NTNU
Norwegian University of
Science and Technology

Performing reproductions to understand the state of reproducibility in current AI research

Odd Cappelen
Martin Mølnå

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

Norwegian University of Science and Technology
Department of Computer Science

Performing reproductions to understand the state of reproducibility in current AI research

Odd Cappelen and Martin Mølnå

June 18, 2018

Abstract

In the last few years, the issue of reproducibility has gained increased attention in many scientific fields, including Artificial Intelligence (AI). Reproducibility of published results is a key concept of the scientific method, yet recent studies in AI and other computational sciences have shown that many experiments cannot be reproduced, and that current documentation practices are insufficient. In this project, reproductions are attempted of experiments from 30 highly cited papers in AI from recent years. The goal is to provide a better understanding of the state of reproducibility in the field, and identify issues limiting reproductions.

Three hypotheses are investigated in the project. First, it is hypothesized that most studies are difficult to reproduce. Secondly, the issues that make reproductions difficult are hypothesized to be similar across different studies. Thirdly, the level of documentation measured for an article is hypothesized to be related to how easily it can be reproduced. From the 30 papers investigated, 22 reproduction attempts were performed, where 10 were partially successful. The results achieved corroborate the first and second hypothesis, and the third hypothesis can neither be rejected nor corroborated.

Lastly, this project presents three contributions. The first contribution is the overview of the current state of reproducibility in AI provided by the results of the reproduction attempts. The second is a model for interpreting research articles in AI and estimating the level of documentation provided. The last is a set of categories of issues intended to cover most issues encountered during reproductions.

Sammendrag

Reproduksjon er et tema som i de senere år har fått økt interesse i flere vitenskapelige områder, deriblant Kunstig Intelligens (KI). Reproduksjon av publiserte resultater er et av nøkkelkonseptene i den vitenskapelige metoden, men nyere studier i KI og datateknologi har vist at mange publiserte eksperimenter ikke kan reproduseres. I dette prosjektet blir eksperimenter fra 30 nyere, høyt siterte artikler fra KI forsøkt reprodusert. Prosjektets mål er å skape en bedre forståelse for tilstanden til reproduksjon i KI i dag, og identifisere problemer som begrenser reproduksjonsforsøk.

Tre hypoteser ble undersøkt i dette prosjektet. Først, det antas at de fleste artikler er vanskelig å reprodusere. Den neste, problemene som gjør reproduksjon vanskelig antas å være liknende imellom artikler. Den siste, nivået av dokumentasjon målt for en artikkel antas å være relatert til hvor enkelt den kan reproduseres. Av de 30 artiklene undersøkt ble 22 forsøkt reprodusert, og av disse var 10 delvis suksessfulle. Resultatene støtter den første og den andre hypotesen, og den siste hypotesen kan hverken bekreftes eller forkastes.

Dette prosjektet har tre hovedbidrag. Det første bidraget er bildet av den nåværende tilstanden til reproduksjon i KI som gis av reproduksjonsresultatene. Det andre bidraget er en modell for å tolke forskningsartikler i KI og estimere dokumentasjonsnivået til en artikkel. Det siste bidraget er et sett med problemkategorier som er tiltenkt å dekke all problemer møtt under reproduksjonene.

Preface

This report is the Master's Thesis for our degrees in Computer Science at the Norwegian University of Science and Technology (NTNU). The project was performed in collaboration with Nicklas Grimstad Nilsen, who produced his own Master's Thesis from the project. Nicklas created the selection of papers for the experiment, while the model for understanding AI articles was developed by us. The methodology for the reproductions and the set of issue categories were created in collaboration. All three students participated equally in the reproduction attempts. The analysis of the results was performed independently by us and Nicklas, and the two reports were written independent of each other. The project was carried out under the supervision of Odd Erik Gundersen. The preliminary work of the project, including the creation of the experiment and collection of the background material, was carried out in the autumn of 2017 as part of a preliminary specialization project.

We wish to thank Odd Erik Gundersen for his help supervising this project.

We wish to thank Nicklas Grimstad Nilsen for his collaboration in this project.

We thank Jan Gulla for providing the LaTeX template used for this report.

Contents

List of figures

List of tables

Chapter 1

Introduction

Reproducibility, or replicability, of studies is a fundamental aspect of the scientific method. Through the process of reproduction researchers can corroborate good results, discard false leads, and build upon the work of others. However, in recent years the reproducibility of published results in many fields have been drawn into question [1]. The so-called "Replication Crisis" [2] has spread from psychology, and today touches most scientific fields, including the computational sciences and our field, AI [3]. This crisis refers not only to that fact that a significant portion of published results are being refuted as a result of reproductions, but also to the problem of many results being neither refuted nor corroborated because the experiments which produced the results are impossible to recreate.

Access to good documentation is a key requirement for reproducibility. In order to successfully reproduce a study researchers need a thorough understanding of the research question, methodology, and experimental setup of the original study. The current "Replication Crisis" is believed to be in part a crisis of documentation, arising from the difficulty of sharing all details of an experiment. However, the majority of experiment in computer science are computational experiments, where the experiment consists of running some program, or code, on some defined problem or data set. Computational experiments can be defined by the code executed, the hardware and software platform used, and, if applicable, one or more data sets. Sharing these resources should allow anyone to replicate the experiment. Models for packaging code and data into experiments that can be executed automatically have been suggested [4]. However, such models have not been widely adopted, and recent studies show major limitations in current code and data sharing policies [5], [6]. These studies indicate that the reproducibility problem has not been solved for AI.

The **goal** of this project is to provide a quantitative overview of the state of reproducibility in our field, AI, and to help produce a better understanding of the problems currently limiting reproducibility. The **method** proposed to achieve this goal is to perform a series of reproduction attempts on 30 recent, highly cited studies in AI. 10 studies are selected from each of the years 2012, 2014, and 2016, and the reproduction attempts are performed in a structured and transparent manner.

Three **hypotheses** are proposed for the project. The *first hypothesis* is that it is difficult to reproduce many of the results achieved in AI research in recent years. The *second hypothesis* is that the issues which make reproducibility difficult are the same across many studies. The *third hypothesis* is that there is a link between the level of documentation provided by an article, and how easy it is to reproduce. A prediction is defined for testing each hypothesis. The *first prediction* is that the majority of studies in this project cannot be reproduced within the limitations of the experiment. The *second prediction* is that it is possible to group the majority of issues encountered in the reproduction attempts into a set of categories. The *third prediction* is that there is a significant correlation between the documentation level measured for an article, and the outcome of the related reproduction attempt.

This study also has three main **contributions**. Firstly, it provides an overview of the state of reproducibility in AI research. Secondly, a model for understanding AI articles is proposed, along with a metric for estimating documentation levels. Third and lastly, a set of issue categories is proposed covering all major issues encountered during the reproductions.

The project builds upon the work done by Sigbjørn Kjensmo during his master project, also under the supervision of Odd Erik Gundersen. Kjensmo studied documentation and reproducibility in AI research through a survey of 400 papers, and developed a method for quantifying reproducibility. The result of Kjensmo's study was presented in a scientific paper [6], which has been accepted for publication in AI Magazine.

The remaining parts of this report are structured as follows. Chapter 2 provides an introduction to the scientific method in AI, followed by an overview of the concept of reproducibility in AI and of related work on the subject. Chapter 3 presents the model for understanding AI articles proposed in this project, followed by an overview of the reproduction procedure used. Chapter 4 presents the results achieved in the experiment, and the categories of issues encountered. Chapter 5 discusses the results and evaluates the project. Lastly, Chapter 6 concludes the report.

Chapter 2

Background

This chapter provides an introduction to the scientific method of AI research, followed by an overview of the concept of reproducibility as it is understood in computational sciences and AI. This also includes a presentation of some suggested methods for differentiating and classifying levels of reproducibility. Lastly, a brief overview of earlier work on reproducibility in computational sciences and AI is provided.

2.1 The Scientific Method in AI Research

The field of AI is highly diverse and covers many different topics such as classification, function optimization, and clustering. Defining exactly what constitutes AI and AI studies is therefore a difficult problem. Cohen [7] defined AI research as the study of **AI programs** and their behaviour. Gundersen and Kjensmo [6] expanded on this definition by introducing the concept of an **AI method**. The AI method is the conceptual algorithm or system which is implemented by an AI program. According to them, the scientific process in AI consists of the formulation and adjustment of beliefs about an AI program through the execution of experiments. Based on some initial beliefs, a set of hypotheses and predictions are made. An experiment is constructed to test the predictions, and the results are compared with the predicted outcomes. Based on the researcher's interpretation of the results, the beliefs are adjusted.

For this project, Gundersen and Kjensmo's model was expanded to explicitly encompass the concepts of AI method and phenomenon. In this model, illustrated in Figure 2.1, the set of initial beliefs are assumed to concern either what an AI method can do, or how to study or solve a phenomenon or task. If the beliefs are focused on the method, a phenomenon is chosen to test the method. If the beliefs are about a phenomenon or task, a method is proposed for investigating or solving it. When the method and phenomenon are chosen, one or more hypotheses are formulated, followed by a set of predictions. The method is implemented in a program, and a data set, or task specification, is created to represent

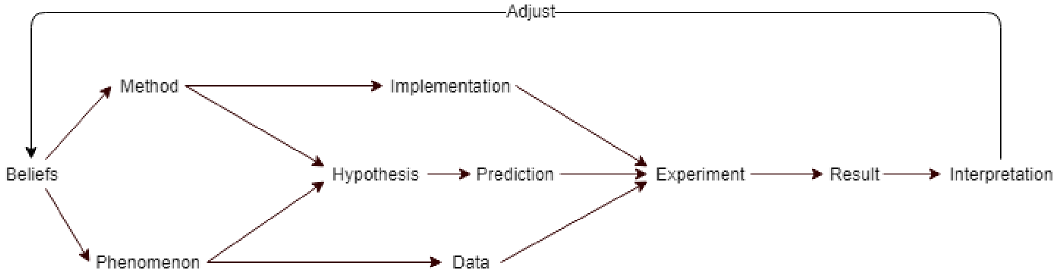

Figure 2.1: The research process in AI as understood in this project

the phenomenon. Using the implementation and data, the predictions are then tested in an experiment which produces results. The results are compared with the predictions and interpreted, and based on the interpretation the initial beliefs are updated. This model forms the basis for our understanding of empirical research in AI.

2.2 Reproducibility

Although reproducibility is an important aspect of research and the scientific method, there exists no commonly agreed upon definition of what reproducibility is. The U.S. National Science Foundation has the following definition of reproducibility, "*Reproducibility refers to the ability of a researcher to duplicate the results of a prior study using the same materials and procedures as were used by the original investigator*" [8]. However, this definition does not distinguish between different degrees of reproducibility, and it remains unclear what the exact goal of a reproduction team should be.

Several researchers have suggested a separation between the terms **replication** and **reproduction** [5], [9]–[11], but no consensus seems to exist on precise definitions. According to Drummond [9], replication is the re-running of the original experiment with minimal changes to the experiment. Reproduction, on the other hand, is testing or corroborating the conclusions of the original study through new and different experiments. The creators of the ReScience Initiative [11] have adopted the reverse position. According to their definition, reproduction is running the same software with the same data and obtaining the same results. Replication is writing a new implementation and aiming at achieving results which are equivalent, but not necessarily identical.

Goodman, Fanelli, and Ioannidis [12] proposes an alternative division of reproducibility into three categories; **methods reproducibility**, **results reproducibility**, and **inferential reproducibility**. In their system, a study is methods reproducible if it is documented well enough to be repeated exactly, using the same experimental setup, code, and tools. It is results reproducible if a reproduction using the same experimental methods yield results which corroborate the results of the original study, and inferential reproducible if a reproduction results in conclusions which are similar to the conclusions of the original study.

In their paper, Gundersen and Kjensmo [6] proposes a new definition for reproducibility in AI research: "*Reproducibility in empirical AI research is the ability of an* independent research team *to produce the* same results *using the same* AI method *based on the* documentation *made by the original research team.*" With this definition they emphasis the importance of the reproducibility being carried out by a team of researchers independent of the original research team. They also introduce the term AI method to refer to the proposed algorithm or method, and distinguishes it from the specific implementation created in the original study. From this definition, Gundersen and Kjensmo, proposes three **degrees of reproducibility**, distinguished by their independence from the original implementation and data set, similar to the separation between reproduction and replication. The three degrees of reproducibility proposed in [6] are:

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 execution of *an alternative implementation of the AI method* produces the same results when executed on *different data.*

R1 reproducibility involves the least independence from the original experiment, using the exact same code and data. Its goals can be considered equivalent to those of replication. *R3* reproducibility involves reproduction independent of both the implementation and data set of the original experiment, and is closer to reproducibility as defined in [9].

For the remaining part of this report, the term reproduction and reproducibility will be used according to the definition proposed by Gundersen and Kjensmo. Furthermore, a modified version of their degrees of reproducibility, presented in Section 3.3, will be used to classify reproduction attempts.

2.3 Related Work

There have been several studies into reproducibility in computational research and AI, revealing different issues about the current state of reproducibility. Various initiatives for improving the situation have also been proposed. Section 2.3.1 discusses some selected studies on concrete reproduction attempts and their outcomes. Section 2.3.2 presents some of the initiatives for increased reproducibility and openness which have been proposed in recent years, as well as surveys documenting the current state of reproducibility and documentation practices.

2.3.1 Reproduction Attempts in Computational Sciences and AI

In 2010, Mende [13] attempted to replicate two studies in Defect Prediction Models (DPM), with the goal of identifying potential problems. Using the same data, but independent code implementation, Mende was able to replicate the results of one of the studies, but not the other. He also produced some recommendations for facilitating replication, including explicit description of data transformation and summaries of data sets.

In 2013, Fokkens et al. [14] in a similar study attempted to reproduce two studies in Natural Language Processing (NLP). They showed that documentation is often too poor to exactly replicate the results of the original studies. Furthermore, they showed that the results of an experiment can be heavily influenced by aspects which are often not thoroughly documented, such as data pre-processing and resource versioning.

In 2015, Topalidou, Leblois, Boraud, and Rougier [15] attempted to reproduce a model from computational neuroscience. In their case, the source code of the original model was provided, but due to missing packages they were unable to compile it. Their effort at recoding the model in a new programming language was successful, but the entire process took approximately 3 months. Their reproduction attempt was published as one of the first in the ReScience Initiative [11].

In 2016, Vitay [16] published a reproduction attempt of a study on recurrent neural networks. This reproduction attempt was successful, and was made possible by the original article being detailed and well documented, and the original source code being available.

In 2017, Manninen, Havela, and Linne [17] studied reproducibility in their field of computational neuroscience. They attempted to reimplement and reproduce results from four computational models of astrocyte excitability using only published information. They were only able to completely reproduce results for one of the models, and found that the three other models did not provide sufficient information for reimplementation.

2.3.2 Initiatives for Increased Reproducibility

Several initiatives for increased reproducibility have been proposed. Since access to documentation, code, and data generally are accepted as important criteria for reproducibility, several of the initiatives have been focused on increased openness. Perspectives calling for increased openness and focus on reproducibility in computational sciences have been published in several leading scientific journals [5], [18], [19]. The goal of these efforts have been to encourage computational scientist to make more of their data and code publicly available, and encourage the scientific community to adopt stricter data sharing policies. As stated by Ince, Darrel, and Graham-Cumming [19], "*..., anything less than release of actual source code is an indefensible approach for any scientific results that depend on computation, because not releasing such code raises needless, and needlessly confusing, roadblocks to reproducibility.*"

The OpenML project [20] is an initiative for increased reproducibility specifically targeting the AI and Machine Learning community. OpenML is a platform for sharing code, data, and experiments, with the goal of making these resources easily accessible to other researchers and encouraging collaborative work.

The ReScience Initiative [11], launched in 2015, is a peer-reviewed journal in computational research focused on the reproduction of previously published results. Their aim is to encourage reproduction of existing science, and provide a journal were such efforts can be published. Furthermore, in order to encourage good documentation practices, all published reproduction attempts must themselves be re-runnable by other researchers. As of May 2018 19 papers on reproduction attempts have been published in ReScience, all reporting successful reproductions [21].

Despite the initiatives for increased reproducibility, recent surveys show that the situation is far from ideal. A survey of data sharing policies for journals showed that as of June 2012, only 38% of journals had explicit data policies, and only 22% had code policies [22]. Furthermore, in a 2010 survey of participants at the Neural Information Processing Systems (NIPS) conference, participants self reported sharing only approximately 32% of their code and 48% of their data online [23].

Even when efforts have been made at making research reproducible, the quality of the documentation may cause problems. Mayer and Rauber [24] studied experiments documented through workflows, systems used for defining and executing a series of computational steps, and showed that many experiments still were difficult to reproduce. Surveying 1 443 publicly available workflows, they found that only 29.2% of the experiments could be executed successfully. Several of the encountered failures were due to inadequate documentation, or missing resources.

Looking specifically at the availability of code, Collberg and Proebsting [25] attempted to estimate the repeatability of several computational studies by attempting to find and build the code used in the original studies. Surveying 402 studies, they were only able to build the code independently in 48.3% of the cases, rising to 54.0% when they contacted the original authors. The study did not attempt reproduction of results, and it is not known what percentage of built code would produce the same results as was published.

Gundersen and Kjensmo [6], did a survey specifically within the field of AI. Attempting to estimate the degree to which studies were *R1*, *R2*, and *R3* reproducible using a set of variables measuring documentation level, they found that out of 400 papers only about 25% were *R1* reproducible, 28% *R2* reproducible, and 30% *R3* reproducible.

Chapter 3

Experiment

This chapter presents and explains the experiment carried out in this project. Section 3.1 introduces the idea of an empirical study in AI as understood in this project. Section 3.2 presents the model for understanding AI articles developed in the project. Section 3.3 presents the classification of reproducibility used in the project. Section 3.4 mentions the selection process for studies, but this is discussed in greater detail in Nicklas Grimstad Nilsen's master thesis. Section 3.5 discusses the methodology and procedure used in the reproduction attempts. Lastly, Section 3.6 gives an overview of the documentation practices of this project.

The goal of the experiment is to perform reproduction attempts of multiple recent, highly cited AI studies in a structured and transparent manner. The aim is to produced quantitative results on the degree to which the selected studies can be reproduced, and on the problems and issues encountered in the reproduction processes. The methodology used in the experiment is intended to be sufficiently structured to enable a good comparison of the outcomes of several different reproduction attempts, while also flexible enough to handle the wide variety of empirical studies published in the field of AI. Furthermore, the proposed methodology is intended to be as transparent as possible, to encourage other researchers to understand and build upon the work.

3.1 Empirical Studies in AI

The focus of this project is empirical studies in the field of AI. For the purpose of this project, AI is interpreted broadly and a study is considered to be within AI as long as the method studied or used is commonly agreed to fall within AI. Empirical studies are interpreted as studies which propose new methods or hypotheses, and perform new experiments. Survey studies, technical guides, or papers which present only a data set and not a method are not classified as empirical in this project.

