tensor completion	0.0	1.0	0.0	0.0	0.0	0.2	1.0
41	telemedicine	0.0	0.0	0.0	1.0	0.0	0.2	1.0
42	target cross-validation	0.0	0.0	0.0	0.0	1.0	0.2	1.0
43	support vector machines	0.0	0.0	0.0	0.0	1.0	0.2	1.0
44	support vector machine (svm)	0.0	0.0	1.0	0.0	0.0	0.2	1.0
45	support vector machine	0.0	0.0	1.0	0.0	0.0	0.2	1.0
46	subspaces	0.0	1.0	0.0	0.0	0.0	0.2	1.0
47	subspace clustering	0.0	1.0	0.0	0.0	0.0	0.2	1.0
48	strips	0.0	1.0	0.0	0.0	0.0	0.2	1.0
49	steering kernel regression	1.0	0.0	0.0	0.0	0.0	0.2	1.0
50	statistics	0.0	1.0	0.0	0.0	0.0	0.2	1.0
51	stacked autoencoder (sae)	0.0	0.0	1.0	0.0	0.0	0.2	1.0
52	spectral clustering	0.0	1.0	0.0	0.0	0.0	0.2	1.0
53	spatio-temporal descriptors	1.0	0.0	0.0	0.0	0.0	0.2	1.0
54	sparse learning	0.0	1.0	0.0	0.0	0.0	0.2	1.0
55	sparse errors	1.0	0.0	0.0	0.0	0.0	0.2	1.0
56	spark	0.0	0.0	0.0	1.0	0.0	0.2	1.0
57	social-based frameworks and applications	0.0	0.0	0.0	0.0	1.0	0.2	1.0
58	social media	0.0	0.0	0.0	0.0	1.0	0.2	1.0
59	smartphones	0.0	0.0	0.0	0.0	1.0	0.2	1.0
60	smartphone	0.0	0.0	0.0	1.0	0.0	0.2	1.0
61	single-link clustering	0.0	0.0	1.0	0.0	0.0	0.2	1.0
62	simulated annealing	0.0	1.0	0.0	0.0	0.0	0.2	1.0
63	short-term forecasting	0.0	1.0	0.0	0.0	0.0	0.2	1.0
64	service science	0.0	1.0	0.0	0.0	0.0	0.2	1.0
65	service orientation	0.0	1.0	0.0	0.0	0.0	0.2	1.0
66	sequence-coupling model	0.0	0.0	0.0	0.0	1.0	0.2	1.0
67	sentiment analysis	0.0	1.0	0.0	0.0	0.0	0.2	1.0
68	semi-supervised learning	0.0	0.0	1.0	0.0	0.0	0.2	1.0
69	semi-auto image tagging	0.0	1.0	0.0	0.0	0.0	0.2	1.0
70	self-organization	1.0	0.0	0.0	0.0	0.0	0.2	1.0
71	self-learning particle swarm optimizer (slpso)	1.0	0.0	0.0	0.0	0.0	0.2	1.0
72	segmentation	0.0	1.0	0.0	0.0	0.0	0.2	1.0
73	search methods	0.0	1.0	0.0	0.0	0.0	0.2	1.0
74	scene text detection	0.0	0.0	1.0	0.0	0.0	0.2	1.0
75	scalable machine learning	0.0	0.0	0.0	0.0	1.0	0.2	1.0
76	rotation-invariant cnn (ricnn)	0.0	0.0	0.0	0.0	1.0	0.2	1.0
77	rotating machinery	0.0	0.0	0.0	0.0	1.0	0.2	1.0
78	robust principal component analysis	1.0	0.0	0.0	0.0	0.0	0.2	1.0
79	retinal image analysis	0.0	0.0	0.0	1.0	0.0	0.2	1.0
80	remote sensing images	0.0	0.0	0.0	0.0	1.0	0.2	1.0

Generated_by: Appendix B: F)

Table 6.12: Both Author keywords and Index keywords, All 2000 papers, top selection of most common keywords

	Keyword	2012	2013	2014	2015	2016	mean	total
1	artificial intelligence	1984.0	1982.0	1988.0	1999.0	1998.0	1990.2	9951.0
2	learning systems	99.0	138.0	518.0	967.0	1017.0	547.8	2739.0
3	algorithms	625.0	601.0	533.0	503.0	348.0	522.0	2610.0
4	article	578.0	582.0	381.0	211.0	172.0	384.8	1924.0
5	humans	362.0	372.0	357.0	199.0	119.0	281.8	1409.0
6	algorithm	343.0	344.0	344.0	200.0	141.0	274.4	1372.0
7	human	317.0	336.0	346.0	216.0	155.0	274.0	1370.0
8	machine learning	116.0	150.0	283.0	369.0	396.0	262.8	1314.0
9	optimization	147.0	146.0	148.0	231.0	309.0	196.2	981.0
10	learning algorithms	76.0	78.0	145.0	276.0	273.0	169.6	848.0
11	priority journal	174.0	173.0	154.0	116.0	130.0	149.4	747.0
12	pattern recognition, automated	245.0	229.0	176.0	74.0	17.0	148.2	741.0
13	decision support systems	175.0	150.0	129.0	136.0	135.0	145.0	725.0
14	classification (of information)	57.0	68.0	133.0	223.0	205.0	137.2	686.0
15	procedures	22.0	127.0	242.0	172.0	82.0	129.0	645.0
16	automated pattern recognition	186.0	185.0	164.0	80.0	20.0	127.0	635.0
17	neural networks	82.0	65.0	79.0	158.0	237.0	124.2	621.0
18	methodology	271.0	260.0	70.0	3.0	7.0	122.2	611.0
19	support vector machines	90.0	68.0	154.0	148.0	127.0	117.4	587.0
20	data mining	60.0	94.0	95.0	172.0	160.0	116.2	581.0
21	forecasting	71.0	62.0	93.0	157.0	181.0	112.8	564.0
22	sensitivity and specificity	145.0	162.0	146.0	60.0	35.0	109.6	548.0
23	reproducibility of results	172.0	153.0	138.0	55.0	17.0	107.0	535.0
24	computer simulation	170.0	148.0	113.0	39.0	25.0	99.0	495.0
25	support vector machine	74.0	66.0	131.0	110.0	96.0	95.4	477.0
26	classification	81.0	70.0	99.0	100.0	81.0	86.2	431.0
27	software engineering	130.0	116.0	81.0	41.0	53.0	84.2	421.0
28	female	93.0	98.0	120.0	66.0	43.0	84.0	420.0
29	computer assisted diagnosis	102.0	128.0	114.0	54.0	22.0	84.0	420.0
30	image interpretation, computer-assisted	114.0	151.0	104.0	36.0	14.0	83.8	419.0
31	controlled study	96.0	84.0	101.0	74.0	62.0	83.4	417.0
32	artificial neural network	99.0	71.0	76.0	79.0	86.0	82.2	411.0
33	decision making	63.0	72.0	70.0	88.0	114.0	81.4	407.0
34	feature extraction	46.0	45.0	78.0	109.0	124.0	80.4	402.0
35	decision support system	122.0	92.0	67.0	59.0	60.0	80.0	400.0
36	reproducibility	103.0	101.0	123.0	55.0	17.0	79.8	399.0
37	male	91.0	90.0	114.0	65.0	39.0	79.8	399.0
38	computer science	24.0	147.0	174.0	12.0	33.0	78.0	390.0
39	semantics	57.0	71.0	56.0	86.0	92.0	72.4	362.0
40	image processing	84.0	56.0	73.0	78.0	64.0	71.0	355.0
41	pattern recognition	29.0	43.0	88.0	111.0	82.0	70.6	353.0
42	decision trees	37.0	43.0	53.0	124.0	90.0	69.4	347.0
43	ant colony optimization	61.0	87.0	59.0	68.0	70.0	69.0	345.0
44	computer vision	72.0	39.0	67.0	79.0	80.0	67.4	337.0
45	computers	23.0	36.0	190.0	51.0	34.0	66.8	334.0
46	image enhancement	87.0	125.0	90.0	24.0	6.0	66.4	332.0
47	particle swarm optimization (pso)	74.0	64.0	57.0	60.0	66.0	64.2	321.0
48	genetic algorithms	56.0	44.0	50.0	77.0	87.0	62.8	314.0
49	evolutionary algorithms	40.0	64.0	55.0	69.0	81.0	61.8	309.0
50	prediction	57.0	63.0	77.0	58.0	53.0	61.6	308.0
51	regression analysis	36.0	48.0	71.0	65.0	81.0	60.2	301.0
52	adult	67.0	66.0	74.0	51.0	31.0	57.8	289.0
53	swarm intelligence	75.0	56.0	57.0	53.0	47.0	57.6	288.0
54	physiology	55.0	47.0	82.0	60.0	26.0	54.0	270.0
55	signal processing	36.0	36.0	72.0	60.0	48.0	50.4	252.0
56	machine learning techniques	8.0	17.0	56.0	81.0	90.0	50.4	252.0
57	social networking (online)	18.0	27.0	57.0	92.0	52.0	49.2	246.0
58	accuracy	69.0	57.0	53.0	42.0	25.0	49.2	246.0
59	iterative methods	22.0	38.0	48.0	68.0	62.0	47.6	238.0
60	neural networks (computer)	75.0	56.0	36.0	40.0	29.0	47.2	236.0
61	clustering algorithms	39.0	36.0	29.0	69.0	63.0	47.2	236.0
62	animals	53.0	68.0	64.0	22.0	21.0	45.6	228.0
63	complex networks	14.0	24.0	34.0	65.0	90.0	45.4	227.0
64	image segmentation	48.0	51.0	41.0	39.0	47.0	45.2	226.0
65	benchmarking	38.0	32.0	36.0	52.0	68.0	45.2	226.0
66	statistical model	65.0	54.0	54.0	30.0	22.0	45.0	225.0
67	diagnosis	31.0	25.0	34.0	70.0	62.0	44.4	222.0
68	automation	48.0	43.0	38.0	53.0	39.0	44.2	221.0
69	software	62.0	54.0	51.0	22.0	24.0	42.6	213.0
70	image analysis	46.0	47.0	36.0	45.0	32.0	41.2	206.0
71	stochastic systems	17.0	15.0	41.0	63.0	68.0	40.8	204.0
72	fuzzy logic	50.0	50.0	32.0	36.0	28.0	39.2	196.0
73	state of the art	20.0	16.0	22.0	68.0	66.0	38.4	192.0
74	problem solving	20.0	25.0	35.0	58.0	54.0	38.4	192.0
75	magnetic resonance imaging	46.0	45.0	55.0	29.0	15.0	38.0	190.0
76	aged	42.0	43.0	45.0	39.0	21.0	38.0	190.0
77	big data	1.0	7.0	39.0	64.0	75.0	37.2	186.0
78	brain	50.0	44.0	45.0	29.0	17.0	37.0	185.0
79	feature selection	20.0	22.0	41.0	52.0	46.0	36.2	181.0
80	ant colony optimization (aco)	59.0	48.0	21.0	27.0	26.0	36.2	181.0

Generated_by: Appendix B: F)

It can be thought that this is because with 15 or 60 papers, the sample size is so small that an author is likely to have their keyword appear higher up on the list if the author adds a somewhat uncommon keyword or even a slightly different spelling of a common keyword (and even moreso if every author does this). In the table of 2000 papers, several rather distinct subfields are present – support vector machines, swarm intelligence, clustering, neural networks, and genetic algorithms to name a few. All of these except clustering are represented at least 3 times in table 6.10 of the top 60 author keywords.

6.3 Evaluation

Having analysed the citation counts of the selected papers as well as other attributes like the source (journal) and keywords, the part of selection hypothesis **HYP-sel 1** concerning the papers being of higher impact is supported, as long as the used selection does not extend towards the lower end of the first 60 papers for each year and beyond.

Papers are represented from a fair number of journals, where a slight trend in the prevalence of keywords of interest to determining which subfields are covered is starting to emerge in the selection of the first 60 papers, the trend being rather apparent when extended to the 2000 papers per year. The evidence supporting this part of the hypothesis weak when limited to the first 15 papers from each year, in large part due to noise. It appears to be difficult to use author-supplied and index-supplied keywords to say something about the diversity of topics covered when the sample size gets smaller, though at the same time, there is no strong evidence that a subfield is grossly overrepresented either.

Chapter 7

Research Method – Reproduction

This chapter covers the process of reproducing the selection of papers that resulted from the method described in chapter 5, using the framework presented by the collaborating research group of Odd Cappelen and Martin Mølnå, as well as the survey presented in [GK18].

7.1 Methodology

In order to address the research questions that pertain to reproducibility, an observational study in the form of a survey ([Oat06, pp. 93-94]) has been devised, which is fed in the data generated by the paper selection process. Two rather similar surveys are used – one made by the collaborating group, and one presented in [GK18] (which the collaborating group's survey builds on).

The surveys cover the collection of data regarding the documentation level of the publications. The attempted reproduction of a publication – as well as the identification of its reproducibility – does not seem as easily described by just the research strategy of surveys (in the sense of the same kinds of data being gathered systematically), as it is not entirely clear what data should be gathered for each attempt, until the attempt has been made and some idea of the difficulties encountered is had (even if the methodology for each paper is systematic). There is also the aspect where the research is studying research, where that research has its own research methods for its experiments – and attempting to reproduce the results may involve in some sense repeating the experiment of the paper (although the purpose of performing the experiment is in our case not related to the phenomenon the publication relates to).

Instead of attempting to force the methodology to be describable by e.g an overall strategy, the process is described in the following.

7.2 Process

7.2.1 Identifying R1

The paper selection process resulted in a set of 2000 papers per year, in our case covering the years from 2012 up to and including 2016. These papers are sorted by citation count, the measure by which the papers were selected.

The initial target was to identify R1[1] papers, and attempt reproducing their results. R2 and R3 papers were identified as well, though were set aside. In addition, a few papers with no experimental results (e.g theoretical papers and review papers) present in the selection of papers were identified as such, and as reproducing (i.e proving) theoretical results is outside our scope, they were skipped.

This was done for year 2012 at first – though after having identified the reproducibility level of 10-15 papers – was expanded to cover years 2014 and 2016 as well. The reason for this was to ensure that not only a single year was represented, and also to have an evenly spaced selection of years (2012, 2014, 2016), so that it would be possible to draw conclusions about how (or whether) documentation has changed over time.

As our groups had covered 10-15 papers and were about to extend the coverage to years 2013 and 2015, it was decided that due to the low number of papers identified as R1, most being identified as R2, it would be appropriate to extend the coverage of the reproduction attempt to cover R2 papers as well. As R2 papers require less documentation, being more decoupled from the exact experiments they present, they provide higher generality, and may provide information about reproducibility than R1 papers alone (which are essentially repeating the experiments on different hardware, perhaps with other slightly different variables). R3 papers were still concluded to require too much effort per paper, as a new dataset would have to be procured.

7.2.2 Reproduction of R1 and R2

For the process of reproducing the results of the R2 papers (as well as the R1 papers), a hard cutoff-limit of 40 hours per paper was applied. This means that once 40 hours of work has been put into a paper, any results obtained so far are “final” (a possible outcome is also that no results are obtained, because a running implementation of the method and experiment has not been achieved within the time limit). It should be noted that this is 40 hours of active effort; a paper that presents a method that requires training, where the training for instance runs for a week on a GPU, is not timed when the implementation has been invoked and runs in the background – after all, it is possible to spend time working on other papers

[1] See chapter 3 for the definition of the degrees of reproducibility.

while this implementation runs.

In addition to the hard time limit, other criteria for when to stop working on a paper are introduced via the process of Odd Cappelen and Martin Mølnåas well – some of them formalizing the process of leaving out papers that do not fit an identification of R1, R2, or R3 – and some related to the documentation presented. A selection of the criteria (from their autumn 2017 project report) are repeated here for reference, with shortened descriptions:

1. Article unrelated to query (i.e, not related to the field of AI).
2. No experiments presented.
3. **Only** qualitative results are presented (these are hard to determine whether have been reproduced).
4. Data sets unavailable [for R1 and R2-D].
5. Required resources unavailable [to us] (e.g if a paper uses a very large cluster of computers).
6. Required hyper-parameter values not provided (e.g if the value is central to the results, and searching for the "right" value would take a lot of effort).

An additional stopping-criteria relating to the experiment description was later introduced. This covers papers where an experiment is presented, but the experiment is described in such a way that it is very difficult or impossible for independent researchers to perform the same experiment (that is to say, the reproduction would have to take guesses at what the experiment was – in a somewhat blunt manner of speech, devise the experiment to test the method on its own).

The survey was populated with information from the process of reproducing each paper, other issues encountered being noted down and discussed, for use in e.g revising the set of analysis methods to use. This includes a selection of identified categories pertaining to problems encountered, assumptions made, and likely causes of error where the achieved results were different from the results achieved in the paper. The details of this are covered in chapter 8, on the results of the reproduction attempts.

Tables 7.1, 7.2, and 7.3 list the papers along with the citation for all the 30 papers that were covered.