3.2 Model for Understanding AI Studies

To aid the work with the studies an *Article Model* is proposed, along with an associated *Article Model Metric*. The model provides a structured overview of the different aspects of an AI study. Given an article, the goal is to be able to use the model to identify the most important aspects of the study, and which aspects are well documented by the article, and which are not. The model is represented as an UML diagram in Figure 3.1. The model is based on the factors and variables proposed by Gundersen and Kjensmo in [6].

For most studies the research article is the main form of documentation provided, and should cover all important parts of the study. The *Article Model* divides an article into a set of *components*, with relations between them. Some of the components are further divided into *sub-components*, which are separate parts of a component. A component is an aspect of the research, such as the AI method implementation or the data sets used, which should be documented in the article. For each component there is a proposed *Component Metric*, which provides a method for estimating the degree to which the component is well documented in the article. Together, *Component Metrics* make up the *Article Model Metric*.

The purpose of the metrics is to provide a quantitative measure of the documentation level of each component in a given article, and the overall documentation level. I.e. they can be used to estimate how well a given article documents the main aspects of its study. Each metric is designed to use values between 0 and 1, with 0 indicating a poor documentation level, and 1 a good documentation level. For components with one or more sub-components, the *Component Metric* of the entire component is the average of the values for the sub-components. Like the model, the metrics selected are heavily influenced by the work of Gundersen and Kjensmo [6].

In the following sections we describe each component and the relations between them.

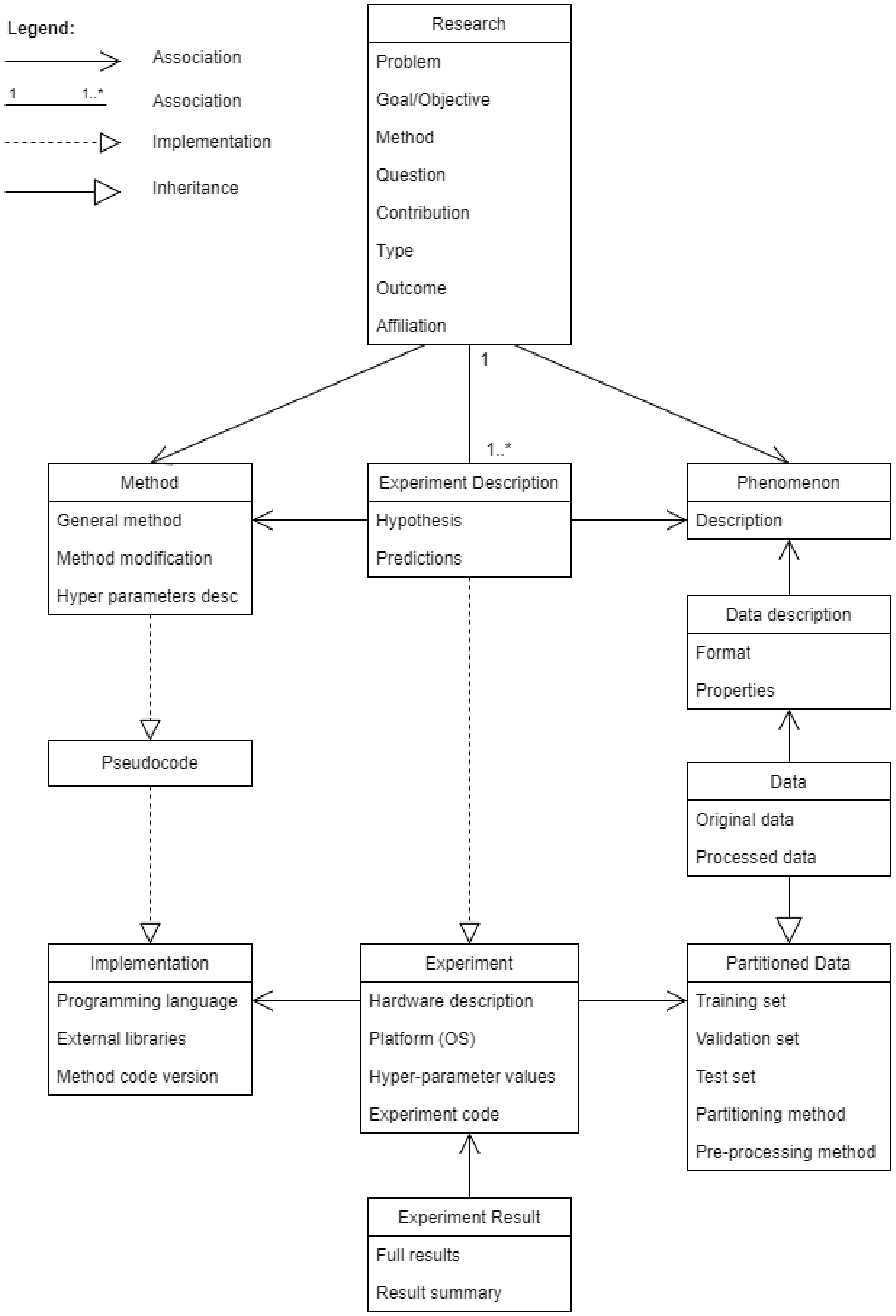

Figure 3.1: UML model of the *Article Model*

3.2.1 Research

The *Research* component of the model gives an overview of the research conducted in the study and should include the groundwork of the study. The first five sub-components covers which problem the article seeks to solve, what the goal of the research is, which research method is employed, what the research question is, and what the contribution of the study is. Additionally, other factors of the research, such as the type, the outcome, and the affiliation of the authors is included in this component.

The component is connected to the *Method* and *Phenomenon* components because each study should focus on one AI method and one phenomenon which the method is applied to. The component is also connected to the *Experiment Description* component because one or more experiments can be performed as part of the study.

A metric is proposed for each sub-component. As stated above, the aim of the *Component Metric* is to provide a quantitative measure for how well the given component or sub-component is documented in the article. The proposed metrics for the sub-components of the *Research* component are listed below. The type, outcome, and author affiliation of the study are not quantified, but rather classified with a set of possible values. The proposed metric for the entire component is the average of the values estimated for the first five sub-components.

Problem: Explicitly mentioned in article (1), or not mentioned (0).

Goal/Objective: Explicitly mentioned in article (1), or not mentioned (0).

Method: Explicitly mentioned in article (1), or not mentioned (0).

Question: Explicitly mentioned in article (1), or not mentioned (0).

Contribution: Explicitly mentioned in article (1), or not mentioned (0).

Type: Experimental (E) or Theoretical (T).

Outcome: Positive (P) or Negative (N).

Affiliation: Academia (A), Collaboration (C), or Industry (I).

3.2.2 Method

The *Method* component covers the AI method, or algorithm, used in the study. The description of the method is divided into two main parts. Many new methods proposed in AI are variations of existing methods, and the first part of the description is the the general method description, which covers the fundamentals of the existing method on which the new method is based. The second part is the method modification description, which covers modifications made to the original method in this particular study, if it is based on a general method. In addition to this, all hyper-parameters used by the AI method should be described as part of the method description.

For every sub-component, a metric is proposed for estimating the level of documentation of that sub-component. These are listed below. The proposed metric for the entire component is the average of the values estimated for the sub-components.

General method: Described (1) or not described (0).

Method modification: Described (1) or not described (0).

Hyper parameter description: All parameters described (1), some parameters described (0.5), or no parameters described (0).

3.2.3 Pseudocode

In many cases a research article in AI will include pseudocode to explain the proposed AI method. The *Pseudocode* component covers this aspect of an article. The pseudocode can be viewed as a formalization of the method description given in the *Method* component, and in the model it is therefore considered an implementation of that component.

In order to measure the level of pseudocode documentation provided by an article, a metric with the following possible values is proposed: Method completely covered by pseudocode (1), method partially covered by pseudocode (0.5), and no pseudocode (0).

3.2.4 Implementation

The *Implementation* component covers the actual implementation of the AI method. When covered in an article, the implementation is well documented if the following aspects are covered. First, the programming language used in the implementation should be mentioned in the article. Second, any external libraries used as part of the implementation should be listed. Both of these aspects need to be documented in order to enable exact re-implementation. In addition to this, the source code of the implementation should ideally be available, to allow researchers to replicate the experiments with the same code. The implementation code is only part of the code necessary for an experiment, which also includes experiment setup and potentially data pre-processing. Details about the experiment and experiment code is found in Section 3.2.10. In some cases the method implementation is further developed after a paper is published. Because of this, the version of the method code used in the article should also be specified.

Since the running program can be viewed as an implementation of pseudocode, the *Implementation* component is viewed as an implementation of the *Pseudocode* component in Figure 3.1.

The degree of documentation provided about the implementation can be estimated using a set of metrics. Below are the proposed metrics for each sub-component. The metric for the entire component is the average of the values for the sub-components.

Programming language: Specified in article (1), or not specified (0).

External libraries: Specified in article (1), or not specified (0).

Method code version: Average of the two following metrics:

1. Code provided with article (1), code available online (0.5), or code not available (0).
2. Code version used specified (1), or code version used not specified (0).

3.2.5 Phenomenon

The phenomenon of a research article is the real world event or concepts that the research tries to study, or apply the AI method to. The *Phenomenon* component covers information about the phenomenon.

To measure the degree to which the phenomenon is documented in the article, a metric with three possible values is proposed. The possible values are: phenomenon described (1), phenomenon mentioned (0.5), and phenomenon not mentioned (0).

3.2.6 Data Description

Most AI studies are performed by running an AI method on some data. When data is used, the article documenting the study should include a description of the data instances. This description should include the format of the data, e.g. the file format such as png or csv. Additionally, the properties or meta-data of the data instances should be described. This can be information about the structure of the data, or a description of the method used to gather or generate the data. The purpose of the data description is to document the properties of the data sufficiently well that independent researchers can recreate equivalent data sets.

Any description of the data used will depend upon the phenomenon studied, and the *Data Description* is therefore associated with the *Phenomenon* component in Figure 3.1.

The quality of documentation provided by the data description depends on the level of documentation provided about the data format and data properties. To estimate the level of documentation provided for the sub-components, the following metrics are proposed. The estimate for the entire *Data Description* component is the average of the estimates for the sub-components.

Format: Specified in article (1), or not specified (0).

Properties: Specified in article (1), or not specified (0).

3.2.7 Data

The *Data* component covers the actual data sets used in the study, and their level of availability to independent researchers. Whenever external data sets are used they should be clearly documented in the article, either through references to the articles which originally presented the data sets, or through links to online repositories where the data sets are available. Having all data sets used in a study be available to other researchers allows these researchers to use the exact same data when performing a reproduction attempt, and is key to enabling replication of studies.

The choice of data sets is dependent upon which properties are desired of the data. As such, the *Data* component is related to the *Data Description* component, which provides information about the data instances.

The level of the documentation provided about the data used in a study is entirely based on availability. More precisely, how many of the data sets used in the original study are available to independent researchers. When estimating the level of availability, the model differentiates between availability for two sub-components. The first sub-component covers the original, or raw form, data. Often a study will use external data sets created by other researchers, or originally gathered for some other experiment. These data may require some pre-processing before they are used in the new experiment. Having access to the original data is important for researchers seeking to reproduce an entire study, starting with the same resources as the original researchers.

The second sub-component covers data in a processed state. As mentioned, many studies will use existing data, but perform some sort of pre-processing or augmentation before passing the data to their AI program. This can involve reducing dimensionality of data, performing image processing, or other tasks. Having access to the processed data is key for researchers seeking to replicate a study, using the exact same data.

When estimating the availability of the data, the availability of the two sub-components are estimated separately. For each data set being used in an experiment, the model differentiates between three levels of availability for that data set, each with a score:

1. Data set is *provided* by article. Either through a link to an online repository, or through references to the original articles describing the data set. (1)

2. Data set is not provided by the article, but it is *retrievable*. This means that the data set is not linked or adequately referenced in the article, but it is possible to find or reconstruct the data set using resources online. (0.5)

3. Data set is *missing*, i.e. not provided or retrievable. (0)

The relationship between these levels of availability is illustrated in figure 3.2.

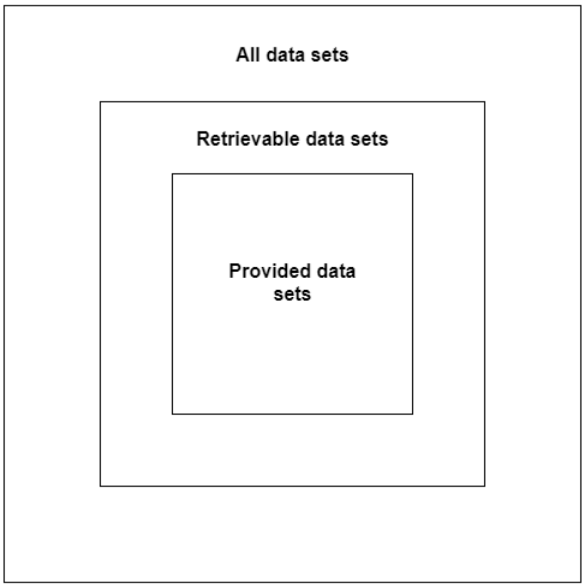

Figure 3.2: Different levels of data availability

The metric for estimating the level of availability, or documentation, for each sub-components is based on the total number of data sets used, and the number of data sets in each of the above categories. For each sub-component, the values for the following set of variables are recorded:

$\mathbf{D}_T$ Total number of data sets used in the original study.

$\mathbf{D}_P$ Number of data sets *provided.*

$\mathbf{D}_R$ Number of data sets *retrievable.*

$\mathbf{D}_M$ Number of data sets *missing.*

Based on these variables, and the scores of each category outlined above, the following metric is proposed for estimating the level of data availability for each sub-component.

$$Availability = \frac{D_P + \frac{1}{2}D_R}{D_T}$$

Using this formula on each of the two sub-components results in two measure of availability. One for original, or raw, data, and one for processed data. The estimate for the documentation level of the entire *Data* component is the average of these two values.

We observe that this metric gives the highest score of 1 to articles where all data sets used are *provided* in both raw and processed form. This is the ideal level of documentation, and gives independent researchers the greatest possibilities for reproducing or replicating the study. Data sets which are not provided, but which have been made available online, i.e. are *retrievable*, are given partial credit. When some data sets are *missing*, they are given no credit, and the total score of the article drops.

3.2.8 Partitioned Data

Several AI methods use parts of the available data for different purposes. A common pattern is to divide data into a training set, a validation set, and a test set. Others only utilize a part of the data by using a subset of the data. The *Partitioned Data* component covers this partitioning, and the methods used to perform it. When documenting the partitioning of the data, a research article should ideally provide all the partitions used in the study. Additionally, the method used to perform the partitioning should be described. In cases where the actual partitions are not provided, the description of the method should allow independent researchers to reconstruct the partitions from the original data sets.

As mentioned in the discussion of the *Data* component, data sets are often pre-processed before being used in an experiment. In these cases, the method used for processing the data should be described. The purpose of this description is to allow other researchers to re-implement the pre-processing method. In the model, the method for pre-processing is viewed as part of the *Partitioned Data* component.

Since the data instances of the partitioned data are the same as those in the original data sets, the *Partitioned Data* component is represented as specialization of the *Data* component in the UML diagram of Figure 3.1.

Ideally, a research article should provide all data used in the study in a processed and partitioned state, in the final form in which it was used in the original experiment. To measure how well this documentation actually is, the following set of metrics has been proposed for the sub-components of the *Partitioned Data* component. The estimate for the entire component is the average of the estimates for the sub-components.

Training set: All training sets provided (1), some training sets provided (0.5), No training sets provided (0).

Validation set: All validation sets provided (1), some validation sets provided (0.5), no validation sets provided (0).

Test set: All test sets provided (1), some test sets provided (0.5), no test sets provided (0).

Partitioning method: Provided for all data sets (1), provided for some data sets (0.5), provided for none of the data sets (1).

Subset of data: All subsets provided (1), some subsets provided (0.5), no subset provided (0).

Pre-processing method: Described (1), or not described (0).

3.2.9 Experiment Description

All empirical studies should contain one or more experiments. Each experiment starts with an overall description outlining the purpose of the experiment. Central to this is the formulation of a hypothesis, and a set of predictions. Documenting these are important for expressing the purpose of the experiment to other researchers. Furthermore, experiments should be related to one AI method performed on one phenomenon, as shown if Figure 3.1.

When measuring how well a research article documents the purpose of an experiment, this model proposes to estimate the level of documentation for the hypotheses and predictions, and use the average of these values as an estimate for the documentation level of the entire *Experiment Description* component. To estimate the documentation of the hypotheses and predictions, the following metrics are proposed.

Hypothesis: Explicitly mentioned in article (1), or not mentioned (0).

Predictions: Explicitly mentioned in article (1), or not mentioned (0).

3.2.10 Experiment

An actual experiment in AI is an implementation of the experiment description discussed above, aiming to test a proposed prediction. In practice, it usually involves running some AI program with some data, in a particular setting. There are several aspects of the experiment which should be documented in order for replication to be possible. Experimental setup factors, such as the hardware running the experiment and the operating system of that hardware are important aspects. Furthermore, most AI programs accept some hyper-parameters which controls the running of the program. Documenting these are vital to enable others to recreate the experiment. Lastly, the execution of the experiment is also usually performed using an experiment program. This program may perform tasks such as reading and pre-processing data, setting hyper-parameters, and calculating results.

In terms of relations to other components, the *Experiment* component can be viewed as an implementation of the *Experiment Description* component. Furthermore, as mentioned above, the experiment usually involves running an implementation of an AI method with some data, often partitioned into training, validation, and test sets. Because of this the **Experiment** component is related to the *Implementation* and *Partitioned Data* components.

The estimate for the documentation level of the experiment provided by an article is based on how well the aspects discussed above are documented. The following metrics are proposed for estimating the documentation level of each sub-component. For the entire component, the estimate is the average of the values for the sub-components.

Hardware description: Provided (1), or not provided (0).

Platform (OS): Provided (1), or not provided (0).

Hyper-parameter values: All values given (1), some values given (0.5), or no values given (0).

Experiment code: Code provided by article (1), code available online (0.5), or not available (0).

3.2.11 Experiment Result

When running an experiment, a set of results is produced. In a research article, these results are usually documented through some form of aggregation or summary, which conveys the most important observations from the experiment. To fully facilitate reproduction, the full results of an experiment should also ideally be made available. The full result is the actual output of the AI program when run on a data set. The *Experiment Result* component covers both these kinds of results. Since the results are directly dependent on an experiment, the component is related to the *Experiment* component i Figure 3.1.

The estimate for the documentation level of the *Experiment Result* component is based on the degree to which result summaries and full results are provided. Below is the proposed metric for evaluating the documentation of these sub-components. For the entire component, the estimate is the average of the values for the sub-components.

Full results: Provided (1), or not provided (0).

Results summary: Provided (1), or not provided (0).

3.3 Levels of Reproducibility

All reproduction attempts are classified according to the reproduction level attempted. As discussed in Chapter 2, Gundersen and Kjensmo [6] proposed three levels of reproducibility, *R1*, *R2*, and *R3*. However, their system did not cover the case where the method implementation used in a study is available, but not the data. To cover this situation, we expanded their system by dividing the *R2* reproduction level into two levels, *R2-D* and *R2-M*. *R2-D* retains the original definition of *R2*, while *R2-M* is introduced to describe reproductions where implementations, but not data, are available. The definitions for the four levels of reproduction used in this project are therefore as follows.