Table 7.1: The selection of papers from 2012 that was covered

Title	Cited by	Level	Researcher	Citation
Context-aware saliency detection	592	R2-D	MM	[GZMT12]
A modified Artificial Bee Colony algorithm for real-parameter optimization	456	R2-D	NN	[AK12]
Measuring the objectness of image windows	424	R1	OC	[ADF12]
Blind image quality assessment: A natural scene statistics approach in the DCT domain	415	R1	OC	[SBC12]
RASL: Robust alignment by sparse and low-rank decomposition for linearly correlated images	292	R1	NN	[Pen+12]
Cooperatively coevolving particle swarms for large scale optimization	287	R2-D	OC	[LY12]
Learning sparse representations for human action recognition	186	R2-D	OC	[GW12]
Single image super-resolution with non-local means and steering kernel regression	175	R3	OC	[Zha+12]
Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease	173	R3	OC	[ZSI+12]
Development and investigation of efficient artificial bee colony algorithm for numerical function optimization	163	R2-D	MM	[LNX12]

Generated_by: Appendix B: E)

Table 7.2: The selection of papers from 2014 that was covered

Title	Cited by	Level	Researcher	Citation
Visualizing and understanding convolutional networks	767	R2-D	OC	[ZF14]
Clustering by fast search and find of density peaks	545	R1	MM	[RL14]
Distributed representations of sentences and documents	266	R2-D	MM	[LM14]
DeCAF: A deep convolutional activation feature for generic visual recognition	257	R2-D	MM	[Don+14]
DeepReID: Deep filter pairing neural network for person re-identification	178	R2-D	NN	[Li+14]
Robust text detection in natural scene images	177	R3	NN	[Yin+14]
Deep learning-based classification of hyperspectral data	171	R1	NN	[Che+14]
Semi-supervised and unsupervised extreme learning machines	158	R2-D	NN	[Hua+14]
Towards end-to-end speech recognition with recurrent neural networks	140	R3	MM	[GJ14]
Facial landmark detection by deep multi-task learning	132	R2-D	NN	[Zha+14]

Generated_by: Appendix B: E)

Table 7.3: The selection of papers from 2016 that was covered

Title	Cited by	Level	Researcher	Citation
Mastering the game of Go with deep neural networks and tree search	571	R3	MM	[Sil+16]
Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning	152	R3	OC	[Shi+16]
MLlib: Machine learning in Apache Spark	98	R3	OC	[Men+16]
XGBoost: A scalable tree boosting system	97	R1	MM	[CG16]
Deep neural networks: A promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data	77	R2-D	NN	[Jia+16a]
ISuc-PseOpt: Identifying lysine succinylation sites in proteins by incorporating sequence-coupling effects into pseudo components and optimizing imbalanced training dataset	71	R2-D	OC	[Jia+16b]
Classification with Noisy Labels by Importance Reweighting	69	R2-D	OC	[LT16]
Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition	52	R1	OC	[OR16]
Generalized Correntropy for Robust newline ?Adaptive Filtering	49	R2-D	MM	[Che+16]
Learning Rotation-Invariant Convolutional Neural Networks for Object Detection in VHR Optical Remote Sensing Images	45	R3	MM	[CZH16]

Generated_by: Appendix B: E)

Chapter 8

Results

This chapter presents the results of attempting to reproduce the selection of papers, and presents some identified categories of problems, assumptions made, and possible sources of errors.

8.1 Papers covered

The selection of papers from the process described in chapter 5 is rather substantial, spanning 2000 papers for each of the years from 2012 up to and including 2016. As was alluded to in chapter 7, only a tiny portion of these have been covered by our research groups.

The total number of papers (only counting papers identified as R1, R2-D, R2-M, or R3) covered per year towards the end of the project was not evenly distributed, though the counts lie in the range $[10, 15]$. It was decided to use the same number of papers per year, and as the counts were all fairly close to 10 (when rounded down), the final selection to finalize evaluations of and include was set to the first 10 (experimental) papers per year, for the years 2012, 2014, and 2016. The final selection of papers is shown in tables 7.1, 7.2, and 7.3, along with a label of which individual ended up covering the paper.

Some papers were revieved by several members of the research groups. One reason why this was done was that the person who went through the paper (in a getting-an-overview-sense), noticed it required access to specific hardware that our groups were in the process of getting access to. When we did get access, the person who had started on the paper was busy with another paper, while a member involved in getting access to the hardware had the opportunity to attempt it, so we agreed to transfer the "responsibility" of that paper (along with the time already spent on it, although the time was never substantial in any of these cases). In those cases, the person with the final (and in our cases larger) responsibility for that paper is the one who is listed in the table.

8.2 Identifying Categories

Other than the proposed metrics (via the framework, and the metric that is proposed in [GK18]), it was difficult to beforehand (that is, before the reproductions were attempted) decide on what to do with the results from going through with the process. This is particularly true as the research process evolved throughout the course of the project, as aspects that were not previously considered surfaced. Nonetheless, as the results of more papers were attempted reproduced, some commonalities in what made that paper difficult started to appear. After having gathered and documented 10 papers for each year (30 in total, only counting R* ones), we went through several iterations of describing what our issues (if any) with reproducing the results were, and to generalize them into categories that might apply to several papers.

The result of this process was three types of categories:

- **Problem:** which problems were encountered during the attempt?
- **Assumption:** which assumptions were made, due to missing or ambiguous information in the published documentation?
- **Error:** if the result from the reproduction attempt is different from the result presented in the paper, what are the more likely causes for this, in our belief?

These categories are in turn covered below, with the larger contribution to the following generalized problem, assumption, and error-category formulations being attributed to the collaborating research group:

8.2.1 Problem categories

Some of the problem categories from table 8.2 may be ambiguous from their shorter description, or may otherwise benefit from further motivation:

- P6: the code not being inspectable. Essentially covers closed source executable code (e.g binaries or protected matlab files). There is not really any way to know whether these implement the method that the paper presents.
- P10: Going from the abstract overview that papers tend to present their method in to an actual implementation often requires certain assumptions about choices that were glossed over. If the number of assumptions needed start to add up, or if significant assumptions are needed, the belief in how accurately the presented method has been implemented decreases. An example of this is [Li+14], where negative training examples are gradually increased as training progresses, but the details around how to implement this are left out.

Table 8.1: The selection of (R1, R2-D) papers for reproduction. The initials indicate who had the responsibility for that paper. NN is the author of this thesis, Nicklas Grimstad Nilsen, while OC and MM are the authors of the collaborating thesis, Odd Cappelen and Martin Mølnå.

ID	Article	Year	R1/R2	Overall outcome	Researcher
1	Measuring the objectness of image windows	2012	R1	Partial success	OC
2	Generalized Correntropy for Robust Adaptive Filtering	2016	R2	Partial success	MM
3	Development and investigation of efficient artificial bee colony algorithm for numerical function optimization	2012	R2	Partial success	MM
4	Blind Image Quality Assessment	2012	R1	Partial success	OC
5	Cooperatively Coevolving particle swarm optimization for large scale optimization	2012	R2	Partial success	OC
6	Learning sparse representations for human action recognition	2012	R2	Failure	OC
7	Visualizing and understanding convolutional networks	2014	R2	No Result	OC
8	iSuc-PseOpt: Identifying lysine succinylation sites in proteins by incorporating sequence-coupling effects into pseudo components and optimizing imbalanced training dataset	2016	R2	Partial success	OC
9	A modified Artificial Bee Colony algorithm for real-parameter optimization	2012	R2	Partial success	NN
10	RASL: Robust alignment by sparse and low-rank decomposition for linearly correlated images	2012	R1	Failure	NN
11	Classification with noisy labels by importance reweighting	2016	R2	Failure	OC
12	Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition	2016	R1	Partial success	OC
13	Context Aware Saliency Detection	2012	R2	Failure	MM
14	Distributed representations of sentences and documents	2014	R2	No Result	MM
15	XGBoost: A scalable tree boosting system	2016	R1	Failure	MM
16	Facial landmark detection by deep multi-task learning	2014	R2	No Result	NN
17	Deep learning-based classification of hyperspectral data	2014	R1	Failure	NN
18	Semi-supervised and unsupervised extreme learning machines	2014	R2	Failure	NN
19	DeepReID: Deep Filter Pairing Neural Network for Person Re-Identification	2014	R2	No Result	NN
20	Deep neural networks: A promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data	2016	R2	No Result	NN
21	Clustering by fast search and find of density peaks	2014	R1	Success	MM
22	DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition	2014	R2	Failure	MM

Generated_by: Appendix B: E)

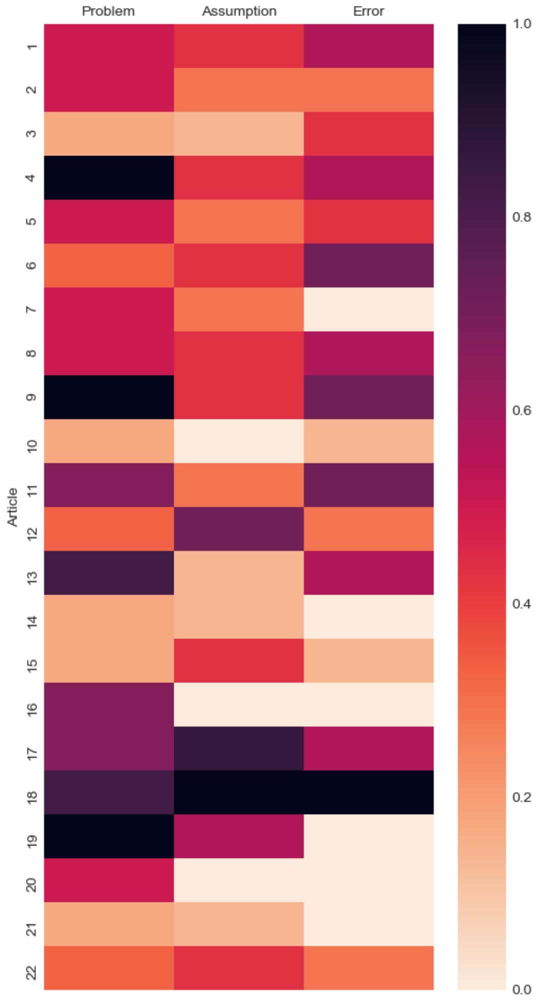


Figure 8.1: A heatmap where each column represents one of the three category groups identified. Each row is one of the papers that were attempted reproduced, where the value (colour) in the cell is the proportion of possible categories in that column that apply to the paper. Each column has been normalized separately, so that if there are 100 possible problems and 50 possible errors, a paper with 90 problems and 45 errors will have the same colour in those two columns.