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-D: 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.

R2-M: Method Reproducible The results of an experiment are method reproducible when the execution of the same implementation of an AI method produces the same results when executed on *different data.*

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

The term *same implementation* is slightly unclear since, in practice, only part of the implementation required for an experiment might be shared. For the purpose of this project, we consider *same implementation* to mean that the method implementation is the same in the original study and reproduced attempt. The implementation of the experiment, or pre-processing of data, might therefore be different.

In practice a published study can contain multiple experiments with different levels of reproducibility. However, in this project we limit ourselves to one level of reproducibility per study. More specifically, the highest level of reproducibility is chosen. Therefore, a study is considered *R1* reproducible if at least one of its experiments are *R1* reproducible, and similarly for *R2-D* and *R2-M.* A study which is neither *R1*, *R2-D*, or *R2-M* reproducible is only *R3* reproducible.

3.4 Selection of Studies

The process for selecting studies to be reproduced was created by Nicklas Grimstad Nilsen, and is discussed in detail in his master thesis. A short summary of the process is provided in this section.

In total 30 papers were used in this study, 10 each from the years 2012, 2014, and 2016. Using the Scopus website[1], a search was performed in each year for empirical papers in AI, and the results were ranked according to the number of citations. The ten most highly cited papers from each year were selected for this project. In the initial read through of the top ranked papers it was discovered that some of the papers produced by the search were not empirical AI studies as described in Section 3.1. These papers were replaced by next most highly cited papers. The final list of papers therefore contains the most highly cited empirical papers in AI from the years 2012, 2014, and 2016. The final list of papers is given in Table 4.1 in Chapter 4.

3.5 Reproduction Procedure

This section describes the procedure used during the reproduction attempts. Having selected a study to reproduce, the article documenting the study is found and read through.

[1] www.scopus.com

The possible levels of reproducibility for the study is then determined. When searching for implementations to determine if a study is *R1* reproducible we first check if code or implementation is linked from the research article. If no link to implementation is provided we perform a search online using the Google search engine[2], searching for the name of the study, and the name of the study followed by the term "github" to specifically check the popular code hosting platform GitHub[3]. Lastly, we check the web pages of the main authors of the study. If an implementation is found we try to determine if this implementation is original, or just another reproduction. To do this we check if the implementation explicitly mentions being part of the original study, or if the code author is one of the authors of the original study. If either of these requirements are met, the implementation is assumed to be original. Implementations shared in a non-inspectable manner, i.e. as compiled programs, are not used in this project since the reproduction team is unable to verify that the program implements the correct method.

A similar procedure is used when searching for data sets. When a data set is mentioned in the study, we check if a link is provided to the web page hosting the data set, or if the paper proposing the data set is referenced. If a link or reference is not provided we perform a search for the data set name using the Google search engine. If no matches are found we then look to referenced articles for information on where to find the data set. When determining if a data set found online is the same as was used in the original study, the data set is assumed to be the same if the name of the data set is identical, or if the original article and the page hosting the data set references the same research paper.

As a policy we do not contact the original authors during our reproduction attempts. Neither to ask for original implementations or data, or to resolve uncertainties in our understanding of the studies. It is our opinion that studies should be reproducible by independent researchers using only the publicly available documentation.

In some cases a study will use a popular data set in one or more of its experiments, but will perform significant pre-processing on the data before using it in the proposed method. Preprocessing may change the data instances or the composition of the data set, and results in a new data set slightly different from the original. However, since a study often uses data sets owned by other researchers or institutions, the authors may have limited ability to share the new data. When encountering studies with data pre-processing where the processed data is not shared we try to re-run, and if necessary re-implement, the data pre-processing. The only exceptions to this are studies where the pre-processing must be re-implemented and involves multiple complex stages, or where the pre-processing requires manual editing. In these cases, the likelihood of us creating a data set which is identical or equivalent to the original data set is deemed low. Studies with these kinds of pre-processing are considered to be *R3* reproducible, on the grounds that the processed data is sufficiently different from the original data set to be considered a new data set.

[2] www.google.com
[3] www.github.com

When the possible levels of reproduction for a study have been determined reproduction is started. As stated in Section 3.3 reproductions are only performed at the highest possible level for each study. I.e. if a study is deemed *R1* reproducible, only *R1* reproduction is attempted. *R3* reproduction is not attempted for any study, due to the difficulty of comparing results on different data sets. Also as discussed in Section 3.3, our definition of *R1* reproducible requires the method implementation to be provided for at least one experiment. As such, *R1* reproduction may involve the writing of new code, primarily implementation of experiments.

In this project each reproduction attempt is limited to a maximum of 40 work hours, or approximately one work week. 40 hours is considered a reasonable effort, and it is our belief that well documented studies should be reproducible within this time frame.

Many published studies include more than one experiment. When attempting reproduction we focus on one experiment at a time. When selecting which experiment to attempt first, emphasis is put on the importance of the experiment in the article, i.e. how much it is discussed, and the order in which the experiments are presented. In most cases, the first eligible experiment is chosen as the first to be reproduced. Some experiments may be excluded as a result of the level of reproduction attempted. For example, when attempting *R1* reproduction only experiments covered by the provided method code is considered eligible for reproduction. If, after having achieved results for the first experiment, there is still time left, we move on to the next eligible experiment.

Counting the number of experiments in an article and differentiating between them may be difficult. Since different articles use different definitions of experiment, there is a need for a common definition. The definition used in this project is that *one experiment* is *one method* run on *one data set or function.* When running multiple methods or using multiple data sets, these are considered to be multiple experiments even if the original article does not classify them as such.

When performing a reproduction attempt, programming language and third-party libraries are chosen to be the same as was used in the original study. When programming language is not specified, a language considered suitable is chosen. The hardware used for reproduction are the personal computers of the reproduction team and one high-end GPU system operated by NTNU. When the use of a third-party library is mentioned in a study we will attempt to use the same library and version. However, if the mentioned library is unavailable or highly impractical to use a substitute library may be used. We also allow the use of third-party libraries not mentioned in the original paper in our implementation in cases where this is considered practical. For example in cases where a study uses a known algorithm as part of its method, but does not describe how this algorithm was implemented. During the reproduction, whenever a random number generator is used, the seed is explicitly set in the code. All results produced by a method, not just metrics and result aggregates, are written to file.

In some cases a reproduction attempt may be aborted before all experiments have been attempted or the time limit is reached. This is done for studies where the hardware demand exceeds what is available to our reproduction team, or in cases where a reproduction attempt has reached a situation where it is deemed impossible to get any results within the remaining time frame of the reproduction attempt.

In order to analyze the results of the project some way of classifying the outcomes of the reproduction attempts is needed. There are two level of outcomes to consider: The outcome of an experiment, and the overall outcome for the reproduction attempt. The classification of experiment results is addressed first.

In cases where the reproduced results are *identical* to the original results, the experiment is considered a success. However, in many cases the reproduced results will not exactly match the original values. Still, the results which are close to the original should be distinguished from those which are far off. Most studies include results from at least one baseline algorithm which is used to compare the results of the new method proposed in the study with the existing state-of-art methods. When this is the case, the outcome of a reproduction attempt can be evaluated using the baseline results. If the results from a reproduction attempt have the same performance relative to all baseline methods as the original results, the reproduced results are considered *consistent* with the original results. In these cases, the conclusions drawn about the performance of the method are the same for the reproduced and original implementation. If the reproduction result is classified as neither *identical* or *consistent*, it is classified as *different*. For this project we therefore define three possible outcomes for each experiment reproduction attempt: *identical*, *consistent*, and *different*. A reproduction attempt is *identical* if all results for that experiment are identical to the original results. The attempt is *consistent* if the performance of the reproduced results, relative to all baseline methods, is the same as the original results. Lastly, results which are neither *identical* or *consistent* are classified as *different*.

The comparison of results is primarily performed on reported aggregated metrics or summary of results. I.e. when comparing two implementations of a classification method, the comparison is performed based on accuracy or error achieved, not on the exact classification of each example in the test set. The main reason for this is that most studies only report a result summary, and do not provide the full results. Furthermore, requiring exact match on the full results is a significantly harder requirement.

When it comes to classifying the overall outcome of a reproduction attempt, four categories were defined. An outcome is defined as a *Success* if all of the performed experiments achieved *identical* results. If at least one of the experiments achieved either *identical* or *consistent* results, the outcome is defined as *Partial Success*. In the case of all experiments being *different*, the outcome is defined as a *Failure*. Lastly, if no experiment were successfully conducted, the outcome is classified as *No Result*. This is judged as a less favourable outcome than *Failure*.

3.6 Documentation

The code developed for a selection of the reproduction attempts is available on GitHub [4], along with code develop for use in this report. The selection of reproduction attempts cover all different outcomes, reproduction levels, and a variety of areas. The results achieved are also provided for these studies, along with a description of the experiment setup. The full results are provided if it does not require including the original data set.

A Google Forms form [5] was used to register data for the *Article Model Metrics*. These scores were entered during each reproduction attempt. The responses from the Google Forms form is also available [6].

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

[5] https://goo.gl/forms/5eXAC9TOuR97nS063

[6] https://docs.google.com/spreadsheets/d/1ciwZ2GW3EZbS9mCHbiCyTpC1J5xMoXkBuPTPjRCsnP4/edit?usp=sharing

Chapter 4

Results

This chapter presents the results achieved in the reproduction attempts. Section 4.1 lists the 30 studies covered in the experiment, and the results achieved for each reproduction attempt. Section 4.2 presents the system for categorizing issues in reproductions which is one of the contributions of this project. This includes statistics on the number of issues encountered for each paper.

4.1 Reproduction Attempts

The experiment conducted involved 30 articles, 10 for each year studied. The articles are listed in Table 4.1. Out of the 30 papers, 7 were found to be *R1* reproducible, 15 *R2-D* reproducible, and 8 *R3* reproducible. No *R2-M* articles were encountered. Table 4.1 lists the title of each article, their reproduction level, the hours spent in the reproduction attempt, and the id used as identification for the articles in the rest of the tables. Time spent only covers time spent on actual reproduction attempts, and *R3* studies are therefor registered with no time, even though some effort was spent reading the papers and attempting to find associated code and data. More information on the articles can be found in Appendix A. The list of the persons responsible for each reproduction attempt can be found in Table A.1. The count of experiments per article and the outcomes are found in Table A.2. The *Article Model Metric* calculated per component is found in Table A.3.

The results of the reproduction attempts are presented in the Table 4.2. The first column identifies which article the row is connected to. The second row shows how many experiments were identified for the article. The following four columns provide statistics on how many experiments were conducted, and the outcomes achieved. Column seven list the reasons why not all experiments were conducted in the reproduction attempt. There are four different reasons found. *Time*: There was not enough time to complete all the experiments. *Code*: The code in the *R1* article did not cover all the experiments. *Data*: One of more experiments needed data sets not available or only available in a form not suited for the experiment.

Id	*Title*	*Type*	*Year*	*Hours spent*
1	Measuring the Objectness of Image Windows [26]	R1	2012	40
2	Generalized Correntropy for Robust Adaptive Filtering [27]	R2-D	2016	40
3	Development and investigation of efficient artificial bee colony algorithm for numerical function optimization [28]	R2-D	2012	40
4	Blind Image Quality Assessment: A Natural Scene Statistics Approach in the DCT Domain [29]	R1	2012	25
5	Cooperatively Coevolving Particle Swarms for Large Scale Optimization [30]	R2-D	2012	40
6	Learning Sparse Representations for Human Action Recognition [31]	R2-D	2012	40
7	Visualizing and Understanding Convolutional Networks [32]	R2-D	2014	40
8	iSuc-PseOpt: Identifying lysine succinylation sites in proteins by incorporating sequence-coupling effects into pseudo components and optimizing imbalanced training dataset [33]	R2-D	2016	22
9	A modified Artificial Bee Colony algorithm for real-parameter optimization [34]	R2-D	2012	40
10	RASL: Robust alignment by sparse and low-rank decomposition for linearly correlated images [35]	R1	2012	10
11	Classification with Noisy Labels by Importance Reweighting [36]	R2-D	2016	40
12	Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition [37]	R1	2016	20
13	Context Aware Saliency Detection [38]	R2-D	2012	40
14	Distributed representations of sentences and documents [39]	R2-D	2014	40
15	XGBoost: A scalable tree boosting system [40]	R1	2016	40
16	Facial landmark detection by deep multi-task learning [41]	R2-D	2014	40
17	Deep learning-based classification of hyperspectral data [42]	R1	2014	8
18	Semi-supervised and unsupervised extreme learning machines [43]	R2-D	2014	40
19	DeepReID: Deep Filter Pairing Neural Network for Person Re-Identification [44]	R2-D	2014	22
20	Deep neural networks: A promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data [45]	R2-D	2016	8
21	Clustering by fast search and find of density peaks [46]	R1	2014	33
22	DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition [47]	R2-D	2014	40
23	Single image super-resolution with non-local means and steering kernel regression [48]	R3	2012	-
24	Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease [49]	R3	2012	-
25	Robust text detection in natural scene images [50]	R3	2014	-
26	Towards end-to-end speech recognition with recurrent neural networks [51]	R3	2014	-
27	Mastering the game of Go with deep neural networks and tree search [52]	R3	2016	-
28	Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning [53]	R3	2016	-
29	MLlib: Machine learning in Apache Spark [54]	R3	2016	-
30	Learning Rotation-Invariant Convolutional Neural Networks for Object Detection in VHR Optical Remote Sensing Images [55]	R3	2016	-

Table 4.1: Information about the articles used in the reproduction attempts.

Id	*Total number of experiments in article*	*% of results that were identical*	*% of results that were consistent*	*% of results that were failures*	*% of experiments that were not conducted*	*Why not all experiments were conducted*	*Overall outcome*
1	18	0%	22%	0%	78%	Time	Partial Success
2	4	25%	0%	25%	50%	Time	Partial Success
3	46	37%	37%	26%	0%	-	Partial Success
4	10	0%	10%	0%	90%	Code	Partial Success
5	21	0%	19%	14%	67%	Time	Partial Success
6	14	0%	0%	7%	93%	Time	Failure
7	20	0%	0%	0%	100%	Time	No Result
8	1	0%	100%	0%	0%	-	Partial Success
9	68	0%	1%	10%	88%	Time	Partial Success
10	31	0%	0%	77%	23%	Code	Failure
11	6	0%	0%	17%	83%	Data, Time	Failure
12	4	0%	25%	25%	50%	Code	Partial Success
13	2	0%	0%	100%	0%	-	Failure
14	3	0%	0%	0%	100%	Time	No Result
15	4	0%	50%	0%	50%	Time	Partial Success
16	12	0%	0%	0%	100%	Time	No Result
17	16	19%	19%	0%	63%	Code	Failure
18	38	0%	0%	13%	87%	Time	Failure
19	6	0%	0%	0%	100%	Time	No Result
20	10	0%	0%	0%	100%	Time	No Result
21	7	86%	0%	0%	14%	Presentation	Success
22	4	0%	0%	25%	75%	Time	Failure

Table 4.2: Summary of the results of the reproduction attempts: The number of experiments in the original article and the status of these in the reproduction attempt. In addition, the overall reproduction outcome for each article.

Presentation: The result of an experiment was presented in such a way that it would be difficult to compare the results from the reproduction with the original article. The last column present the one of four possible outcomes of the reproduction. *Success*: All experiments conducted had identical results. *Partial success*: At least one experiment reproduced had either identical or consistent results with the article. *Failure*: All of the experiments conducted were failures. *No Result*: No experiment was fully conducted in the reproduction. It must be noted that the overall outcome of a reproduction attempt reported here should not be interpreted as a definitive statement about whether the study can or cannot be reproduced. The reproduction attempts in this project are carried out with limited time and resources, and by a limited group of reproducers. Several of the studies on whose reproduction we report *Failure* or *No Result* may be reproducible given more time, or by other researchers. However, since it is our belief that reproductions should be easy and well facilitated, we do not believe that these disclaimers invalidate the results presented in the following parts of this report.

4.2 Discrepancies

During the reproduction attempts several kinds of issues were encountered, complicating the reproductions. The issues cover everything from poor documentation to lack of resources, and are characterized by the fact that they increase the difficulty of the reproductions and introduce sources of error in the outcomes. The **second hypothesis** of this project is that the issues which complicate reproductions are the same across studies. We predicted that the issues encountered in the reproductions could be grouped into categories. To test this, a system of categories was created, with the categories covering all major issues encountered. The system provides a method for systematically and quantitatively identifying how a reproduction attempt diverged from the original experiment it aimed to reproduce, and provides a way of identifying the most common issues over all reproductions.

The issues encountered in the reproductions are referred to as *discrepancies* since they introduce differences between the original and reproduced experiment. Three main types of discrepancies were identified immediately after the reproductions: *Problems*, *Assumptions*, and *Errors*. Each of these types cover a broad number of issues, and are further divided into sub-categories, with each category describing a particular type of issue.

This section presents the system of discrepancy categories proposed in this project. The categories are based on the issues observed during the reproduction attempts. Each type of discrepancies is discussed, and their sub-categories are listed and explained. Statistics on the number of observed instances of each discrepancy category is also reported. Lastly, statistics are given on the number of discrepancies identified for each paper which was attempted reproduced. Since discrepancies are discovered during reproductions, the statistics are limited to the 22 papers which were attempted reproduced. Complete data on which discrepancies were encountered for which papers are reported in Appendix B.

4.2.1 Problem Categories

Problem discrepancies are difficulties encountered as a result of missing or unclear documentation, or the reproducers not having access to necessary resources such as hardware. Many of these issues would be improved by more openness and more thorough descriptions from the original authors. During the reproductions it was found that most papers had one or more problems. Furthermore, identical or similar problems were often found across different studies. After the completion of all reproduction attempts the problems encountered were therefore grouped into categories. In total, 20 categories of problems were found. Table 4.3 lists the identified problem categories. As seen by the descriptions, some of the problems are only relevant for *R1* reproducible studies. All other problems affect both *R1* and *R2-D* reproducible studies.

Description of Problem Categories

This section contains a more detailed description of each problem category, along with examples which were encountered during our reproduction attempts.

Problem category *P1* concerns the situation where the authors of an article have shared an implementation of their method, but not the experiment code. Alternatively, if an implementation of experiments is provided, it does not cover all experiments. Category *P1* was observed in several *R1* papers, including "Measuring the Objectness of Image Windows" [26]. When reproducing the experiments of this paper the method implementation could be used without modification, but the experiment, including loading of data and comparison of predictions with ground truths, had to be reimplemented.

Category *P2* covers instances where the authors of an article have shared method code, but where the implementation does not cover the entire method. This was an issue for "Blind Image Quality Assessment: A Natural Scene Statistics Approach in the DCT Domain" [29]. For this paper, which trains a model for evaluating image quality, the method for testing the model was provided, but not the method for training it. *R1* reproduction attempts are therefore limited to using those learned models which were provided with the code.

Category *P3* covers issues of poor documentation in implementations. In some cases where researchers have shared an implementation of their method or experiments, the code lacks good documentation. This makes the code unnecessary difficult to understand and interpret. It is particularly impactful in cases where the reproducers are uncertain about how well the available implementation covers the proposed method or experiments. A problem of category *P3* was encountered during the reproduction of "Blind Image Quality Assessment: A Natural Scene Statistics Approach in the DCT Domain" [29]. Some sections of the method are commented out without sufficient explanation, and it is difficult to determine if these sections were used in the original experiment.

Category *P4* concerns situations where method or experiment code is shared, but where not all parameters are included, or where the included parameters differ from the parameters described in the associated article. This problem was encountered during the reproduction of "Deep Convolutional and LSTM Recurrent Neural Networks for Multi-modal Wearable Activity Recognition" [37]. Implementations of both the method and one of the experiments were provided for this paper. As part of the experiment code, parameters for sliding window length and step size were defined. However, their values appear to differ from those given in the paper. Furthermore, when using the parameters provided with the available implementation, 9894 instances are classified. In the paper, it is reported that 16 900 instances were classified.

Category *P5* covers issues relating to versioning of provided implementations. Ideally, any piece of code shared from a study should be versioned, and the research paper documenting the study should mention which version was used in the experiments. This allows reproducers to ensure that the code they have available is the same as was used in the original experiments. In many cases the shared code is not versioned, but even when it is, such as for "Measuring the Objectness of Image Windows" [26], the gain is limited if the research paper does not state which version was used in the experiments.

Category *P6* covers the issue of code being shared in a compiled or non-inspectable form. In these cases the program can be executed, but the source code cannot be inspected, and reproducers cannot verify that the code implements the method described in the associated paper. When attempting reproduction of "Context Aware Saliency Detection" [38], a MATLAB implementation was available. However, it was shared in the MATLAB protected format, which is obfuscated, and *R1* reproduction could therefore not be performed.

Category *P7* concerns issues relating to random numbers and random number generators. Many methods and experiments in AI involve some level of randomness, for example in random initialization of weights or random partitioning of data. However, the random numbers used, or the random number generator and its seed, is rarely provided. How sensitive a method is to variations in random initialization may vary greatly, and even if a method is insensitive to random initialization, experiments often employ random partitioning of data which may influence the results, e.g. when partitioning data into training and test sets in supervised learning. Problem category *P7* is intended to cover all problems arising from the lack of documentation of random numbers. Instances are only recorded for studies where random numbers are used in a significant way, i.e. where change in the random numbers influence the results. *P7* was a problem during the reproduction of "Cooperatively Coevolving Particle Swarms for Large Scale Optimization" [30]. The paper's function optimization method is tested on seven test functions, each of which is shifted with a random value in each dimension. However, these random values were not shared. Sharing these random vectors would have ensured that any reproduction attempt would have access to the exact same functions as was used in the original experiment.

Category *P8* concerns issues relating to how the method of a study is described in the research article. Even though the presentation of a new method often is the main purpose of a research article, some aspects of the method may be accidentally omitted, or poorly described. This may cause significant issues for reproduction teams, particularly when reimplementation of the method is required, because the reproduction team is forced to make assumptions about these aspects. A problem of category *P8* was encountered during the reproduction of the above mentioned "Cooperatively Coevolving Particle Swarms for Large Scale Optimization" [30]. The proposed particle swarm optimization method searches for a vector of variables which minimizes the given function. Each variable in the vector has a range of legal values it can take, but the method for ensuring that a variable remained within its range was not described in the paper.

Category *P9* is similar to *P8*, but covers issues relating to the description of the experiment, rather than the method. In order to correctly reproduce results, the experiments of the original study must be correctly recreated. However, in some cases parts of an experiment are not described, or poorly described. For example, in "Visualizing and Understanding

Convolutional Networks" [32], the method for estimating the accuracy of the validation data after data pre-processing was not well described.

Category *P10* concerns problems arising from unclear description of the implementation of the method or experiment. Even if a method or experiment is well described, issues and ambiguities may arise during implementation. An example of this problem category is provided in "Classification with Noisy Labels by Importance Reweighting" [36]. The method of this study, for performing classification in the presence of noisy or corrupted labels, is well described through mathematical equations, but how it was implemented in practice was difficult to understand. The lack of a clear description of the implementation caused significant delays to the reproduction.

Category *P11* covers a specific type of problem where a paper suggests multiple methods for solving a task, but does not mention which method was used in their implementation. The suggested methods are often references to other scientific papers. While providing alternative methods for solving a problem may be positive, a paper should mention which method was used to help reproducibility. A problem of category *P11* was encountered in "iSuc-PseOpt: Identifying lysine succinylation sites in proteins by incorporating sequence-coupling effects into pseudo components and optimizing imbalanced training dataset" [33]. In this paper three methods for inserting hypothetical examples into training data were suggested, all referencing different papers, but the paper did not mention which variation was used in the experiments.

Category *P12* covers the problem where a set of trained weights or parameters from a supervised learning method are shared online, but where these weights are not the same as the ones used in the original experiments. Training a supervised learning model may take significant time or, in the case of *R1* reproduction attempts, be impossible given the available code. Researchers sharing the weights they have trained is therefore helpful for the purpose of reproduction, and for other researchers who wish to use the code. However, in some cases the weights which are shared are not the same as the ones used in the original experiment. For example, they might be trained on different data sets. This was an issue during the *R1* reproduction of "Blind Image Quality Assessment: A Natural Scene Statistics Approach in the DCT Domain" [29]. As mentioned in the discussion of category *P2*, this paper had not shared the code for training the weights. Several sets of weights trained on different subsets of the data set had been included, making the testing part of the method useful, but only one of the subsets matched a subset used in the original experiments. As a result, only one experiment from the article could be *R1* reproduced.

Category *P13* covers the issue of some experiment- or hyper-parameters not being provided. Having the correct values for experiment- and hyper-parameters is highly important for reproducing an experiment. However, in some cases not all of these values are provided in the research paper. For example, in "XGBoost: A scalable tree boosting system" [40], the number of epochs for which the model was trained was not provided, even though the number of epochs significantly influenced the results. The absence of experiment- and hyper-parameters force reproducers to estimate these themselves, and this introduces unnecessary difficulties in the reproduction process.

Category *P14* covers problems arising from there being an error in a paper. In this project, only one possible case was identified, but problems of this category may occur. The case encountered was during the reproduction of "A modified Artificial Bee Colony algorithm for real-parameter optimization" [34], where the mathematical formula for one of the test functions (the Schwefel function) did not match the common definition.

Category *P15* concerns the situation where a data set is found online whose name matches a data set referenced in a paper, but where there is a mismatch between the content of the two data sets. In cases where the available data set is larger than the one described in the paper, the correct data set can sometimes be extracted. In other cases the mismatch is too significant to reconcile. The Weizmann Action and Robustness data set [56] as described in "Learning Sparse Representations for Human Action Recognition" [31] contains 90 examples. However, the version available online contains 93 examples. In this case the additional three examples appeared to be alternative versions of other examples, and a subset of 90 examples could be extracted.

Category *P16* also concerns problems of data availability. It specifically concerns the issue of a paper using a subset of a known data set, but that subset not being shared. Not having access to the correct subset of a data set can in some cases be equivalent to not having access to the data set. In one of the experiments of "Context Aware Saliency Detection" [38], a subset of 100 examples was used from a data set of 1003 examples, but the subset was not specified in the referenced article.

Category *P17* concerns the problem of augmented and pre-processed data not being shared. In some situations a study will use a known data set in its experiments, but will pre-process the data or augment it with hypothetical data before use. The processed data created in this process is in many cases not shared, and since the pre-processing procedure is not an integral part of the method, it may in some cases be described less thoroughly than other aspects. Category *P17* was encountered during the reproduction of "iSuc-PseOpt: Identifying lysine succinylation sites in proteins by incorporating sequence-coupling effects into pseudo components and optimizing imbalanced training dataset" [33]. In this study, the original data set, which was shared, is augmented by the insertion of hypothetical positive examples, and the removal of redundant negative examples. The augmented data set was not shared and the description was ambiguous at times, for example in the description of the method for inserting hypothetical examples, as mentioned in the discussion of problem category *P11*.

Category *P18* covers problems relating to the partitioning of data into training, validation, and test sets. In supervised learning methods, a data set is usually partitioned into training, validation, and test sets. This partitioning may either be part of the original data set, or be performed as part of each study. The k-fold cross-validation method is a popular method for partitioning data and for testing supervised learning methods. Partitioning of data affects which data is used in the training and test stages of a method, and will in many instances affect the results of a method. Still, in cases where the partitioning is not a part of the original data set, the partitioning is rarely shared. Furthermore, the process by which the data is partitioned is not always well described. In "Classification with Noisy Labels by Importance Reweighting" [36], each data set is partitioned independently 10 times. During each partitioning 75% of the data is assigned for training, and 25% for testing. However, this description is not sufficient for reproducing the exact partitions.

Category *P19* concerns the problem of result presentations. In order for results to be reproducible they need to be quantitative in some manner, and it must be possible for a reproducer to estimate the error between reproduced and original results. However, even when results are presented quantitatively, they may be presented in a manner which makes exact comparison of results difficult. Specifically, when data is presented only in a visual plot, reading the correct values may be difficult and may introduce unnecessary error. This was a problem during the reproduction of "Generalized Correntropy for Robust Adaptive Filtering" [27], where the results were presented in a series of plots.

Lastly, category *P20* covers problems caused by reproducers lack of access to specialized hardware and software. Some experiments proposed are highly resource demanding, and can only be executed on very powerful hardware. Similarly, some methods may rely on specialized software which is not publicly available. Having such requirements in an experiment may in some cases be unavoidable, but it can make reproduction significantly more difficult, and limit who can attempt reproduction. In this project, *P20* was only an issue during the reproduction of, "Visualizing and Understanding Convolutional Networks" [32], whose experiments were performed on large data sets requiring powerful hardware.

Problem Category Results

Figure 4.1 shows the number of instances of each problem category encountered during the reproduction attempts. *R1* and *R2-D* studies are separated and stacked to show the difference in problems encountered by these types of reproductions. Categories *P1-P5* concerns problems only applicable for *R1* studies. It is observed that categories *P9* and *P10* are the most frequent categories, with 12 of 15 *R2-D* studies having at least one issue of category *P10*. Categories *P9* and *P10* cover problems encountered due to poor documentation of the experiment or method implementation.

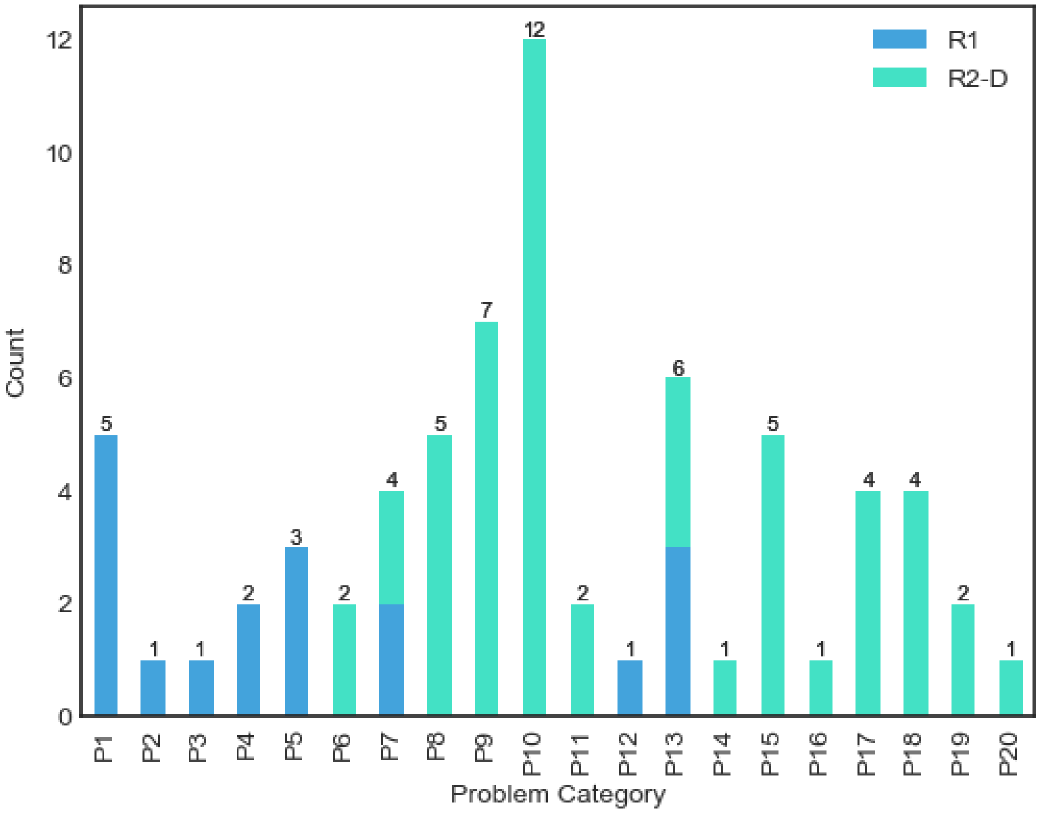

Figure 4.1: Stacked bar plot of number of observed instances of each problem category

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

Table 4.3: List of problem categories identified in the project

4.2.2 Assumption Categories

Assumption discrepancies are potential deviations introduced consciously by the reproducers as a result of ambiguities or missing access to documentation or resources. Assumptions may in some cases be unavoidable, and may be directly related to problem discrepancies. Similarly to problems, the assumptions made for different studies often have similarities, and they can be grouped into assumption categories. 15 assumption categories were found across all reproduction attempts, and are listed in Table 4.4.

Description of Assumption Categories

This section provides a description and example of each assumption category.

Assumption category *A1* covers the assumption that the code which is shared and found online is the same as was used in the original experiments of a study. This assumption is necessary in almost all *R1* reproductions, since few papers link directly to a shared implementation and provide proper versioning. As discussed in section 3.5, measures should be taken when searching for implementations to increase the validity of this assumption. The specific form of the assumption may vary between studies. For example, for "Measuring the Objectness of Image Windows" [26] the main assumption is that the version available online is equivalent to the version used in the original experiments.

Category *A2* is similar to *A1*, but concerns assumptions about experiment- and hyper-parameters shared being the same as were used in the original experiments. These parameters may be part of the code shared online. This assumption should be tested by inspecting the parameter values in the code, and comparing them with those reported in the associated paper. Ideally, if all parameters are clearly given in both the code and paper, and have identical values, then no assumption is necessary. In the code shared for "Deep Convolutional and LSTM Recurrent Neural Networks for Multi-modal Wearable Activity Recognition" [37], all experiment parameters are provided. Some of these values appear to be different from the values described in the paper, but are assumed to be correct.

Category *A3* covers the specific assumption that when using code created by others, a minor change to the code does not impact the results. By minor change, it is meant a change which does not affect the method and which does not cause the implementation to deviate from the experiment description. The change should only be made to ensure that the code can be executed on the available system. This assumption was made during the reproduction of "Deep Convolutional and LSTM Recurrent Neural Networks for Multi-modal Wearable Activity Recognition" [37], when an import statement and system flag set operation was added to the top of the code to facilitate running the Theano library on our available GPU.

Category *A4* concerns assumptions made about terms or concepts in the face of ambiguous descriptions. In some cases the language of a paper may be slightly unclear, introducing ambiguity about what is actually meant by a term. In these cases an assumption must be made about how to interpret the ambiguity. In the reproduction of "Learning Sparse Representations for Human Action Recognition" [31], the concept "improved Harris keypoint detector" was interpreted as the Harris-Laplace keypoint detector.

Category *A5* is related to problem category *P8*, and concerns the assumptions which are made about an aspect of a method which is not well described. However, the assumptions categorized under category *A5* are based on how similar aspects have been treated in other papers. Assumptions based on other research papers are believed to be better founded than assumptions not based directly on examples from other papers. An example of this kind of assumption comes from the reproduction of "Cooperatively Coevolving Particle Swarms for Large Scale Optimization" [30]. As mentioned in the discussion of problem category *P8*, the method for ensuring that all variables remained within their legal range was not described in the paper. It was therefore assumed that the paper used the same procedure as was used in the update function described in "Development and investigation of efficient artificial bee colony algorithm for numerical function optimization" [28].

Category *A6* covers assumptions which are similar to those of *A5*, except that they are not based on descriptions from other research papers. Ideally, assumptions should be based on examples from other research papers, but such examples are not always available. The distinction between assumptions of category *A5* and *A6* is considered important because assumptions of category *A5* are believed to have a stronger foundation. "Cooperatively Coevolving Particle Swarms for Large Scale Optimization" [30] also provides an example of an *A6* assumption. In the method proposed in the paper, the termination criterion is the number of function evaluations performed. The pseudocode of the method indicates four function evaluations per variable (particle) per iteration. However, in the practical implementation the number of evaluations can be reduced to approximately one per particle per iteration. In this situation the practical implementation with minimum number of function evaluations was chosen.

Category *A7* is directly related to problem category *P11*. When a paper suggests multiple methods for solving a task, an assumption must be made about which to use. This assumption should be based on which method appears best or most likely, but this is not always obvious. As mentioned in the discussion of problem category *P11*, "iSuc-PseOpt: Identifying lysine succinylation sites in proteins by incorporating sequence-coupling effects into pseudo components and optimizing imbalanced training dataset" [33] references three possible methods for inserting hypothetical examples. In the reproduction, two of the methods were implemented, and the best performing method was assumed to be correct.

Category *A8* concerns assumptions about the use of third party libraries. In some situations a paper will use a commonly known algorithm or function as part of its method, but not provide details on how this was implemented. For these algorithms, public third party implementations are often available. A reproducer should in these cases be able to use a third party implementation rather than reimplementing the algorithm. However, the implicit assumption made by the reproducer is then that the third party implementation is equivalent to the implementation used in the original experiment. Category *A8* aims to make these assumptions explicit. In "Classification with Noisy Labels by Importance Reweighting" [36], the Kullback-Leibler Importance Estimation Procedure was used for density ratio estimation, but no implementation details were given. In the reproduction it was therefore decided to use a third party implementation of this algorithm, and it was assumed that this implementation would produce results equivalent to the implementation used in the original study.

Category *A9* also concerns third party libraries, but specifically the use of third party libraries which are mentioned in the original study. In some cases a study will use a third party library and provide the name of that library. However, the version of the library used may not be provided. In these cases it may be reasonable to use the newest version of the library available, and assume that this version produces results which are equivalent to the version used in the original study. In the reproduction of "DeCAF: A Deep Convolutional Activation Feature for Generic VisualRecognition" [47] the Caffe library [57] was used in place of its predecessor DeCAF, which was deprecated at the time of reproduction.

Category *A10* is related to problem category *P12*, and concerns assumptions about shared trained weights and parameters. In cases where a study based on supervised learning shares weights or trained parameters, using these may significantly ease reproduction. However, the weights must then be assumed to be the same as were used in the original experiments. Pre-trained parameters were included in the code from "Measuring the Objectness of Image Windows" [26]. These were assumed to be the same as were used in the original experiments, and were used in the reproduction.