Generated_by: Appendix B: E)

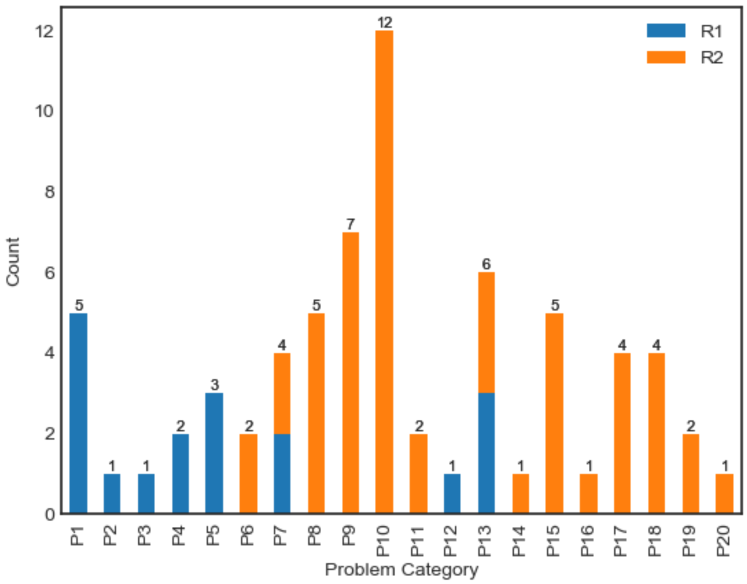


Figure 8.2: A stacked bar-chart covering each of the problem categories from table 8.2, and how many times it applied to either an R1 paper or an R2 paper.

Generated_by: Appendix B: E)

- P11: The paper [Hua+14] presents the method in a rather abstract way, providing suggestions for how to e.g normalize a matrix, but without actually stating which method was used to produce their own results.

- P14: The paper [AK12] has various errors in the tables, e.g inserted additonal symbols, and parameter values being shifted so that it is hard to tell which parameters were used for which results.

- P15: The paper [Jia+16a] uses a dataset with health conditions for machinery, but the way the dataset is described in the paper (as having 200 signals with 2400 data points), does not match the dataset that is available online, and it is not clear what the paper did to make the dataset into that representation.

For R1-papers, the most common issue was that the code for running one or more experiments was not shared, making it difficult to perform the R1 reproduction level for the paper (as experiments have to be reimplemented). Another issue is versioning, where the authors may have made several versions of the code available (through e.g GitHub), or a version has been made available but it is not certain whether that was the version they used, or if the version is a new implementation from a later point in time.

Table 8.2: The problem categories that were identified, and the indices of the papers that each are considered to apply to. The paper corresponding to each index is listed in table 8.1

Category Id	Problem Category	Papers
P1	For R1 study, the experiment code is not shared, or the experiment code does not cover all experiments	1, 4, 12, 10, 17
P2	For R1 study, the method code does not cover the entire method as described in the paper	4
P3	For R1 study, the code is poorly documented and difficult to interpret	4
P4	For R1 study, the parameter values shared with the code are not complete, or differ from the values given in the paper	12, 17
P5	For R1 study, code was not versioned, or the paper did not state which version was used in experiments	1, 4, 17
P6	An implementation of the method or experiment is shared, but the code is not inspectable.	13, 16
P7	Random numbers are used in a significant way, but the numbers, random number generator, and random seed are not shared	1, 4, 5, 9
P8	An aspect of the method is not described, or described in a manner difficult to understand	5, 9, 18, 19, 22
P9	An aspect of the experiment is not described, or described in a manner difficult to understand	2, 7, 11, 13, 9, 19, 20
P10	The implementation of the method, or an aspect of it, is not described, or difficult to understand	2, 3, 5, 7, 8, 11, 14, 9, 16, 18, 19, 22
P11	Multiple methods or implementations are suggested, but which variation is used is not stated	8, 18
P12	Trained weights or trained parameters shared online are not the same as was used in original experiments	4
P13	Not all parameter or hyper-parameters needed are given	6, 9, 15, 17, 18, 21
P14	There is a possible error in the paper	9
P15	There is a mismatch between a data set as described in the paper and as available online	6, 13, 16, 19, 20
P16	A necessary subset of a data set is not shared	13
P17	Augmented or pre-processed data set is not shared, and the method for pre-processing and data augmentation is not clearly described	8, 11, 16, 19
P18	Partition of data into training, validation, and test set is not shared, and the method for performing the partition is not clearly described	11, 18, 19, 20
P19	Results are presented in a manner unsuitable for reproduction	2, 13
P20	Significant resource demands (hardware or software) make reproduction complicated	7

Generated_by: Appendix B: E)

For R2 papers, the most common issue is by far that their description of how the AI method or experiment is implemented/performed is unclear or not described at all. The more assumptions that has to be made, the more chances there are for various errors or differences between interpretations or understandings to occur. An example of this is [Li+14], where the various training strategies are only summarily described. They are described well enough for other researchers to use the training strategy themselves in their own research, but not well enough that an independent researcher would be able to implement the strategy same same way; we could conceivably have gone through with spending the rest of the allotted time for the paper to attempt implementing the strategies, using assumptions where unclear, but the number of assumptions required for how to run the experiment became so numerous that it was decided it would fall under the stopping criteria for the experiment not being documented well enough (see section 7.2.2).

Datasets are another large problem. They can be difficult to understand how to use (sometimes so difficult that they were the reason reproduction was not achiveved, as in [Jia+16a]), other times the dataset appears to have been changed since it was used in the paper, as appears to be the case with [Che+14] – the number of training samples per category when running the experiment code being different from what is presented in the paper, even with the random seed that the authors provided set.

Random seeds are another difficult issue, as they can sometimes be difficult or less meaningful to share (e.g for methods with other sources of nondeterminism than the random number generator, for instance in cases of multiprocessing), though for research where the experiment code is released where setting a random seed causes results to be consistent across runs, it can be very helpful in removing variables for why the result might be different. In our implementation of the Artificial Bee Colony optimization presented in [AK12], when the program was invoked without a random seed set, the results were wildly different between our own invocations (far beyond what a Welch t-test would indicate). With a seed set, the results between runs were the same, though still significantly different from the results presented in the paper (the only result somewhat consistent with the results paper was for the simplest function (sphere), hence the paper is labeled as "partial success" in table 8.1).

8.2.2 Assumption Categories

Similarly to in the subsection on problem categories, here some of the assumption categories from table 8.3 are elaborated on:

- A3: This concerns for instance [Che+14] (an R1 paper with experiment code available as well), where the paper mentions the code runs faster on a GPU, but can be run on a CPU if desirable, though it does not appear to run on a GPU out of the box. A few

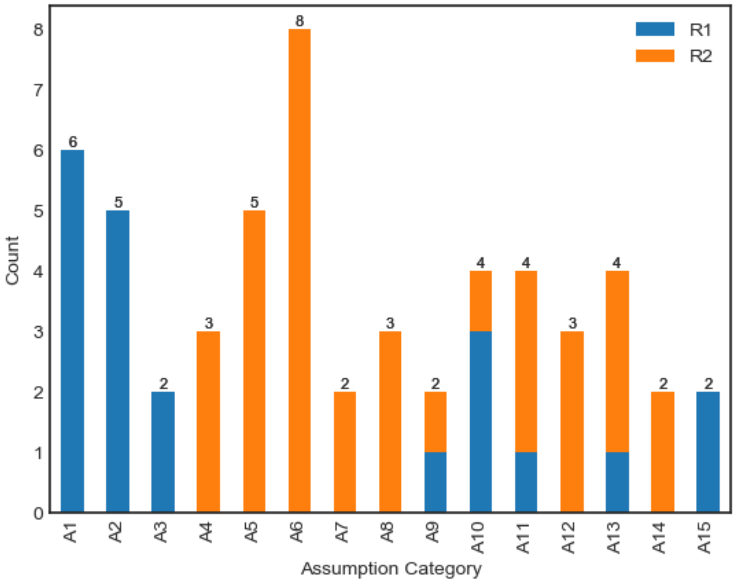


Figure 8.3: A stacked bar-chart covering each of the assumption categories from table 8.3, and how many times it applied to either an R1 paper or an R2 paper.

Generated_by: Appendix B: E)

Table 8.3: The assumption categories that were identified, and the indices of the papers that each are considered to apply to. The paper corresponding to each index is listed in table 8.1

Category Id	Assumption category	Papers
A1	For R1 study, the code available is assumed to be the same as was used in the original experiments	1, 4, 12, 15, 21, 17
A2	For R1 study, the parameter values in available code are assumed to be the same as was used in the original experiments	1, 4, 12, 15, 17
A3	For R1 study, a minor change to the code to facilitate running on our hardware is assumed to not affect results	12, 17
A4	An assumption is made about how to interpret a term or concept which is ambigous	2, 6, 18
A5	An assumption is made about how to treat an aspect of the method which is not well described, based on how that aspect is treated in another paper	5, 7, 14, 18, 19
A6	An assumption is made about how to treat an aspect of the method which is not well described, but this assumption is not based on how that aspect is treated in another paper	2, 3, 5, 8, 19, 18, 9, 22
A7	For aspects where multiple methods are suggested, an assumption is made about which to use	8, 18
A8	A third party implementation of a method is assumed to be similar to the original implementation	7, 11, 18
A9	When using a third party library or framework used in the original experiment, it is assumed that the version used in the reproduction can produce the same results as the version used in the original experiment	17, 22
A10	Trained weights or trained parameters shared online are assumed to be the same as was used in the original experiments	1, 4, 12, 22
A11	Assumption is made about one or more parameter values	6, 17, 18, 9
A12	Augmented or pre-processed data set is assumed to be equivalent to original data set used in experiment, even if it is not identical	8, 11, 19
A13	An assumption is made about how to partition a data set	6, 15, 19, 18
A14	An assumption is made about the usage of a data set	13, 9
A15	The hardware is assumed to not significantly influence the results of the experiment	12, 17

minor changes were made in the code to make it utilize the GPU we had available.

- A5,A6,A7: In [Hua+14], the construction of the graph laplacian has assumptions, both based on other papers and separately. The paper is rarely explicit on how it was done for the results presented, but rather suggests many ways things can be done.
- A8: e.g by using a library for K-nearest neighbours, where the original method says it uses this method – the assumption being that the implementations of K-nearest are equivalent.
- A15: This mostly applies when the hardware is specialized, e.g using GPUs for training neural networks.

A lot of the assumption categories stem from the problems that we identified as having in attempting to reproduce the results of the papers. One of the papers covered that necessitated a lot of assumptions is [Hua+14]. It follows the trend of describing the method in a way that somebody may be able to implement it for a separate problem, though not enough that the experiment can easily be reproduced. It should also be noted that our achieved results from running our implementation perform poorly enough that there likely are unknown errors in the implementation.

On the topic of assuming the version of the code that was retrieved is the same as was used in the paper, another related issue presented itself: For at least one paper, code was found online (where there was no reference to such supplementary materials in the paper), but we were unable to determine whether the author of the code was the same author as the author of the paper. Looking through license files and following links to the homepage of the code author (on a university web server) did not lead to any author or website-owner names.[1] Running code that is not beyond reasonable doubt verifiably written by one or more of the paper authors is not within our scope of R1-papers, though it could perhaps fall under the scope of R2, although it was not something we decided to do.

8.2.3 Error Categories

Similarly to in the subsection on problem categories, and on assumption categories, here some of the error categories from table 8.4 are elaborated on:

- E2, E3, and E4 cover the cases where our assumptions that we had reason to believe were fair, perhaps were not as fair as we at first thought.

[1] The paper that this concerned is [Hua+14], the repository that was found being https://github.com/ExtremeLearningMachines/SS-US-ELM

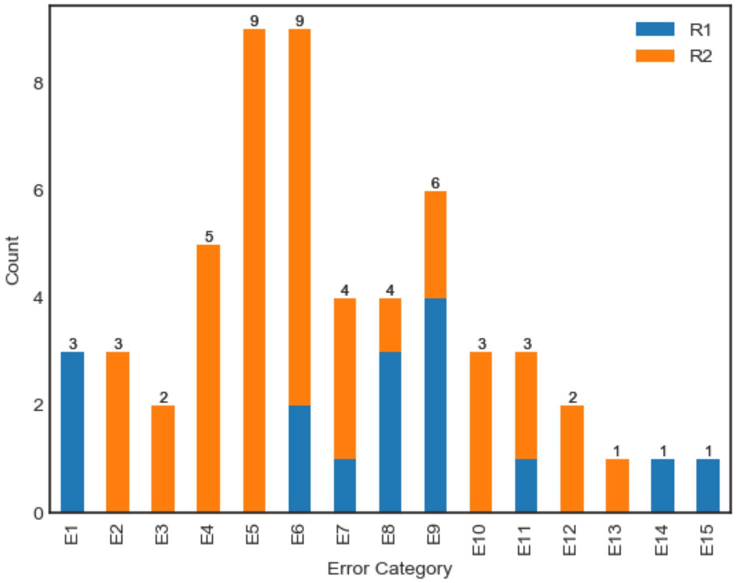


Figure 8.4: A stacked bar-chart covering each of the error categories from table 8.4, and how many times it applied to either an R1 paper or an R2 paper.

Generated_by: Appendix B: E)

Table 8.4: The error categories that were identified, and the indices of the papers that each are considered to apply to. The paper corresponding to each index is listed in table 8.1

Category Id	Error Category	Papers
E1	For R1 study, the code available is not exactly the same as was used in the original experiments	1, 4, 17
E2	There are errors in our assumptions about the method	5, 18, 9
E3	There are errors in our assumptions about the experiment	6, 18
E4	There are errors in our assumptions about the implementation	3, 5, 13, 18, 9
E5	There are unknown errors in our implementation of the method	2, 3, 5, 6, 8, 13, 18, 9, 22
E6	There are unknown errors in our implementation of the experiment	1, 2, 3, 4, 6, 8, 13, 18, 9
E7	An implementation in a third party library is not equivalent to the implementation used in the original experiment	6, 11, 17, 18
E8	The trained weights or trained parameters shared are not the same as was used in the original experiment	1, 4, 12, 22
E9	The parameters are not the same as was used in the original experiment	1, 4, 6, 15, 18, 17
E10	The randomness in the method or experiment, and the lack of shared random numbers and random number generator, influences the result	8, 11, 9
E11	Augmented and pre-processed data set used in reproduction is not equivalent to the data set used in original experiment	8, 11, 17
E12	The data subset used in reproduction is not the same as was used in original experiment	11, 13
E13	The partitioning of data into training, validation, and test set is not the same as was used in the original experiment	11
E14	Differences in hardware influenced the result	12
E15	The reproduced results are difficult to compare to the original results	10

- E5 and E6 cover unknown errors in our implementation. This can be anything from a wrong sign in a mathematical expression to errors in the way datasets are read in and used.

  These are, by definition, fairly hard to give examples of that have not been fixed before trying to run the implementation again; if the fix really fixed the issue, then a paper would likely not be listed here.

- E12: In case a method only uses a portion of a dataset, e.g if a dataset is very large and the paper decides to only select 10% of the elements, but does not say which 10%.

- E15: The paper [Pen+12] presents a facial alignment method, and has published code for some of their experiments. Not all of these experiments have ground truths, and are thus qualitative (as mentioned in the paper). The paper does present some quantitative results for a dataset with ground truths, though the experiment code provided does not produce these quantitative values, and it is difficult to say whether the output of the code is the same as is presented in the paper (as the output is qualitative here too, though this time quantitative results should be attainable, although it is not clear how to make the code achieve this).

Errors E2-E6 are somewhat related, pertaining to the implementation of the method and experiment. Of these, E4, E5, and E6 are the most common, indicating that the method and experiment (E2 and E3) appear to be described reasonably well, so that that a researcher can understand the method and the experiment – but not well enough that the step from method to implementation (E4) is achievable without sizable room for errors in the necessary assumptions. Furthermore, a lack of experience within a field may increase the chances of unknown implementation errors (E5, E6).

Likewise related, errors E8 and E9 (which concern parameters and weights) have fairly commonly been identified as likely to be the cause of the error, pertaining to the parameters and weights used. Even if a hyperparameter is not provided (or not provided explicitly), it might still be a hyperparameter with a range of values that it is reasonable to estimate. Even so, it is not unlikely that different parameters may affect the performance of the method, especially if the researcher implementing the method has a limited amount of time to try out different parameters.