Category *A11* concerns assumptions about parameter values. In some situations a study does not include all experiment- or hyper-parameter values. In these instances an assumption must be made about the value of the parameter, with the goal being to use parameters which are as close to the ones used in the original study as possible. In "Learning Sparse Representations for Human Action Recognition" [31] one of the hyper-parameter values, *error threshold*, was not provided. The paper stated that the parameter value was found empirically, and provided a suggestion for estimating it. The suggested method was used to estimate the parameter value, and an assumption of category *A11* was made.

Category *A12* is related to problem category *P17*. When a study pre-processes or augments a data set before use, the processed data is sometimes not shared. In these instances the augmentation and pre-processing steps must be performed by the reproducers. Even if the reproducers cannot recreate identical processed data, the recreated processed data can be assumed to be equivalent to the original data. As mentioned in the discussion of problem category *P17*, the paper "iSuc-PseOpt: Identifying lysine succinylation sites in proteins by incorporating sequence-coupling effects into pseudo components and optimizing imbalanced training dataset" [33] employs data augmentation. The augmentation process had to be reimplemented during the reproduction attempt, but the new data was assumed to be equivalent to the original for testing purposes.

Category *A13* concerns assumptions made about how to partition a data set. In supervised learning methods, a data set is commonly partitioned into training, validation, and test sets. In some instances this process is not well described, and assumptions must be made about how it should be performed. In "XGBoost: A scalable tree boosting system" [40] two test sets were provided for one of the data sets, but the paper did not state which was used. Based on their performance, the second test set was assumed to be the correct one.

Category *A14* concerns assumptions about how to use a data set in the face of unclear description. In some cases the correct data set is available, but a study is unclear about how it is used. An assumption must then be made about how to use the data. In "Context Aware Saliency Detection" [38] one of the data sets provided four ground truths for each image, but it was not stated how they were used. An assumption was made that each prediction was compared to all ground truths, and that the average score for all ground truths was used as result.

Category *A15* covers the specific assumption that when running an experiment where the hardware is described in the study, the experiment can be reproduced on different hardware. For "Deep Convolutional and LSTM Recurrent Neural Networks for Multi-modal Wearable Activity Recognition" [37], information about the GPU used to run the convolutional neural networks was provided. Since the exact GPU was not available to our reproduction team, it was assumed that similar results could be achieved on a different GPU.

Assumption Category Results

Figure 4.2 shows the number of instances of each assumption category encountered during reproduction, with *R1* and *R2-D* separated and stacked. Assumptions *A1-A3* are only applicable for *R1* studies. Assumption category *A6*, which covers assumptions about the method not taken directly from other articles, is the most common, being encountered in over half of the *R2-D* studied investigated.

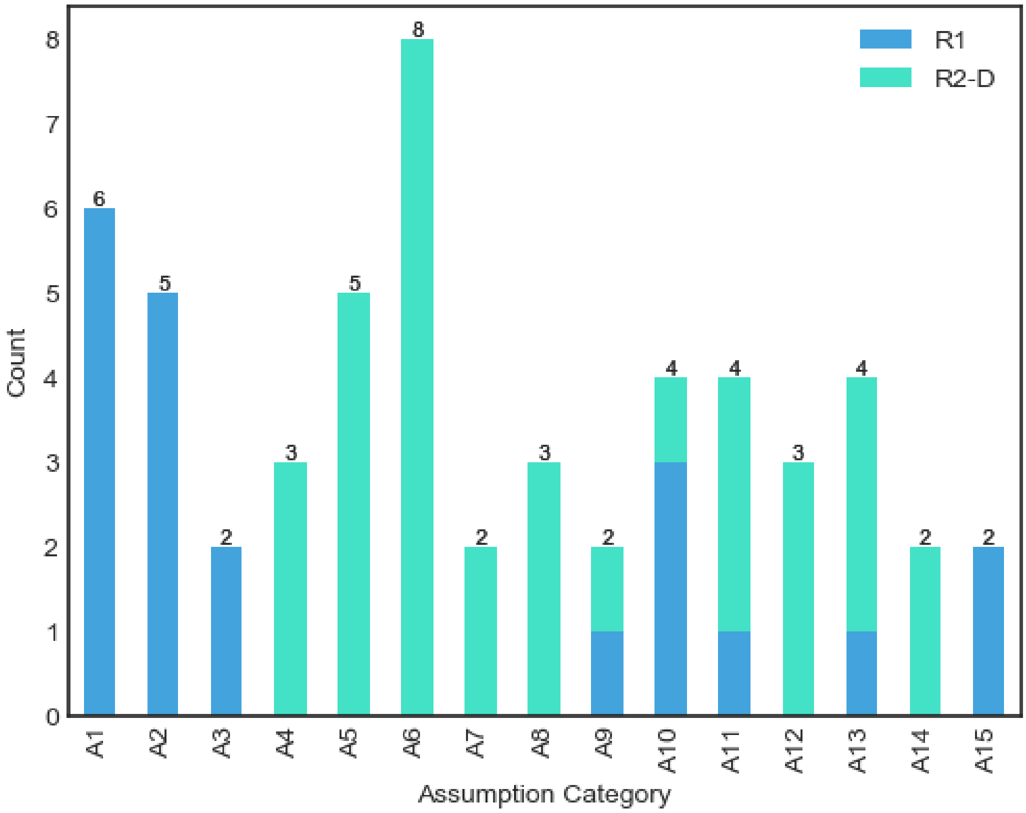

Figure 4.2: Stacked bar plot of number of observed instances of each assumption category

Id	*Description*
A1	For R1 study, the code available is assumed to be the same as was used in the original experiments
A2	For R1 study, the parameter values in available code are assumed to be the same as was used in the original experiments
A3	For R1 study, a minor change to the code to facilitate running on our hardware is assumed to not affect results
A4	An assumption is made about how to interpret a term or concept which is ambiguous
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
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
A7	For aspects where multiple methods are suggested, an assumption is made about which to use
A8	A third party implementation of a method is assumed to be similar to the original implementation
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
A10	Trained weights or trained parameters shared online are assumed to be the same as were used in the original experiments
A11	Assumption is made about one or more parameter values
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
A13	An assumption is made about how to partition a data set
A14	An assumption is made about the usage of a data set
A15	The hardware is assumed to not significantly influence the results of the experiment

Table 4.4: List of assumption categories identified in the project

4.2.3 Error Categories

As discussed in Section 3.5, a reproduction attempt will often produce results which are not identical to the results of the original study. Error discrepancies are the possible error sources which may be the direct cause of differences in results. The exact cause of difference in results is rarely known, but predicting the most likely errors is useful to understand what the most common error sources are. Just as for problems and assumptions, the errors identified for one study often share similarities with errors identified for another study, and they can be grouped into categories. Table 4.5 lists the 15 error categories identified in this project.

Some of the error discrepancies cannot be determined to be present with full certainty, for example *E5* and *E6*. However, they are counted if they are deemed likely to be present. Even though certainty cannot be achieved, identifying error discrepancies can still be valuable, as it gives an estimate of the most probable sources of difference between original and reproduced experiment.

Description of Error Categories

This section provides a description and example of each error category. Each error category represents one *possible* source of error, and not a definitive error in a reproduction.

Error category *E1* covers the possible errors resulting from code used in *R1* reproduction not being identical to the code used in the original study. As discussed in the paragraph on assumption category *A1*, code available online is sometimes assumed to be identical to the code used in the original study. However, in some cases this assumption can be challenged, and error category *E1* then becomes a possible source of error. An *E1* error is believed to be among the most likely sources of deviation in the reproduction of "Measuring the Objectness of Image Windows" [26], since we lack information about the version of the code used in the original experiment.

Category *E2* covers possible errors resulting from faults in the reproducers assumptions about the method. As mentioned in the discussion on assumption categories *A5* and *A6*, reproducers must sometimes make assumptions about an aspect of the method which is not well described. While these assumptions may be well founded, they introduce possible error in the reproduction process. In the reproduction of "Cooperatively Coevolving Particle Swarms for Large Scale Optimization" [30] two major assumptions were made about the method. Error category *E2* is therefore a likely source of error for this reproduction.

Category *E3* covers errors resulting from faults in the reproducers assumptions about the experiment. Just as for the method, assumptions must sometimes be made about the experiment of a study, and these introduce possible deviation in the results. In the reproduction of "Learning Sparse Representations for Human Action Recognition" [31] different assumptions were made about the experiment. These can have resulted in the reproduced experiment being different from the original, and be the source of differences in results.

Category *E4* covers errors resulting from faults in the reproducers assumptions about how to implement the method. When implementing a method, some assumptions often have to be made, even in the presence of good method description. However, these assumptions can be mistaken and result in deviations. In the reproduction of "Cooperatively Coevolving Particle Swarms for Large Scale Optimization" [30], an assumption is made about the number of function evaluations per iteration, as mentioned in the discussion of assumption category *A6*. If this assumption is wrong, it is probable that it influences the results of the reproduction.

Category *E5* is a generic error category covering all errors caused by unknown faults in the reproducers implementation of the method of a study. During *R2-D* reproduction of a study there is a possibility of unknown errors appearing in the implementation of a method. These errors might be the result of implicit assumptions from the reproducer, wrong interpretation of a concept, or simple mistakes. *E5* errors are possible sources of deviation in most reproductions, but are more likely in complex reimplementations, such as that for "Learning Sparse Representations for Human Action Recognition" [31].

Category *E6* is similar to *E5*, but covers errors caused by unknown faults in the reimplementation of an experiment rather than method. These errors can also be caused by wrong implicit assumptions, misinterpretations, or pure mistakes. *E6* errors are likely to be sources of deviation for reproductions where the experiment implementation is complex and involves multiple steps, for example in data pre-processing. It is believed to be one of the most probable error sources for the reproduction of "iSuc-PseOpt: Identifying lysine succinylation sites in proteins by incorporating sequence-coupling effects into pseudo components and optimizing imbalanced training dataset" [33].

Category *E7* covers errors resulting from the use of third party libraries in reproduction. As mentioned in the discussion of assumption category *A8*, third party implementations of known algorithms can be used in reproduction attempts. These are then assumed to be equivalent to the implementations used in the original study. However, if this assumption is mistaken, the use of these third party implementations can result in deviations in result. In the reproduction of "Classification with Noisy Labels by Importance Reweighting" [36], an implementation of the Kullback-Leibler Importance Estimation Procedure is used. This implementation might be different from the implementation used in the original experiments, and may cause difference in results.

Category *E8* covers errors resulting from the use of trained weights and parameters which are not identical to those which were used in the original study. In certain cases a study will share trained weights or model parameters, which can be assumed to be the same as those which were used in the original experiment. These are useful for reproduction, but if they are not the same weights as were used in the original study, their use may result in outcome deviations. In the reproduction of "Blind Image Quality Assessment: A Natural Scene Statistics Approach in the DCT Domain" [29] available trained parameters were used. The reproduction did not receive results which were identical to those reported in the original paper, and it is probable that this was due to the trained parameters not being identical.

Category *E9* concerns errors resulting from the use of experiment- or hyper-parameters which are different from those used in the original study. Using different hyper-parameters are likely to result in different outcomes, and in cases where an assumption has been made about a parameter value, error category *E9* is a likely source of deviation. In the reproduction of "Learning Sparse Representations for Human Action Recognition" [31], an assumption is made about one of the parameter values, and the assumed value is unlikely to be identical to the value used in the original study. *E9* is therefore a probable source of deviation in results for this paper.

Category *E10* covers errors resulting from the use of different random numbers in the original study and reproduction. As mentioned in the discussion of problem category *P7*, many methods and experiments include some random elements. If the random numbers selected have significant influence on the results the use of different random numbers can cause deviation in results between the original study and its reproduction. Random numbers can have particularly important influence when they are used to augment data, since using two different random value sets will result in two different data sets. An error of category *E10* is suspected to be a factor in the failure of the reproduction of "Classification with Noisy Labels by Importance Reweighting" [36], where noise is added to the data through a random process.

Category *E11* concerns errors resulting from the use of pre-processed or augmented data. As mentioned in the discussion of problem category *P17*, some studies will augment or pre-process their data before use in experiments, but may not share the processed data. If this data is not shared, and the augmentation or pre-processing is reimplemented by reproducers, this creates a possible source of deviation. The data pre-processing of "iSuc-PseOpt: Identifying lysine succinylation sites in proteins by incorporating sequence-coupling effects into pseudo components and optimizing imbalanced training dataset" [33] involves both the insertion of hypothetical examples and removal of redundant examples. Since this process was reimplemented in the reproduction it is a likely error source.

Category *E12* concerns errors arising from the use of a wrong data set. If the data set used in the reproduction of an experiment is different from the data set used in the original experiment, the reproduction is likely to produce different results. In most cases data sets should be referenced well enough that no ambiguity exists. However, if a study uses a subset of a data set, the description of the subset selection might be unclear, and introduce uncertainty. In the reproduction of "Context Aware Saliency Detection" [38] the subset used in one of the experiments was not precisely defined, and it is likely that a different subset was used in the reproduction.

Category *E13* concerns errors resulting from wrong partitioning of data into training, validation, and test sets. When the partitioning used in a reproduction is different from the partitioning of the original experiments, it introduces a possible source of difference in results. *E13* might be a likely source of deviation in the reproduction of "Classification with

Noisy Labels by Importance Reweighting" [36], where random selection was used to partition data into training and test sets.

Category *E14* concerns errors resulting from the use of hardware which is not identical to the hardware of the original experiment. For most experiments hardware is assumed to not influence the results. However, in cases where no other difference between original and reproduced experiment is identified, hardware differences becomes the most likely source of result deviation. This is the case for the reproduction of "Deep Convolutional and LSTM Recurrent Neural Networks for Multi-modal Wearable Activity Recognition" [37], where neither experiment nor method was reimplemented in reproduction.

Category *E15* concerns errors resulting from difficulties in comparing the results of a reproduction with those presented in the original study. This is primarily an issue in the reproduction of studies where results are presented qualitatively, or in a quantitative manner unsuitable for reproduction. It is considered the most important error source in the reproduction of "RASL: Robust alignment by sparse and low-rank decomposition for linearly correlated images" [35], where the output of the shared method could not be compared with the results presented in the original study.

Error Category Results

Figure 4.3 shows the number of instances encountered for each error category, split between *R1* and *R2-D* studies. Category *E1* is only applicable to *R1* studies. The most common error categories are *E5* and *E6*, which cover the possible errors resulting from mistakes in the implementation of method and experiment. Since reproduction attempts with the outcome *No result* do not report error type issues the total set of papers is limited to 17. It should again be noted that counted errors are not guaranteed to be present, but are considered likely.

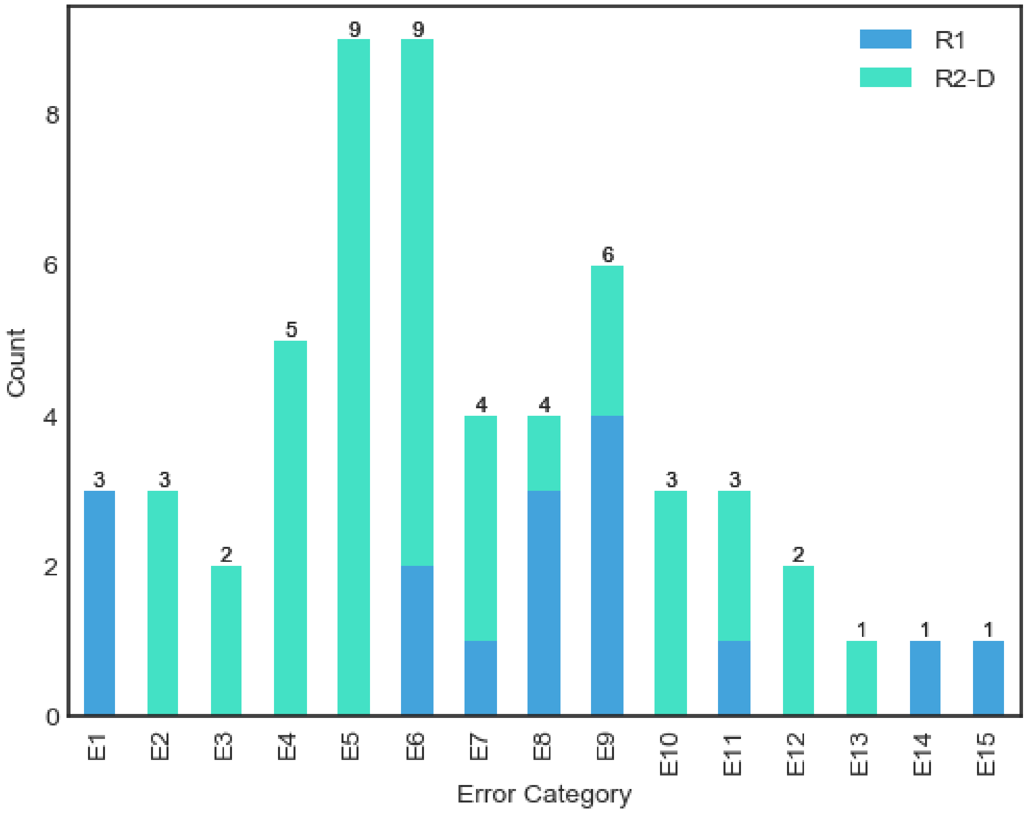

Figure 4.3: Stacked bar plot of number of observed instances of each error category

Id	*Description*
E1	For R1 study, the code available is not exactly the same as was used in the original experiments
E2	There are faults in our assumptions about the method
E3	There are faults in our assumptions about the experiment
E4	There are faults in our assumptions about the implementation
E5	There are unknown faults in our implementation of the method
E6	There are unknown faults in our implementation of the experiment
E7	An implementation in a third party library is not equivalent to the implementation used in the original experiment
E8	The trained weights or trained parameters shared are not the same as were used in the original experiment
E9	The hyper-parameters are not the same as were used in the original experiment
E10	The randomness in the method or experiment, and the lack of shared random numbers and random number generator, influences the result
E11	Augmented and pre-processed data set used in reproduction is not equivalent to the data set used in original experiment
E12	The data subset used in reproduction is not the same as was used in original experiment
E13	The partitioning of data into training, validation, and test set is not the same as was used in the original experiment
E14	Differences in hardware influenced the result
E15	The reproduced results are difficult to compare to the original results

Table 4.5: List of error categories identified in the project

4.2.4 Number of Discrepancies Per Paper

Figure 4.4 shows a heatmap of discrepancy distribution across the papers. The distribution is provided for each discrepancy type, and the combination of all discrepancies. To make the comparison easier, the data is normalized by the highest occurrence within each column. It should be noted that papers whose reproduction ended in *No Result* do not report error discrepancies. The total number of discrepancies are therefore artificially low for these papers. We observe that there is significant variation in the the distribution of discrepancies across the papers. Paper 18, "Semi-supervised and unsupervised extreme learning machine" [43], has the most assumption and error discrepancies, and most overall. In comparison, paper 10, "RASL: Robust alignment by sparse and low-rank decomposition for linearly correlated images" [35], has very few discrepancies of any type. It is also observed that every paper in the study reports at least one discrepancy of the problem type, but not necessarily an assumption.