8.3 Metrics from AAAI-Paper

The other research group has developed a framework based on the survey and metric presented in [GK18]. This newly developed metric could conceivably have been used in this

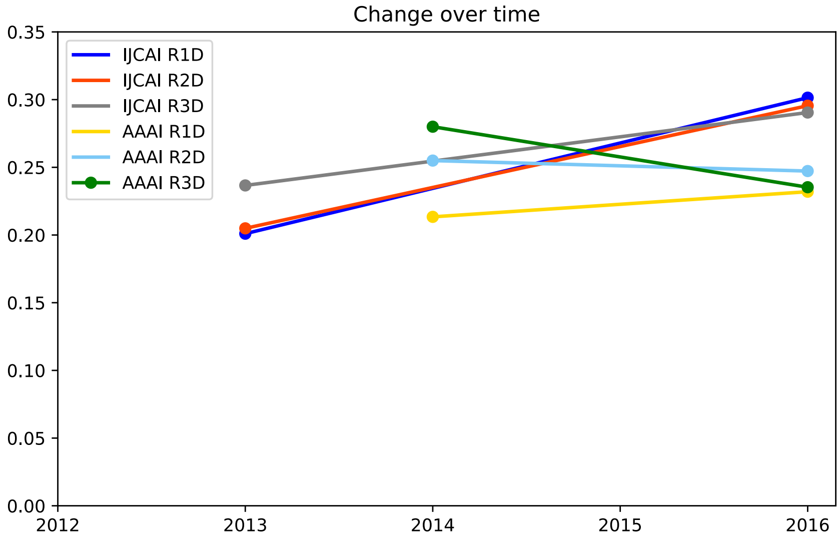


Figure 8.5: The reproducibility metric score presented in AAAI over time (duplicated here for reference), used on the papers from the conferences presented in the same paper, repeated here for reference.

Generated_by: Appendix B: D)

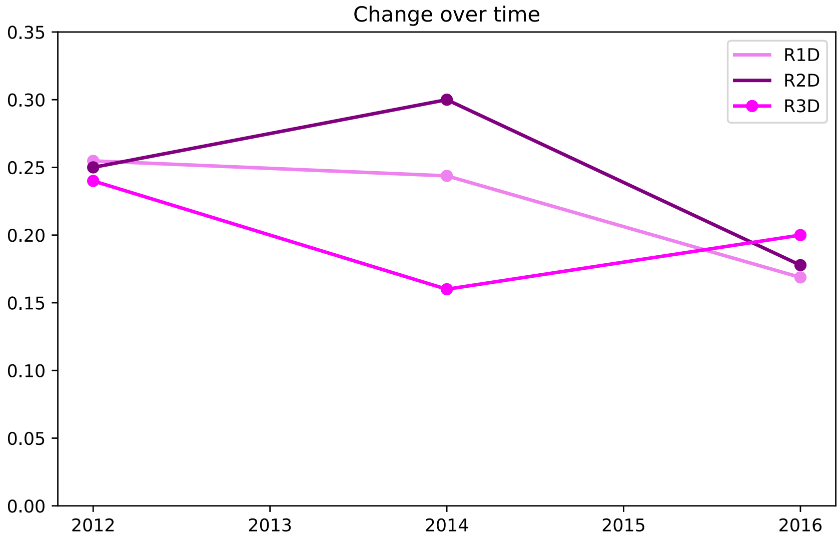


Figure 8.6: The reproducibility metric score presented in AAAI over time, used on our selection of 30 papers, grouped by year.

Generated_by: Appendix B: D)

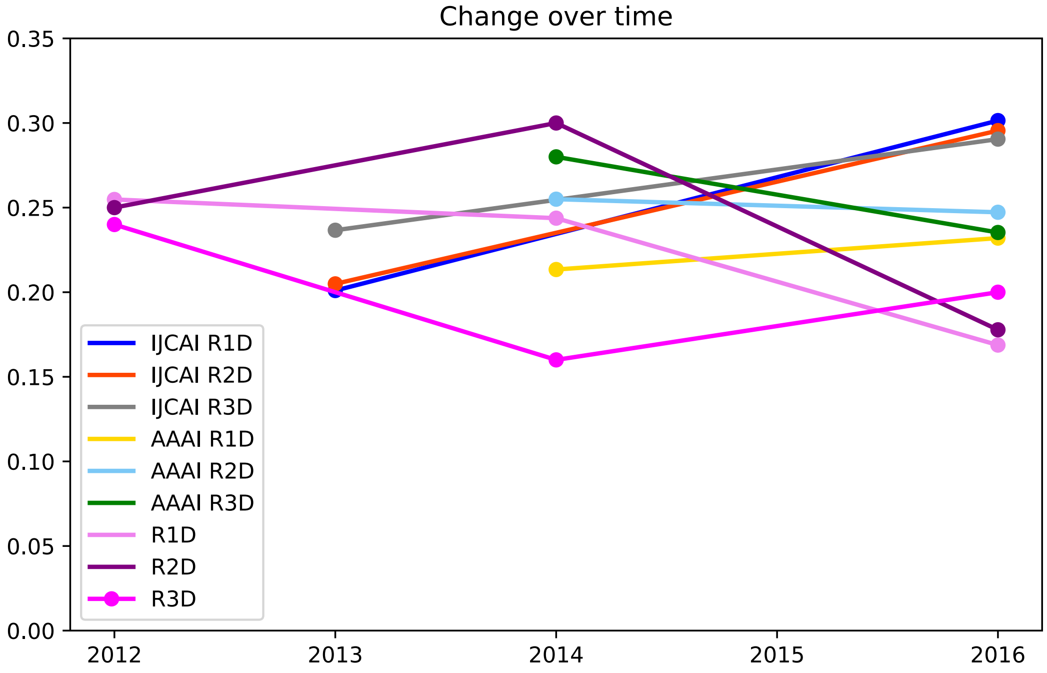


Figure 8.7: The data presented in figures 8.5 and 8.6, here plotted together.

Generated_by: Appendix B: D)

thesis, though as it is already presented in the thesis of the collaborating group, it appears more beneficial to apply the metric presented in the AAAI paper here. One advantage of this is that the AAAI paper, which covers two conferences at two different points in time each, presents a development of the metric over time, grouped by the conference for each year. The goal of this was to give an idea of whether documentation practices have improved over time. By repeating such an analysis here, results can be compared for e.g consistency.

The survey used to generate the data in the AAAI paper has a lot of similarities with the newly developed survey. Additionally, this newly developed survey (while not presented in detail here) has been used during the reproduction attempts to gather the data. As such, it would both save time and limit the possiblity of accidental inconsistencies to use as much of the data gathered in the new survey as possible when entering it in the original survey. See appendix A for a description of how the fields in the survey from [GK18] and [Kje17] were determined, based on the values collected by the newer survey. Nonetheless, it should be noted that the field for evaluation criteria have been left blank due to the description of when to set it being hard to understand. Additionally, the field for experiment setup has been treated the same way as in their presentation (i.e, a check for descriptions of hyper parameters).

There does not appear to be any consistency in the change in reproducibility scores over time, except for year 2016 consistently being worse on all three measures (R1D, R2D, R3D). For year 2014, R2D has one of the highest scores out of all the datapoints for all years, while simultaneously for 2014, R3D has the worst score of all datapoints for any year. Compared to the scores presented in for the IJCAI-samples, the scores for the 30 papers we have surveyed show no such upwards trend. This may well be because our 10 samples per year is a very tiny sample.

Chapter 9

Discussion

9.1 Paper Selection, in Light of Reproductions

The selection of papers, having been made with respect to citation count, does cover a certain breadth of subfields within artifical intelligence; neural networks (deep and convolutional, shallow variants, and other variations like extreme learning machines) are certainly well-represented, though research on other topics like swarm intelligence, decision trees, statistical methods for signal processing, and other new algorithms have some representation as well. Nonetheless, neural networks are present in some regard in the larger portion of the 22 papers covered. This makes sense when one takes into consideration that neural networks and deep learning have seen a rise in popularity this decade, though it still makes the representation skewed.

A potential issue with the method the papers was selected by, is that it does not result in a random sample; the method used to generate it non-probabilistic purposive-sampling [Oat06, pp. 96-98], with our purpose being to obtain a selection of highly cited papers. Because the sample is not random from the population of all Artifical Intelligence-papers, it is not possible to generalize from our observations to that of all papers in the field either – any generalizations apply to the most highly cited papers.

The previous raises another concerning issue – by only giving exposure to the issues of reproducing highly cited papers (and also by covering only R1 and R2 papers, the two categories of papers that already are releasing the most documentation), it may give off the impression that these papers document their research poorly compared to other papers. It is difficult to say whether highly cited papers have different documentation practices than papers with fewer citations [1], and performing such comparisons is not the intention of using this selection method, but rather to get a selection of important (i.e highly cited) papers; after all, highly cited papers have had some impact on the field, so issues with documentation found here may well apply elsewhere, even if it is not statistically possible to draw such a conclusion.

9.2 Reproduction Methodology

Our methodology changed a few times throughout the project, in particular from reproducing R1 papers to including R2 papers as well. The survey used has evolved somewhat from its first iteration to reflect aspects discovered when reviewing papers that did not fit entirely within the framework of the existing survey. Had the later methodology been known from the start, the process may have been slightly more organized, although such hindsight is rarely a luxury one has.

9.3 Reproducibility

Even though a large portion of the selection in some way pertains to neural networks, the problems encountered (in particular related to the implementation of the method and experiment) – as well as the ones related to the source code – are shared between papers from most topics covered. Neural network papers, in addition, are likely to be candidates for having the categories that cover training, validation, and test set specifications. These categories could in some sense be considered more specific than categories that relate to just the dataset, and thus having a lot of neural network papers does not necessarily limit the insight when it comes to datasets, as datasets are a prerequisite for these more specialized categories. The problem categories related to datasets in general (mismatches in descriptions, uncertainty about whether the dataset is the same, or other pre-processing not being included) are in fact among the larger categories that remain after the method description categories have been considered.

Papers that publish their method (even less the experiment) code are not too common among the ones we have covered. [2] Most of the problems we encountered in some sense

[1] the documentation metrics presented in figure 8.7 for both our covered papers and the papers covered by [GK18] from AAAI and IJCAI indicate that this is not the case; an informal search on a select few papers from the list of 400 papers shows citation counts in the range 2-10, compared to most of ours having 50 or more citations – the metrics in the figure are a mix of both better and worse across the selection for each year.

[2] It may be appropriate to add a disclaimer here that authors should not in general be given any blame for code not being released, as the authors might not be able to release their code even if they would like to due to e.g licensing requirements. It could nonetheless be argued that it should be a goal to strive towards.

pertain to this unavailability, as a consequence of the code being unavailable is that an implementation will have to be made, which brings along any issues with understanding the documentation of the method and experiment. The unavailability of code does not just hinder the R1-reproduction of a paper, but may also hinder R2 and R3 reproduction – even if a researcher is intending to implement their own version of the method, while using new data (R3) – having the original "reference implementation" of the authors available can be helpful for eliminating variables that may contribute to the reproduced results not corresponding to the presented ones.

The previous also applies to any other documentation, where variables of uncertaintly about whether the AI method is the deciding factor or not may be introduced. This is in part the reason for the desire to have a metric for reproducibility. The reproducibility levels R1, R2 (R2-D, R2-M), and R3 are far from having common consensus in the scientific community. They are, after all, one possible metric intended to formalize and measure the concept of the reproducibility of a research body (with respect to the documentation). Reading graphs like 8.6 should therefore be done with the consideration in mind that this metric is not absolute, though provided the surveying of papers that was used to generate it is consistent with respect to evaluation criteria, it can be used to say how the documentation level in the papers has evolved over time.

With this in mind, we are not seeing any of the results of [GK18]. That is to say, there is no consisent increase (as with the IJCAI papers), nor is the documentation scores (R1, R2, R3) fairly similar to each other, R2 and R3 diverging for the year 2014. The conclusion in their research, based on how only one of the two conferences showed improvement over time, is that their hypothesis of documentation improving over time is not supported. We too have to come to the same conclusion, that documentation scores for our selection has not improved over time – however, as was mentioned, our sample of 30 papers (10 per year) may be too small, compared to their sample of 100 papers per year per conference (400 in total).

Chapter 10

Conclusions

10.1 Conclusions and Suggestions for Future Work

This theis, in collaboration with a second separate thesis running alongside it, has presented an investigation into the degree of reproducibility and issues encountered during attempting to reproduce a selection of highly cited research within the field of artifical intelligence. A process for selecting such highly cited papers, as well as an evaluation of the topics they cover was presented.

The selection of papers was used in reproduction attempts, with various issues encountered having been documented and generalized. With reference to ealier work on measuring the change of documentation levels of papers over time, a metric for documentation score was used to estimate the documentation level of each paper reviewed.

The sample of papers reviewed is fairly small, so one natural direction for future work is to expand the sample, building on the methodology used, as well as considering different methods for generating a sample of papers, for instance covering subfields of artificial intelligence beyond the more popular ones.

Appendix A

Converting from new survey to AAAI-survey

Below is a listing of the fields in the evaluation of our 30 papers using the survey of [GK18] and [Kje17], and how each was determined:

The following are equivalent to the definitions of the newer survey, and the value can be copied over: research_type, affiliation, problem_description, goal/objective, hypothesis, prediction, contribution

The following need some closer consideration:

- **result_outcome (novelty):** All the research appears to present novel results, other than one paper ([Men+16]) which appears to mostly be a presentation and benchmark of a library than novel research.
- **research_method:** This has a stricter definition, the evaluation guide presented in [Kje17] saying that there has to be an explicit mention of the word "research method"
- **research_question:** Same as above.
- **pseudo_code:** Whether any pseudocode is present – set to 1 if the newer survey says partial or more.
- **open_source_code:** Needs to be explicitly referenced in paper through e.g supplementary materials. If it is, set to 1 if the newer survey mentions method code.
- **open_experiment_code:** Same as above
- **train, validation, test:** This is a very cumbersome set of fields to evaluate:
    - NaN for train, NaN for validation, and 1 for test if dataset is available and the method does not use training/validation. This is manually done.

  - NaN for train, NaN for validation, and 0 for test if dataset is unavailable and no training/validation.
  - If the method uses training/validation/test, set these to 1 if the newer survey fields for (training, validation, test)-partition is set to "some" or "all" (because the survey used in the AAAI-paper is not as strict about the entire split needing to be provided).
  - If any of these fields in the newer survey are set to not applicable, manually determine it by going through the paper.

- **hardware_specification:** This is stricter than in the newer study, requiring very explicit mentions of models. The newer study considers the hardware specification with respect to the paper, e.g if its code mainly runs on a GPU then the CPU is not of *as* much interest (although it can still be very useful to eliminate variables that might affect the result. As it is stricter, set 0 where "none" in the newer study, but if the newer study says "specified", then it has to be manually checked to conform with this stricter requirement.
- **software_dependencies:** Whether there is a readme/requirements file (or in the paper) that mentions software versions. This is manually done.
- **third_party_citation:** Manual, but fairly quick as any citations to code or datasets makes this 1.
- **experiment_setup:** This field was problematic in [Kje17], as it turned into a check for hyperparameters. In order to be consistent to enable comparisons with the results from their survey, this field is treated the same way here – i.e mentions of hyperparameter values or descriptions ("some" in newer survey) suffices.
- **evaluation_criteria:** Unfortunately, this field was vaguely described, and even looking through example papers from the survey of [Kje17], I was unable to determine how this field was set. As such, I have left it blank.

Appendix B

Code versions used

This appendix contains references to the repositories that were used to generate the tables, figures, and other data in this thesis, as well as which particular version/commit was used.

A) Implementation of the paper “A modified Artificial Bee Colony algorithm for real-parameter optimization” [AK12], as well as results from running the implementation.

Version: v1.0

Commit SHA-1: c929558502ba36771cb22bf9889ae5a0c23e3291

URL: `https://github.com/AIReproducibility2018/AModifiedABCAlgorithmOpt`

B) Modifications made to the retrieved implementation of the paper “Deep learning-based classification of hyperspectral data”, [Che+14], as well as results from running the implementation.

Version: v1.0

Commit SHA-1: b824e17219cee03a35ce64ebdf8ea323daf1cf5c

URL: `https://github.com/AIReproducibility2018/DLClassifHypspec`

C) Implementation of the paper “Semi-supervised and unsupervised extreme learning machines” [Hua+14], as well as results from running the implementation.

Version: v1.0

Commit SHA-1: a1076e040a02919939bb9aaa998037dac72dd9cb

URL: `https://github.com/AIReproducibility2018/SemiSupELM`

D) A Jupyter Notebook with a modified version of the code used to generate the reproducibility metric presented in [GK18] over time, adapted to include the metric over time for the papers we have covered.

Version: v1.0

Commit SHA-1: e1a28edeb19a1db9c318bc4bbd0862da622c7258

URL: https://github.com/AIReproducibility2018/UTILS_AAAI_metrics

E) The source code for generating various tables showing statistics about categories, and exporting them to figures. Also contains rather questionable code for converting Pandas dataframes to tex-tables with wrapped text columns.

Version: v1.0

Commit SHA-1: f404ef569b8a624e3e71dd9b56df6b0fc152ce20

URL: https://github.com/AIReproducibility2018/UTILS_tablegen

F) A Jupyter Notebook which analyses various attributes of the selection of papers, like citation counts and keyword proportions.

Version: v1.0

Commit SHA-1: 272d0553f3585a7c591aa861e6697333a0bbcb68

URL: https://github.com/AIReproducibility2018/UTILS_paper_sample_analysis

Bibliography

[ADF12] B. Alexe, T. Deselaers, and V. Ferrari. “Measuring the Objectness of Image Windows”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34.11 (2012), pp. 2189–2202. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2012.28.

[AK12] Bahriye Akay and Dervis Karaboga. “A modified artificial bee colony algorithm for real-parameter optimization”. In: *Information Sciences* 192 (2012), pp. 120–142.

[Bak16] Monya Baker. “REPRODUCIBILITY CRISIS?” In: *Nature* 533 (2016), p. 26.

[BD98] Jonathan Buckheit and David Donoho. “WaveLab and Reproducible Research”. In: Vol. 103 (Nov. 1998).

[BI15] C.G. Begley and J.P.A. Ioannidis. “Reproducibility in science: Improving the standard for basic and preclinical research”. In: *Circulation Research* 116.1 (2015). cited By 200, pp. 116–126. DOI: 10.1161/CIRCRESAHA.114.303819. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84920769131&doi=10.1161%2fCIRCRESAHA.114.303819&partnerID=40&md5=a0b59f3042d389aced6a500c27712dce.