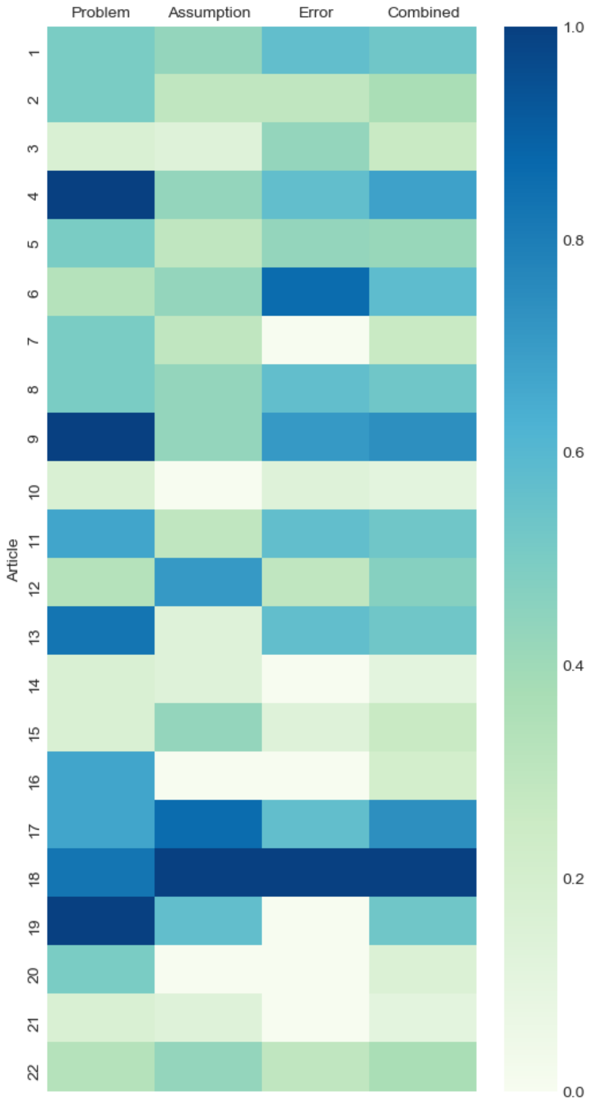

Figure 4.4: Heatmap of discrepancies per paper. Data is normalized by the highest value in each column. Papers 7, 14, 16, 19, and 20 ended in *No result* and do not report error discrepancies.

Figure 4.5 shows the actual number of observed discrepancies of each type encountered in the reproduction of each paper. Though no new pattern is immediately present, we would expect there to be a correlation between the number of discrepancies observed of each type. I.e. the reproduction of a paper with many problem discrepancies would be expected to have more error discrepancies than other papers. Table 4.6 shows the correlation matrix for the number of problems, assumptions, and errors observed in a reproduction. The correlation is computed using the Spearman correlation method. Reproductions ending in *No Result* are excluded since they do not report error discrepancies. There seems to be some positive correlation between the number of problems and the number of error sources. However, the number of assumptions encountered do not appear to be correlated to either the number of problems or errors.

	#Problems	*#Assumptions*	*#Errors*
#Problems	1.00	0.366	0.737
#Assumptions	0.366	1.00	0.473
#Errors	0.737	0.473	1.00

Table 4.6: Correlation matrix for number of observed discrepancies of each type in a paper. Reproductions ending in *No Result* are not included.

Discrepancies Grouped by Reproduction Type

The boxplot in FIgure 4.6 displays data about the number of discrepancies of each type observed for the different reproduction levels *R1* and *R2-D*. Based on the boxplot there does not appear to be a significant difference in the number of discrepancies observed for the two reproduction levels. The median number of discrepancies only differs by 1 for each discrepancy type, and for the total. When looking at all discrepancies, it can also be observed that the edges of the lower and upper whiskers are the same for both reproduction levels. However, the second and third quartile groups are larger for the *R1* reproductions.

Discrepancies Grouped by Outcome

The boxplot in Figure 4.7 displays data about the number of discrepancies of each type observed for the different reproduction outcome levels. Since only one reproduction attempt ended in *Success* there is no variation in the data for that category. Furthermore, since reproduction attempts ending in *No result* do not report error discrepancies, comparing the total number of discrepancies in this category to the other categories is difficult. The two remaining categories, *Partial success* and *Failure*, are the biggest categories. Looking at the data, we observe that their median values are close for all discrepancy types, and total number of discrepancies. However, the reproductions ending in *Failure* have greater variation in the total number of discrepancies, as seen by the larger quartile groups.

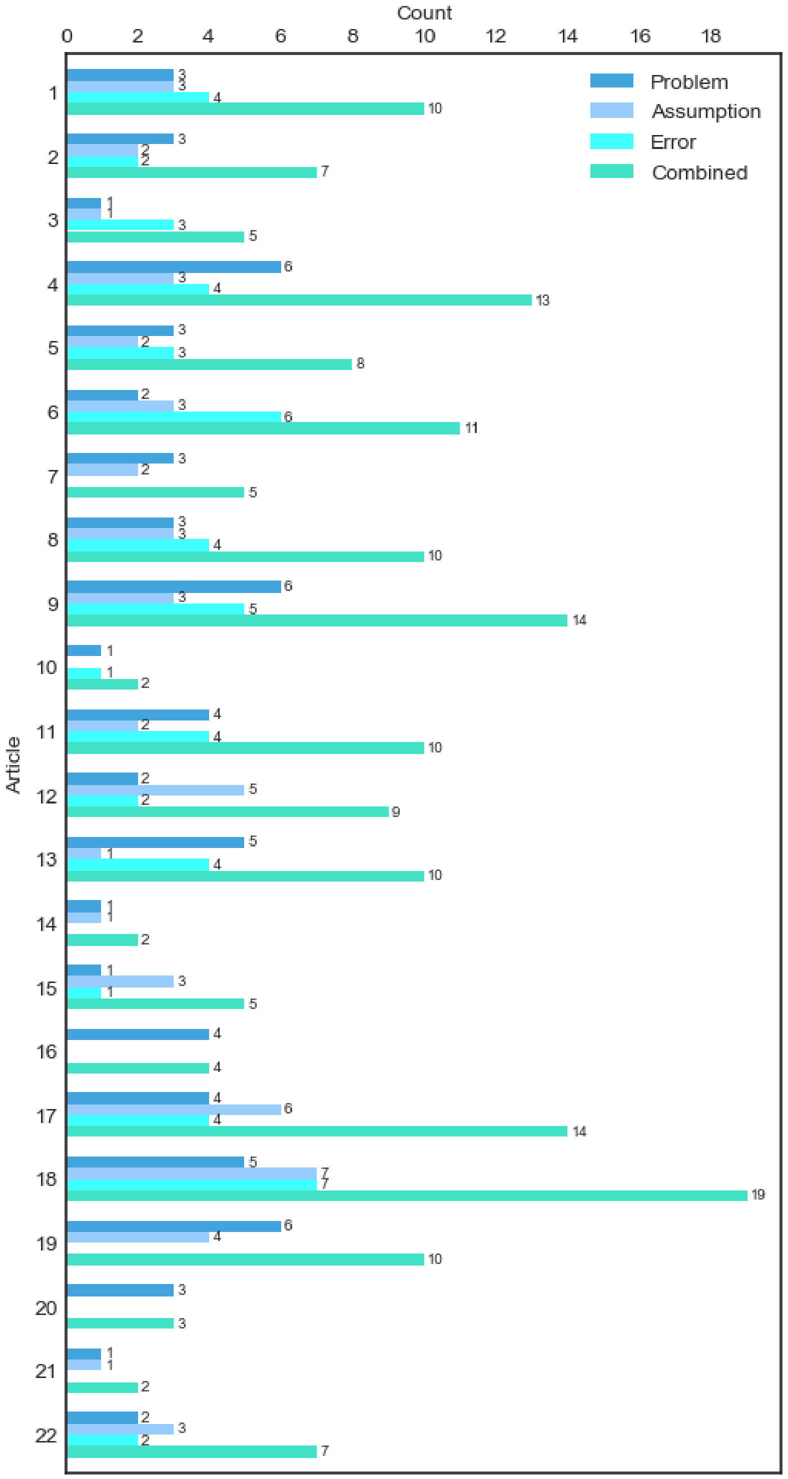

Figure 4.5: Number of discrepancies of each type encountered per paper

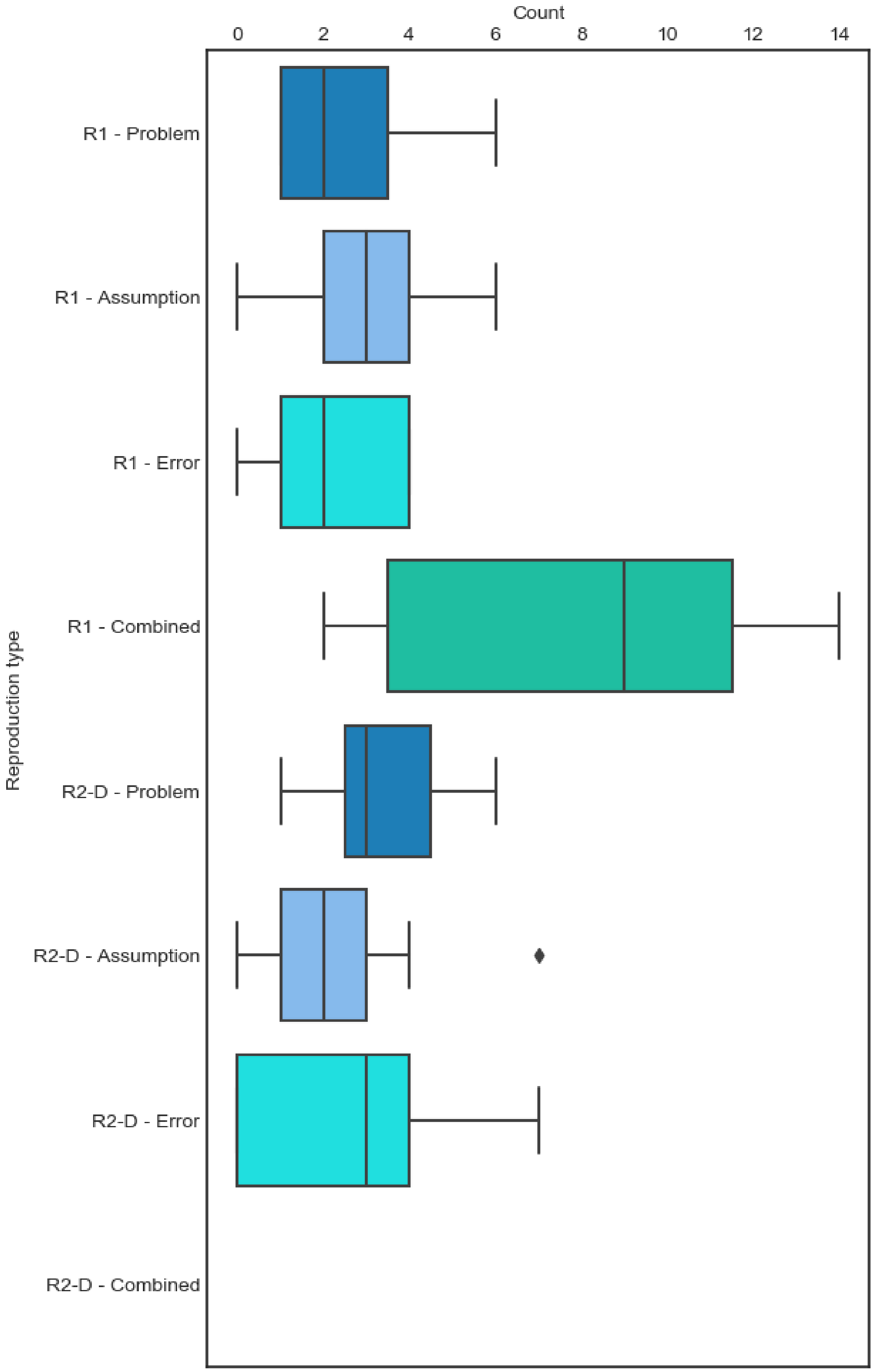

Figure 4.6: Boxplot showing statistics on the different types of discrepancies for each reproduction level: *R1* and *R2-D*

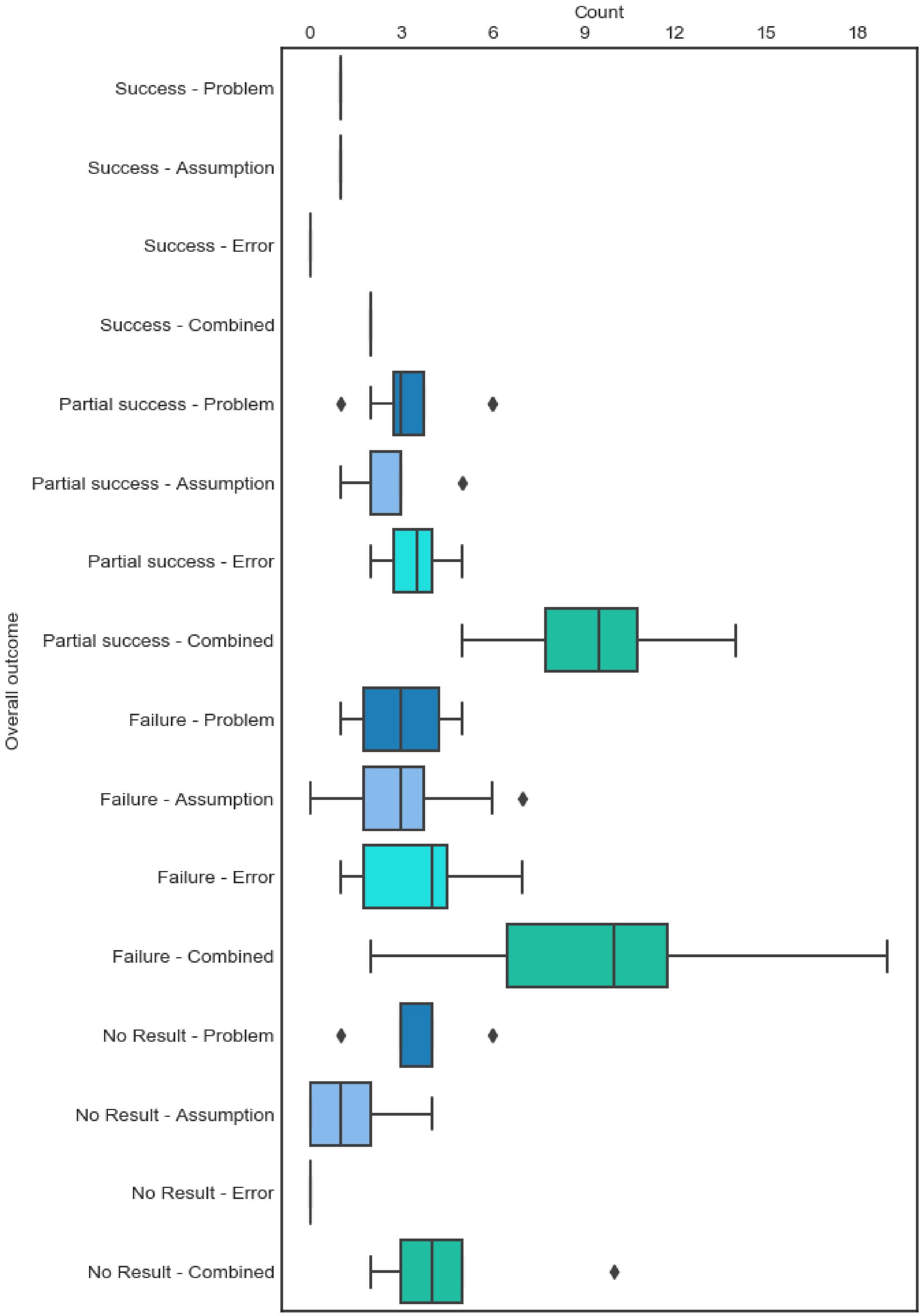

Figure 4.7: Boxplot showing statistics on the different types of discrepancies for each outcome category

Chapter 5

Discussion

This chapter discusses the results presented in Chapter 4 and the project overall. Section 5.1 provides a deeper overview and discussion of the results from the reproduction attempts. Section 5.2 discusses the reproduction procedure of Chapter 3 and how it affected the reported outcomes. Section 5.3 discusses the results for the discrepancies, and provides an evaluation of the discrepancy system. Section 5.4 presents the results achieved for the metrics of the *Article Model* of Section 3.2, and discusses how well these metrics predict the outcome of a reproduction attempt. The metrics are also compared to the score system presented in Gundersen and Kjensmo's paper [6]. Lastly, Section 5.5 discusses further work that could be done based on the results of this project.

5.1 Summary of Results

From the 30 articles chosen for possible reproduction, 7 *R1* and 15 *R2-D* reproductions were attempted. The result were 1 *Success*, 9 *Partial Successes*, 7 *Failures*, and 5 with *No Results*. These categories will be discussed in detail in the following subsections. Table 5.1 gives the percentage of reproduction attempts in each outcome category by reproduction level. More than 2/3 of the *R1* reproductions achieved either partial or full success. In comparison 1/3 of *R2-D* reproductions achieved the same results.

Outcome Category	*Percentage of total reproduction attempts*	*Percentage of R1*	*Percentage of R2*
Success	4.55%	14.29%	0.00%
Partial Success	40.91%	57.14%	33.33%
Failure	31.82%	28.57%	33.33%
No Result	22.73%	0.00%	33.33%

Table 5.1: The percentage of reproduction attempts that fell into each of the four outcome categories

5.1.1 Success

Only one of the 22 reproduction attempts ended in a successful outcome. Therefore, "Clustering by fast search and find of density peaks" [46] will be discussed in some detail. This was an *R1* reproduction attempt where the method code was linked from the article itself. The article ran the method on multiple data sets, with different measurements for the results. In total seven experiments were counted, and six were reproduced. The experiment not conducted (Figure 3B in [46]), was disregarded because comparing the results from the reproduction with the original article would be very difficult. The article presented the clusters created in the experiment by coloring the different points. This was also done for three other experiments in the article, but these contained more separated points, which made it possible to distinguish between each point. Of the six experiments conducted, two used an accuracy score, and in the rest the output from the reproduction attempts was matched with the article's cluster plot. The accuracy scores were presented with two decimal precision.