[CG16] Tianqi Chen and Carlos Guestrin. “Xgboost: A scalable tree boosting system”. In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. ACM. 2016, pp. 785–794.

[Che+14] Yushi Chen et al. “Deep learning-based classification of hyperspectral data”. In: *IEEE Journal of Selected topics in applied earth observations and remote sensing* 7.6 (2014), pp. 2094–2107.

[Che+16] Badong Chen et al. “Generalized correntropy for robust adaptive filtering”. In: *IEEE Transactions on Signal Processing* 64.13 (2016), pp. 3376–3387.

[CPW15] Christian Collberg, Todd Proebsting, and Alex M Warren. “Repeatability and benefaction in computer systems research”. In: (2015).

[CZH16] Gong Cheng, Peicheng Zhou, and Junwei Han. "Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images". In: *IEEE Transactions on Geoscience and Remote Sensing* 54.12 (2016), pp. 7405–7415.

[Don+14] Jeff Donahue et al. "Decaf: A deep convolutional activation feature for generic visual recognition". In: *International conference on machine learning*. 2014, pp. 647–655.

[Els17] Elsevier. *Scopus - Content*. 2017. URL: https://www.elsevier.com/solutions/scopus/content (visited on 04/18/2018).

[GFI16] S.N. Goodman, D. Fanelli, and J.P.A. Ioannidis. "What does research reproducibility mean?" In: *Science Translational Medicine* 8.341 (2016). cited By 108. DOI: 10.1126/scitranslmed.aaf5027. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84973100833&doi=10.1126%2fscitranslmed.aaf5027&partnerID=40&md5=2ce2d3467c834109d4a33bb0eccae8c2.

[GJ14] Alex Graves and Navdeep Jaitly. "Towards end-to-end speech recognition with recurrent neural networks". In: *International Conference on Machine Learning*. 2014, pp. 1764–1772.

[GK18] Odd Erik Gundersen and Sigbjørn Kjensmo. "State of the Art: Reproducibility in Artificial Intelligence". In: (2018). URL: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17248.

[Gun15] Odd Erik Gundersen. "Towards Scientific Benchmarks: On Increasing the Credibility of Benchmarks". In: *Eurographics Workshop on 3D Object Retrieval*. Ed. by I. Pratikakis et al. The Eurographics Association, 2015. DOI: 10.2312/3dor.20151059.

[GW12] T. Guha and R. K. Ward. "Learning Sparse Representations for Human Action Recognition". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34.8 (2012), pp. 1576–1588. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2011.253.

[GZMT12] Stas Goferman, Lihi Zelnik-Manor, and Ayellet Tal. "Context-aware saliency detection". In: *IEEE transactions on pattern analysis and machine intelligence* 34.10 (2012), pp. 1915–1926.

[Hua+14] Gao Huang et al. "Semi-supervised and unsupervised extreme learning machines". In: *IEEE transactions on cybernetics* 44.12 (2014), pp. 2405–2417.

[Jia+16a] Feng Jia et al. “Deep neural networks: A promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data”. In: *Mechanical Systems and Signal Processing* 72 (2016), pp. 303–315.

[Jia+16b] Jianhua Jia et al. “iSuc-PseOpt: Identifying lysine succinylation sites in proteins by incorporating sequence-coupling effects into pseudo components and optimizing imbalanced training dataset”. In: *Analytical Biochemistry* 497 (2016), pp. 48 –56. ISSN: 0003-2697. DOI: https://doi.org/10.1016/j.ab.2015.12.009. URL: http://www.sciencedirect.com/science/article/pii/S0003269715005667.

[Kje17] Sigbjørn Kjensmo. *Research method in AI - Reproducibility of results*. eng. 2017. URL: http://hdl.handle.net/11250/2478230.

[Li+14] Wei Li et al. “Deepreid: Deep filter pairing neural network for person re-identification”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2014, pp. 152–159.

[LM14] Quoc Le and Tomas Mikolov. “Distributed representations of sentences and documents”. In: *International Conference on Machine Learning*. 2014, pp. 1188–1196.

[LMS12] R. Leveque, I. Mitchell, and V. Stodden. “Reproducible research for scientific computing: Tools and strategies for changing the culture”. In: *Computing in Science and Engineering* 14.4 (2012). cited By 37, pp. 13–17. DOI: 10.1109/MCSE.2012.38. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84864270397&doi=10.1109%2fMCSE.2012.38&partnerID=40&md5=bfaf31c10aa171f6bfecbcc8d33ac4fb.

[LNX12] Guoqiang Li, Peifeng Niu, and Xingjun Xiao. “Development and investigation of efficient artificial bee colony algorithm for numerical function optimization”. In: *Applied soft computing* 12.1 (2012), pp. 320–332.

[LP15] Jeffrey T Leek and Roger D Peng. “Opinion: Reproducible research can still be wrong: Adopting a prevention approach”. In: *Proceedings of the National Academy of Sciences* 112.6 (2015), pp. 1645–1646.

[LT16] T. Liu and D. Tao. “Classification with Noisy Labels by Importance Reweighting”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38.3 (2016), pp. 447–461. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2015.2456899.

[LY12] X. Li and X. Yao. “Cooperatively Coevolving Particle Swarms for Large Scale Optimization”. In: *IEEE Transactions on Evolutionary Computation* 16.2 (2012), pp. 210–224. ISSN: 1089-778X. DOI: 10.1109/TEVC.2011.2112662.

[Men+16] Xiangrui Meng et al. “Mllib: Machine learning in apache spark”. In: *The Journal of Machine Learning Research* 17.1 (2016), pp. 1235–1241.

[Oat06] Briony J. Oates. *Researching Information Systems and Computing*. SAGE Publications, 2006.

[OR16] Francisco Javier Ordóñez and Daniel Roggen. “Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition”. In: *Sensors* 16.1 (2016). ISSN: 1424-8220. DOI: 10.3390/s16010115. URL: http://www.mdpi.com/1424-8220/16/1/115.

[Pen+12] Yigang Peng et al. “RASL: Robust alignment by sparse and low-rank decomposition for linearly correlated images”. In: *IEEE transactions on pattern analysis and machine intelligence* 34.11 (2012), pp. 2233–2246.

[Pen11] R.D. Peng. “Reproducible research in computational science”. In: *Science* 334.6060 (2011). cited By 323, pp. 1226–1227. DOI: 10.1126/science.1213847. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-82755174123&doi=10.1126%2fscience.1213847&partnerID=40&md5=a818476504a1de27146bfa7e6f87855

[RL14] Alex Rodriguez and Alessandro Laio. “Clustering by fast search and find of density peaks”. In: *Science* 344.6191 (2014), pp. 1492–1496.

[SBC12] M. A. Saad, A. C. Bovik, and C. Charrier. “Blind Image Quality Assessment: A Natural Scene Statistics Approach in the DCT Domain”. In: *IEEE Transactions on Image Processing* 21.8 (2012), pp. 3339–3352. ISSN: 1057-7149. DOI: 10.1109/TIP.2012.2191563.

[Shi+16] Hoo-Chang Shin et al. “Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning”. In: *IEEE transactions on medical imaging* 35.5 (2016), pp. 1285–1298.

[Sil+16] David Silver et al. “Mastering the game of Go with deep neural networks and tree search”. In: *nature* 529.7587 (2016), pp. 484–489.

[VK11] J. Vitek and T. Kalibera. “Repeatability, reproducibility and rigor in systems research”. In: *2011 Proceedings of the Ninth ACM International Conference on Embedded Software (EMSOFT)*. 2011, pp. 33–38. DOI: 10.1145/2038642.2038650.

[VKV09] Patrick Vandewalle, Jelena Kovacevic, and Martin Vetterli. “Reproducible research in signal processing”. In: *IEEE Signal Processing Magazine* 26.3 (2009).

[Yin+14] Xu-Cheng Yin et al. “Robust text detection in natural scene images”. In: *IEEE transactions on pattern analysis and machine intelligence* 36.5 (2014), pp. 970–983.

[ZF14] Matthew D. Zeiler and Rob Fergus. "Visualizing and Understanding Convolutional Networks". In: *Computer Vision – ECCV 2014*. Ed. by David Fleet et al. Cham: Springer International Publishing, 2014, pp. 818–833. ISBN: 978-3-319-10590-1.

[Zha+12] Kaibing Zhang et al. "Single image super-resolution with non-local means and steering kernel regression". In: *IEEE Transactions on Image Processing* 21.11 (2012), pp. 4544–4556.

[Zha+14] Zhanpeng Zhang et al. "Facial landmark detection by deep multi-task learning". In: *European Conference on Computer Vision*. Springer. 2014, pp. 94–108.

[ZSI+12] Daoqiang Zhang, Dinggang Shen, Alzheimer's Disease Neuroimaging Initiative, et al. "Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease". In: *NeuroImage* 59.2 (2012), pp. 895–907.