This was the only article where the reproduction obtained identical results for every experiment conducted. There was one area where this article differed from the others: The article used for the most part data sets where it achieved the optimal results. Therefore it is less ambiguity whether the reproduction achieved the same results as the article; if the output of the reproduction is optimal, then it is the same. The article also used data sets where the results where measured in percentage accuracy. In these there were two factors that contributed to achieving identical results: A deterministic method, and the precision of the accuracy. First, the method is deterministic which makes it give the same results for each run with the same hyper-parameters. Therefore, if the reproduction is run with the same data and hyper-parameters then it should achieve the same accuracy as the article. The second important factor is the precision of the accuracy. Two decimal precision were used in the article. This makes it so two outputs are seen as *identical* even if their accuracy differ in the third or later decimal. In other articles these differences leads to the results being seen as *consistent.* One of the experiment conducted in the reproduction of "XGBoost: A scalable tree boosting system" [40], achieved an accuracy of 0.8323 while the article reported 0.8304. These were similar to the second decimal, but since the article included a higher precision, the results were deemed not *identical.*

5.1.2 Partial Success

Among the four outcomes, the biggest group was the *Partial Success* containing nine reproduction attempts. Out of these, four were *R1* and five were *R2-D* reproductions. This makes the *R1* reproductions over-represented compared to the *R2-D* reproductions, considering there are seven *R1* articles and 15 *R2-D* articles. There are 40% more *R1* articles among the partial success than expected by the count.

$$\frac{\frac{4}{4+5}}{\frac{7}{7+15}} \approx 1.40$$

5.1.3 Failure

Seven of the reproduction attempts were classified as *Failures*. Out of these, two were *R1* and five were *R2-D* reproductions. The ratio of articles are consistent with the total number of *R1* and *R2-D* articles. The *Failure* category encompasses a broad range of outcomes when it comes to the distance from the goal. Some attempts are very far from the desired results, like "DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition" [47] where the classifier in the reproduction outputs the same class for all test inputs, meaning there is almost certainly a fundamental flaw in the implementation of the reproduction. Other reproductions are close to the results presented in the original article, but the results are not consistent with the conclusions in the article. The reproduction of "Classification with Noisy Labels by Importance Reweighting" [36] was deemed not *consistent* because the result differed too much from the baseline algorithms on two out of 12 tests. Furthermore, the "Context Aware Saliency Detection" [38] article used two data sets to test the performance of the method, and the reproduction ended up with a lower score on one data set and higher on the other.

5.1.4 No Result

There were five reproduction attempts that ended in *No Result*, and all of these were *R2-D* reproductions. There was not a unified cause for this among the two papers covered by the authors of this report. In "Visualizing and Understanding Convolutional Networks" [32] the running time and resources needed ended up being too significant to allow reproduction. In "Distributed representations of sentences and documents" [39] the method code was not completed within the 40 hours allocated. However, time was an important factor in both reproductions, as both used all of the time allocated. In *R2-D* reproductions it is needed to both recreate the method and the experiment, unlike *R1* reproductions where only the latter is needed. Fitting both of these into the time allocated can be difficult. This could be one of the reasons why there were only *R2-D* articles that ended up with *No Results*. However, time is not the only reason for *No Result* as there was two reproduction attempts that did not spend all of the allocated time. These were performed by Nicklas Grimstad Nilsen, and will not be discussed in this report.

5.1.5 Comparison With Other Studies

Counting reproduction attempts ending in *Success* and *Partial Success* as successes, a success rate of 45.5% was achieved. Though no comparable study on reproducibility in AI is available, the results are mostly in line with those reported by other researchers on smaller reproduction studies and in related computational fields, for example in [13], [17]. It should however be mentioned that the lack of common methodology, and in particular common criteria for success, makes comparison across studies difficult. Clearly defining what the requirements for a successful reproduction are and agreeing on a common methodology are therefore important further steps for the AI community.

Looking at the number of *R1* and *R2* reproductions, 23% of the reproductions were *R1* and 50% were *R2-D*. In other words, 73% of the papers shared enough data to be either *R1* or *R2* reproducible, which is better than the 49% reported by Gundersen and Kjensmo [6]. This might be because this project uses on average higher cited articles, which are assumed to generally be better documented.

5.2 Reproduction Procedure

This section discusses the reproduction procedure and methodology used in this project. Specifically the areas which are believed to have had the greatest impact on the results achieved, and the areas which could have been improved.

The first area which should be discussed are the possible error sources in the reproduction procedure. Most reproduction attempts were carried out by only one reproducer, and with limited time. Furthermore, AI is a broad field, and some of the studies investigated concern problems or methods previously unknown to the reproduction team. It is therefore possible that mistakes have been made in the reproductions, either as a result of the reproducers misunderstanding the documentation, or making implementation errors. Since *R2-D* reproductions generally involve more implementation than *R1* reproductions, the mistakes are more likely in *R2-D* reproductions. The time allocated per article is also an important factor in the outcome. 14 out of 22 reproductions used all of the allocated time. If available, more time could have been used to discover errors and figuring out solutions. This in turn could have helped achieve a better outcome. Based on the results, it appears that 40 hours was an unreasonably low time restriction for each reproduction attempt. However, the strict time restriction was necessary to cover a significant number of articles within the time allocated to this project. Because of these factors, it must again be stated that we make no definite claim about the reproducibility of the studies for which our reproductions ended in *Failure* or *No Result*. In order to state that a study is not reproducible, more time should be devoted to the reproduction, and several reproducers should review it.

An important factor that affects the outcome of the reproduction attempts are the definitions of outcomes used. This is divided into reproduction outcomes, and experiment outcomes. Out of the four categories of reproduction outcomes, *Partial Success* is the one with most ambiguity. As an example: Out of the eight experiments performed in the reproduction of "A modified Artificial Bee Colony algorithm for real-parameter optimization" [34], one result was *consistent* and seven were *different*. This reproduction had the same outcome category as "Deep learning-based classification of hyperspectral data" [42], where out of the six performed experiments three achieved *identical* results and three were *consistent*. The last one is a lot closer to the *Success* category, but they are both labeled as *Partial Success*. Even though there are questionable cases, the definition of *Partial Success* covers the area between *Success* and *Failure* as intended. To remove the ambiguity, more categories likely had to be added, which would make the number of reproductions in each outcome category small. The definitions for *Success*, *Failure*, and *No Result* outcome categories were clearer, and achieved their intention.

When it comes to experiment outcomes there were two considerations: What constitutes *identical*, and what constitutes *consistent*. In this project *identical* was defined as having an exact match. I.e. it is not enough to show that achieving a exact match is probable. An example of this is "XGBoost: A scalable tree boosting system" [40], where an experiment was rerun with two different epoch values and achieved higher and lower accuracy than the article. By the definition, this was classified as *consistent*. An argument could have been that the exact result could have been achieved given that a higher and a lower result was possible.

Since *identical* was defined to be exact match up to the decimal precision used in the article, articles using higher precision have a harder time being labeled *identical*. As mentioned in Section 5.1.1: Both "Clustering by fast search and find of density peaks" [46] and "XGBoost: A scalable tree boosting system" [40] had experiment results identical to their articles when rounded to two decimals. However, they were labeled differently. The reason was that they used two and four decimal precision respectively. A standard could have been set in the reproduction where the results would be rounded to two decimals in all articles. With a wider definition of *identical* the number of reproductions with *Success* as outcome would increase.

The leniency of the *consistent* category is indirectly decided by the articles themselves. The more stringent the conclusions drawn in the article, the more difficult to achieve *consistent* results. This often comes down to the choice of baseline method used as comparison. Choosing a baseline method which achieves good results makes it more difficult to achieve a *consistent* reproduction. Two articles that has experiments demonstrating the effect of the baseline method used is "Measuring the Objectness of Image Windows" [26] and "Classification with Noisy Labels by Importance Reweighting" [36]. In the first, the baseline method achieves between 12 and 27 percentage points lower accuracy than the method presented in the article. In the second the baseline method is within 0.3 percentage points from the method presented.

Choosing a baseline method which is not strictly worse than the proposed method also makes it difficult to achieve *consistent* results. In one of the experiments from "Context Aware Saliency Detection" [38] the baseline method was worse and better than the proposed method on different points. The reproduction achieved the same or better on every point compared to the proposed method. This was not categorized as *identical*, because it achieved a different accuracy than the original article. However, the outcome was also not classified as *consistent*. The proposed method was not strictly better than the baseline method, but the reproduction was. Therefore the experiment was classified as *different*.

A way to avoid the issues related to definitions of outcomes would be to not classify them, but use the error rate for each experiment to analyze the results. However, there are multiple challenges with this approach. Some articles like "Generalized Correntropy for Robust Adaptive Filtering" [27] use graphs to display their results. Therefore, a quantitative measure would be difficult to calculate. Even among the experiments with numerical results, this choice of metric could prove difficult. In articles dealing with function optimization, like "Development and investigation of efficient artificial bee colony algorithm for numerical function optimization" [28], the values reproduced can be of a huge negative magnitude. In the paper, the experiment output range from 0 to a couple of thousand, and the highest negative order of magnitude was -197. The error rate of these type of experiments could be in different orders of magnitude compared to experiments where the result is presented as accuracy.

Another alternative is to use statistical tests to determine the significance level of difference in results for experiments. This is most viable for experiments where the reported metric is the average of multiple runs, and where the standard deviation is also reported. In such instances it is possible to use e.g. a Student's t-test to test for significant difference in the average result achieved by original study and reproduction. If the difference is statistically different, the reproduction of an experiment can be classified as a failure, and if not it can be considered a success. Studies in the area of function optimization, such as "Cooperatively Coevolving Particle Swarms for Large Scale Optimization" [30], do in many cases report the average minimum value found across several runs along with the standard deviation, enabling the use of statistical tests for comparing results. However, must studies investigated in this project do not report results over multiple runs.

An area of the reproduction procedure which could have been improved was the documentation practices during reproductions. With the exception of the Google Forms form for registering data related to the *Article Model Metrics*, no unified system for taking notes was used during the reproductions, and all reproducers used different systems. Having a better unified system would have resulted in less work after the reproduction attempts, when the documentation had to be combined for registering e.g. discrepancies.

Lastly, it should be noted that the selection of papers as presented in Section 3.4 was not performed in a randomized way. Ideally, when attempting to understand a phenomenon in a large population, in this case reproducibility in AI studies, the samples selected for use in the experiment should be randomly selected. This increases the likelihood that the results observed in the experiment generalizes to the population as a whole. The main reason why randomized selection was not used was because we wished to understand how the situation was at the top of AI research. The underlying assumption here is that the number of citations achieved by a study is a valid measure of its quality.

5.3 Discrepancies

The system of discrepancies was proposed to quantify the number of issues encountered during reproduction. It is intended to provide a quantitative measure of which issues most commonly affect reproductions, as well as an estimate for how difficult a reproduction attempt was, and how confident the reproducer is about the results. This section discusses the results presented in Section 4.2, and evaluates the value of the discrepancy system for reproducibility.

Figures 4.1 to 4.3 give some interesting insights about the prevalence of the different discrepancy categories. When investigating which discrepancies are the most common, it is observed that the most common problem category (*P10*) and the most common error categories (*E5* and *E6*), all concern issues relating to the implementation of a method or experiment. The most common assumption category (*A6*) is also related to this issue, though it is more directly tied to the method. When performing the implementation of a method or experiment, questions often arise which are not obvious from the description of the method or experiment. Furthermore, since the goal of the paper usually is to present the method, comparatively little space is often used for explaining the implementation. As an example, 19 of the 30 papers covered in this project failed to even mention the programming language used in the implementation. Based on the number of discrepancies arising from poor implementation description, it appears that this is an area of documentation practices where additional focus should be placed.

The majority of reproduction attempts performed have been *R2-D* reproductions. However, some observations are also made specifically about *R1* studies. It is observed from Figure 4.1 that every *R1* study except one has a discrepancy of type *P1*, i.e. even when some method code is shared, experiment code is mostly not shared, at least not for all experiments. This indicates that even when researchers have taken steps to share their code, their effort can and should be improved to increase reproducibility.

Based on the results in Table 4.6 it is unclear whether there is any definite correlation between the number of problems, assumptions, and error discrepancies encountered in a reproduction attempt. The correlation coefficient between the number of problem and error discrepancies of 0.737 is quite strongly positive, but the correlation coefficients between assumption discrepancies and problems and error discrepancies are only moderately positive. As stated in Section 4.2.4, it was assumed that there would be a significant positive correlation between the number of discrepancies. This was based on the intuition that problems create the need for assumptions, and that both problems and assumptions create possible error sources. However, based on the observed results and the limited data set of 17 reproduction attempts, we cannot confirm this intuition.

There is no clear pattern in the distribution of discrepancies between papers either. As mentioned in the discussion of Figure 4.7 there does not appear to be any significant difference in the number of discrepancies between reproduction attempts reporting *Partial* *success* and *Failure*. With the assumption that more discrepancies increase the difficulty of reproduction attempts, we would have expected reproduction attempts ending *Partial success* to have fewer discrepancies than those with *Failure*. However, based on the results currently available, the number of discrepancies observed during a reproduction attempt does not appear to be indicative of the outcome of the reproductive attempt. The number of discrepancies does not appear to be significantly different for *R1* and *R2-D* reproductions either, as shown in Figure 4.6.

The current system of discrepancies cover most of the major issues encountered during our reproduction attempts. Furthermore, it is believed to be rather comprehensive with respect to possible issues that can be encountered during reproductions. However, some important limitations have already been encountered. These limitations might partly explain why the number of discrepancies is a poor indicator of reproduction outcome, and why the system might only partially achieve its goal of quantitatively describing how a reproduction attempt diverges from an original study.

Firstly, the current system and method for counting discrepancies does not account for multiple instances of a discrepancy. In some reproduction attempts a discrepancy will occur several times, and the severity of the issue will be greater than for reproduction attempts where the discrepancy only occurs once. As an example, during the reproduction of "Learning Sparse Representations for Human Action Recognition" [31] three different third-party libraries were used, resulting in three assumptions of category *A8*. During the reproduction of "Classification with Noisy Labels by Importance Reweighting" [36] only one third-party library was used, producing only one assumption of category *A8*. Despite this, the current system counts both papers as having one discrepancy of category *A8*.

A second issue is that the current system for counting discrepancies does not provide a way of quantifying the magnitude of the discrepancy. Several of the discrepancies identified in this project will have significantly differing consequences in different reproduction attempts. In some cases, a study may have many discrepancies, but still be partially reproducible if the discrepancies are manageable or do not affect all experiments. On the other hand, a study might have only one discrepancy, but it might be so severe as to make any reproduction attempt extremely difficult. For "Distributed representations of sentences and documents" [39], the only reported discrepancy was of category *P10*. This problem category was also reported for eleven other papers. However, the proposed method in the paper was so complex and the implementation so difficult to understand that the single discrepancy *P10* resulted in the reproduction attempt resulting in *No result*. This can be compared with "iSuc-PseOpt: Identifying lysine succinylation sites in proteins by incorporating sequence-coupling effects into pseudo components and optimizing imbalanced training dataset" [33], which was also believed to be *R2-D* reproducible, but where three problem discrepancies, including *P10*, were encountered. In this instance, all experiments could be reproduced, and the reproduced results were found to be consistent with the original results.

The discrepancies identified in this project are not necessarily equal in their expected severity either. For example, assumptions of category *A9*, assuming the version of a third-party library used does not affect its output, will in most cases be easier to justify than assumptions from category *A6*, which concerns assumptions about how to treat an aspect of a method which do not have support in similar papers.

Lastly, a combination of discrepancies may give rise to difficulties greater than the sum of each individual discrepancy. For "Blind Image Quality Assessment: A Natural Scene Statistics Approach in the DCT Domain" [29], only 10% of experiments could be conducted, even though 15 hours remained of the 40 hour limit. This was the result of a combination of problem discrepancies encountered where the necessary weights for most of the experiments could not be recovered within the limits of an *R1* reproduction attempt. Specifically, the method code which was shared did not include the procedure for training weights (problem category *P2*), and even though some weights were shared, they were not the same weights as the ones used in the original experiments (problem category *P12*). If either of these problems had been resolved, most of the experiments could have been conducted by a reproducer. The combination of them resulted in only 1 of 10 experiments being reproducible.

As a result of the above four issues, reporting the number of discrepancies encountered during a reproduction attempt may give an inaccurate picture of how difficult the study is to reproduce. The system do still partly provide a quantitative measure of which issues are most common, and might be used to identify where initiatives for better documentation policies should be focused. In this sense, the results presented in Figures 4.1, 4.2, and 4.3 may be more informative than those of Figure 4.5. If the system is ensured to be comprehensive, i.e. the discrepancies cover all possible issues, it may still be useful for identifying the most pressing documentation issues, and may form the basis for a checklist that can be used by researchers to ensure that their science is easily reproducible.

5.4 Article Model

Predicting on unseen data is a good way to evaluate a model. During the preliminary work for the reproduction attempts, the *Article Model Metric*, see Section 3.2, was developed. The goal of this metric was to quantify the level of documentation in an article, with the hypothesis that an article with a high score was more likely to be reproduced successfully. The metric consists of a set of scores for different components of an article, called *Component Metrics*, which can be used to highlight possible problem areas for reproduction. This sections evaluates the *Article Model Metric* with the reproduction attempts conducted during the project. The *Article Model Metric* and *Component Metrics* for each article can be found in Table A.3 in Appendix A. The *Article Model Metric* was influenced by the metric proposed by Gundersen and Kjensmo [6], and their metric is also evaluated. The overall score from both metrics are shown in Table 5.2, where the average is calculated per result category.

The metrics does not match well with the outcome categories. The predicted ordering of the

Outcome	*Article Model Metric*	*Gundersen and Kjensmo [6] Metric*
Success	0.44	0.23
Partial Success	0.47	0.32
Failure	0.50	0.30
No Result	0.41	0.37

Table 5.2: The average documentation level per outcome category, using the *Article Model Metric* and Gundersen and Kjensmo [6] Metric

categories by declining metric should be: *Success*, *Partial Success*, *Failure*, and *No Result.* Neither of the metrics match this, and the categories with highest scores are respectively *Failure* for the *Article Model Metric* and *No Results* for Gundersen's and Kjensmo's metric. Performing a Speraman's Rank Correlation test on the correlation between the *Article Model Metric* score of a paper and the outcome of the reproduction produces a correlation value of 0.02, indicating almost no correlation between the two variables. These results makes it difficult to use the metrics as a way to predict the outcome of a reproduction attempt. There are however multiple possible factors contributing to this poor result.

As mentioned in Section 5.1.3 the *Failure* category can contain articles where the results from the reproduction had a lower error rate than some articles in the *Partial Success* category, but because of stricter conclusions were not consistent with the article's claims. This can be a contributing factor to these two categories having similar score in the *Article Model Metric.*

The selection of articles is small, and there is only one article in the *Success* category. This make the score for the category an unreliable representation of the performance of the metrics. The small sample size is not limited to the *Success* category, but it stands out from the rest with four less articles than the second smallest category: *No Result.* When excluding the *Success* category, the metric created by Gundersen and Kjensmo ends up having *Failure* as the lowest scoring category instead of *Success.*

If both of these factors are taken into consideration, in an advantageous way for the metrics, so *Success* is excluded and *Partial Success* and *Failure* are seen as equal, the usability of the metrics are still questionable. The usability of Gundersen's and Kjensmo's metric does not change, while the *Article Model Metric* seem to be able to separate articles where reproduction attempts will give results or not.

Another factor that could affect the *Article Model Metric* is the importance given to each *Component Metric.* These 11 sub-metrics are weighted equally originally. The average score for each of these can be seen in Figure 5.1. The case might be that some components are more important than others, and changing the weight given to that *Component Metric* could give an overall more consistent metric. When looking at the figure, one can observe that the *Success* category outperforms the others on *Method*, *Implementation*, and *Partitioned Data.* These might be some of the most important components in the model, but it is difficult to draw any conclusions with the limited sample size. There are few components where the

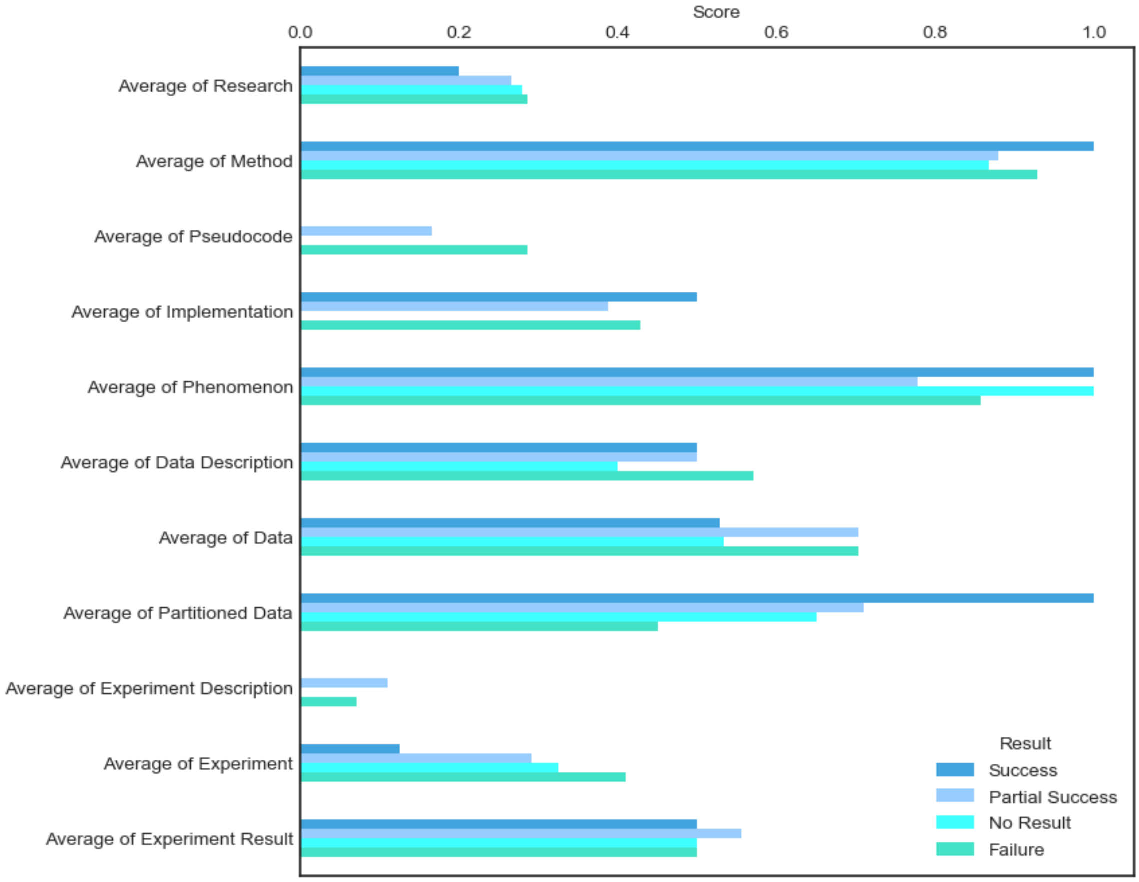

Figure 5.1: Bar plot of the score of all *Component Metrics* by type of outcome

division of *Partial Success* and *Failure* is clearly visible. In terms of the *No Result* category some components stands out, by having a score of zero: *Pseudocode*, *Implementation*, and *Experiment Description*. Out of these the *Implementation* is the only one where all the others get a nonzero score. Again it is difficult to draw conclusions given the sample size.

Neither the *Article Model Metric* or Gundersen's and Kjensmo's metric allow for complex relationships between components, which might limit their predicting capabilities. By complex, it is specifically the interactions between different components that is being highlighted here. E.g. deficiencies in some components by an article can be made up by containing certain other component parts. This is relevant for most articles because few, none in this project, will score maximal on all parts of the documentation. Therefore, most articles will be deficient in some aspects according to the metrics. Then it is important to be able to separate between how big these impacts will be. The complex interaction between the parts makes this task even more difficult. This will not be studied in this project, but more on this can be read in Section 5.5.

5.5 Further Work

This project has studied the current state of reproducibility in AI. A model for understanding empirical studies in AI have been proposed, along with metrics for estimating the reproducibility of a study. The discrepancy system has also been proposed for categorizing issues encountered during reproduction. However, further effort is required to enhance these tools and improve the AI community's understanding of reproducibility.

Constructing a metric for the documentation level in an article which corresponds more closely to the outcome of a potential reproduction would be a valuable tool. If the metric can make the complex interaction between the different components comprehensible, then it could be used as part of the writing process of an article. When writing an article it might be of interest to restrict the length or limit technical detail. In those circumstances a metric like this could help in selecting the most important elements to include in the paper.

Two avenues for improving the *Article Model Metric* will be suggested in this section. In both approaches it would be vital to have a larger sample of reproduction attempts, and keep some attempts as a test to evaluate the performance of the metric. One way of improving the *Article Model Metric*, would be to take into account the effect each component has on the reproduction. This could be done by assigning weights to the *Component Metrics*, so more important components have a bigger impact. Furthermore, this could also be done on the sub-component level. This would take into account the effect of each separate sub-component on the reproduction. To calculate the proper weights, the average metric per outcome category could be calculated, similar to Figure 5.1. These could be used to adjust the weights. This could also be done with sub-components. Then the sub-components that were more prevalent in *Success* and *Partial Success* could get a higher weight, while others would get a lower weight.

Another way of improving the *Article Model Metric* could be to find interactions between sub-components. As an example: The sub-components A and B are often both scored highly in the *Success* category, but when only one of them is scored high they are almost never found in that category. These could be relevant to give a document score that reflects the outcomes to a bigger extent. To discover patterns like these, or larger groupings, correlation analysis could be used.

The system of discrepancies for categorizing issues in reproductions has some significant limitations, and the number of discrepancies reported for a reproduction appears to have little correlation to the outcome of the reproduction. The categories might need adjustments and additions, but if ensured to be comprehensive and adopted for continued use in reproduction attempts, it might be a valuable tool for identifying the most pressing issues in reproducibility in AI. Furthermore, it might also be used by researchers as a guide for ensuring that their research is not affected by common issues.

The lack of a clear methodology for performing a reproduction attempt, and the lack of consensus on what constitutes an successful reproduction are problems which make the study of reproducibility difficult. To better understand the current state of reproducibility, and encourage further efforts in reproducibility, these issues should be further considered.

Lastly, the overarching goal of this project and its related efforts are to improve the reproducibility of AI research. The reproduction attempts carried out in this project have shown that the current state of reproducibility is problematic, and that reproducing studies is more difficult than it should be. The lack of clear documentation and resource publishing guidelines means that reproductions still are difficult, and that some studies are impossible to truly reproduce. Establishing better guidelines and encouraging better documentation practices are therefore still the most important steps that need to be taken by the AI community to improve reproducibility.

Chapter 6

Conclusion

The ability to reproduce published results is a cornerstone of the scientific method. The **research goal** of this project was to provide a quantitative overview of the state of reproducibility in AI. We believe the results from the review of 30 articles, and reproduction attempt of 22, provide interesting insights into the situation. As expected, the situation was found to be less than ideal.

Three hypotheses were proposed for the project, along with three associated predictions. The **first prediction** was that the majority of studies covered in the project would not be reproducible within the limitations of the reproduction procedure. I.e. within 40 hours and with the available resources. Out of the 30 articles covered, only 10 were reproduced with *Success* or *Partial Success.* The majority of the articles were not found to be reproducible, and the **first hypothesis**, that reproduction is difficult, is therefore considered to be supported. Though a disclaimer should be made about the possible error sources of the project as discussed in Section 5.2.

The **second prediction** of the project was that it was possible to group the issues encountered during reproductions into general categories. This prediction is less precise than the others, and therefore more difficult to test. However, an attempt was made using the discrepancy system, which was used to group issues. When investigating which issues had affected which papers, it was found that 12 out of the 22 articles attempted reproduced had had at least one issue of problem category *P10.* We interpret this and the other results from the discrepancy system as indicating that there is significant similarity in issues encountered between reproduction attempts. We therefore consider the **second hypothesis**, that the issues which make reproduction difficult are the same across studies, to be largely supported.

The **third prediction** was that there was a significant correlation between the documentation level measured for an article, and the observed outcome of the associated reproduction attempt. Quantifying documentation levels is a difficult task, and was in this project performed using the proposed *Article Model Metric.* When computing the average metric values for the different outcome levels the results were not higher for better reproduction outcome levels. The correlation test also indicated no correlation between the metric score of an article and the outcome of the reproduction. However, the results showed that *R1* studies, i.e. studies providing at least some code, were more likely to be successfully reproduced than *R2* studies. Though not conclusive, this indicates that documentation level do have some relation to reproduction outcome. More documentation is almost always helpful for reproductions, and increased sharing of data and code is positive. The results achieved here more likely indicate that the *Article Model Metric* is unable to correctly estimate the documentation levels of articles. We therefore do not believed that the **third hypothesis**, that the level of documentation and the ease of reproduction is related, should be rejected, though our results can not corroborate it.

In addition to the results of the reproduction attempts, which provide statistics on reproducibility given the scope of this project, the project has two additional **contributions**. The first is the *Article Model* which is intended as a framework for understanding all components of a research article in AI. The associated metrics are intended to be used as a system for quantifying the documentation level of a paper, which was hypothesized to predict the reproducibility of the paper. However, within the parameters of this study, the metric did not achieve its goal. The second tool developed in the project is the discrepancy system for categorizing and counting issues encountered during reproductions. This tool can be used for identifying which issues commonly limit reproducibility.

This project, along with earlier studies, show that there are significant difficulties in reproducing research within AI. Similar issues often prevent studies from being reproduced, and this provides an indication of were better documentation practices are most needed. We believe further work is needed in the area of reproducibility, most importantly in the construction of better documentation guidelines and policies.

Bibliography

[1] M. Baker, “1,500 scientists lift the lid on reproducibility”, *Nature*, vol. 533, pp. 452–454, 2016.

[2] H. Pashler and E. Wagenmakers, “Editors’ introduction to the special section on replicability in psychological science: A crisis of confidence?”, *Perspectives on Psychological Science*, vol. 7, no. 6, pp. 528–530, 2012, PMID: 26168108. eprint: `https://doi.org/10.1177/1745691612465253`.

[3] M. Hutson, “Artificial intelligence faces reproducibility crisis”, *Science*, vol. 359, no. 6377, pp. 725–726, 2018. eprint: `http://science.sciencemag.org/content/359/6377/725.full.pdf`.

[4] D. L. Donoho, A. Maleki, I. U. Rahman, M. Shahram, and V. Stodden, “Reproducible research in computational harmonic analysis”, *Computing in Science Engineering*, vol. 11, no. 1, pp. 8–18, 2009-01.

[5] V. C. Stodden, “Trust your science? open your data and code”, *Amstat News*, vol. 409, 2011.

[6] O. E. Gundersen and S. Kjensmo, “State of the art: Reproducibility in artificial intelligence”, *AAAI*, 2017.

[7] P. R. Cohen, *Empirical Methods for Artificial Intelligence*. Cambridge, MA, USA: MIT Press, 1995.

[8] K. Bollen, J. T. Cacioppo, R. Kaplan, J. Krosnick, and J. L. Olds, *Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science*. National Science Foundation, Arlington, VA, 2015.

[9] C. Drummond, “Replicability is not reproducibility: Nor is it good science”, 2009-01.

[10] J. Vitek and T. Kalibera, “Repeatability, reproducibility, and rigor in systems research”, in *Proceedings of the Ninth ACM International Conference on Embedded Software*, ser. EMSOFT ’11, Taipei, Taiwan: ACM, 2011, pp. 33–38.

[11] N. P. Rougier, K. Hinsen, F. Alexandre, T. Arildsen, L. A. Barba, F. C. Benureau, C. T. Brown, P. De Buyl, O. Caglayan, A. P. Davison, *et al.*, “Sustainable computational science: The rescience initiative”, *PeerJ Computer Science*, vol. 3, e142, 2017.

[12] S. N. Goodman, D. Fanelli, and J. P. A. Ioannidis, "What does research reproducibility mean?", *Science Translational Medicine*, vol. 8, no. 341, 341ps12–341ps12, 2016. eprint: `http://stm.sciencemag.org/content/8/341/341ps12.full.pdf`.

[13] T. Mende, "Replication of defect prediction studies: Problems, pitfalls and recommendations", in *Proceedings of the 6th International Conference on Predictive Models in Software Engineering*, ser. PROMISE '10, Timioara, Romania: ACM, 2010, 5:1–5:10.

[14] F. Fokkens, M. Van Erp, M. Postma, T. Pedersen, P. Vossen, and N. Freire, "Offspring from reproduction problems: What replication failure teaches us", in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. Association of Computational Linguistics, 2013, pp. 1691–1701.

[15] M. Topalidou, A. Leblois, T. Boraud, and N. P. Rougier, "A long journey into reproducible computational neuroscience", *Frontiers in Computational Neuroscience*, vol. 9, p. 30, 2015.

[16] J. Vitay, "[Re] Robust timing and motor patterns by taming chaos in recurrent neural networks", *ReScience*, vol. 2, no. 1, 2016-10.

[17] T. Manninen, R. Havela, and M.-L. Linne, "Reproducibility and comparability of computational models for astrocyte calcium excitability", *Frontiers in Neuroinformatics*, vol. 11, p. 11, 2017.

[18] R. D. Peng, "Reproducible research in computational science", *Science*, vol. 334, no. 6060, pp. 1226–1227, 2011. eprint: `http://science.sciencemag.org/content/334/6060/1226.full.pdf`.

[19] D. C. Ince, L. Hatton, and J. Graham-Cumming, "The case for open computer programs", *Nature*, vol. 482, 2012.

[20] J. Vanschoren, J. N. van Rijn, B. Bischl, and L. Torgo, "Openml: Networked science in machine learning", *SIGKDD Explorations*, vol. 15, no. 2, pp. 49–60, 2013.

[21] ReScience. (). Reproducible science is good. replicated science is better., [Online]. Available: `http://rescience.github.io` (visited on 2018-05-30).

[22] V. Stodden, P. Guo, and Z. Ma, "Toward reproducible computational research: An empirical analysis of data and code policy adoption by journals", *PLOS ONE*, vol. 8, 2013.

[23] V. Stodden, "The scientific method in practice: Reproducibility in the computational sciences", MIT Sloan, Tech. Rep. 4773-10, 2010-02.

[24] R. Mayer and A. Rauber, "A quantitative study on the re-executability of publicly shared scientific workflows", in *2015 IEEE 11th International Conference on e-Science*, 2015-08, pp. 312–321.

[25] C. Collberg and T. A. Proebsting, "Repeatability in computer systems research", *Commun. ACM*, vol. 59, no. 3, pp. 62–69, 2016-02.

[26] B. Alexe, T. Deselaers, and V. Ferrari, "Measuring the objectness of image windows", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2189–2202, 2012-11.

[27] B. Chen, L. Xing, H. Zhao, N. Zheng, J. C. Prı, *et al.*, "Generalized correntropy for robust adaptive filtering", *IEEE Transactions on Signal Processing*, vol. 64, no. 13, pp. 3376–3387, 2016.

[28] G. Li, P. Niu, and X. Xiao, "Development and investigation of efficient artificial bee colony algorithm for numerical function optimization", *Applied soft computing*, vol. 12, no. 1, pp. 320–332, 2012.

[29] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind image quality assessment: A natural scene statistics approach in the dct domain", *IEEE Transactions on Image Processing*, vol. 21, no. 8, pp. 3339–3352, 2012-08.

[30] X. Li and X. Yao, "Cooperatively coevolving particle swarms for large scale optimization", *IEEE Transactions on Evolutionary Computation*, vol. 16, no. 2, pp. 210–224, 2012-04.

[31] T. Guha and R. K. Ward, "Learning sparse representations for human action recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 8, pp. 1576–1588, 2012-08.

[32] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks", in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., Cham: Springer International Publishing, 2014, pp. 818–833.

[33] J. Jia, Z. Liu, X. Xiao, B. Liu, and K.-C. Chou, "Isuc-pseopt: Identifying lysine succinylation sites in proteins by incorporating sequence-coupling effects into pseudo components and optimizing imbalanced training dataset", *Analytical Biochemistry*, vol. 497, pp. 48–56, 2016.

[34] B. Akay and D. Karaboga, "A modified artificial bee colony algorithm for real-parameter optimization", *Information Sciences*, vol. 192, pp. 120–142, 2012.

[35] Y. Peng, A. Ganesh, J. Wright, W. Xu, and Y. Ma, "Rasl: Robust alignment by sparse and low-rank decomposition for linearly correlated images", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2233–2246, 2012.

[36] T. Liu and D. Tao, "Classification with noisy labels by importance reweighting", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 3, pp. 447–461, 2016-03.

[37] F. J. Ordóñez and D. Roggen, "Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition", *Sensors*, vol. 16, no. 1, 2016.

[38] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 10, pp. 1915–1926, 2012.

[39] Q. Le and T. Mikolov, "Distributed representations of sentences and documents", in *International Conference on Machine Learning*, 2014, pp. 1188–1196.

[40] T. Chen and C. 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.

[41] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, “Facial landmark detection by deep multi-task learning”, in *European Conference on Computer Vision*, Springer, 2014, pp. 94–108.

[42] Y. Chen, Z. Lin, X. Zhao, G. Wang, and Y. Gu, “Deep learning-based classification of hyperspectral data”, *IEEE Journal of Selected topics in applied earth observations and remote sensing*, vol. 7, no. 6, pp. 2094–2107, 2014.

[43] G. Huang, S. Song, J. N. Gupta, and C. Wu, “Semi-supervised and unsupervised extreme learning machines”, *IEEE transactions on cybernetics*, vol. 44, no. 12, pp. 2405–2417, 2014.

[44] W. Li, R. Zhao, T. Xiao, and X. Wang, “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.

[45] F. Jia, Y. Lei, J. Lin, X. Zhou, and N. Lu, “Deep neural networks: A promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data”, *Mechanical Systems and Signal Processing*, vol. 72, pp. 303–315, 2016.

[46] A. Rodriguez and A. Laio, “Clustering by fast search and find of density peaks”, *Science*, vol. 344, no. 6191, pp. 1492–1496, 2014.

[47] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, “Decaf: A deep convolutional activation feature for generic visual recognition”, in *International conference on machine learning*, 2014, pp. 647–655.

[48] K. Zhang, X. Gao, D. Tao, and X. Li, “Single image super-resolution with non-local means and steering kernel regression”, *IEEE Transactions on Image Processing*, vol. 21, no. 11, pp. 4544–4556, 2012.

[49] D. Zhang, D. Shen, A. D. N. Initiative, *et al.*, “Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in alzheimer’s disease”, *NeuroImage*, vol. 59, no. 2, pp. 895–907, 2012.

[50] X.-C. Yin, X. Yin, K. Huang, and H.-W. Hao, “Robust text detection in natural scene images”, *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 5, pp. 970–983, 2014.

[51] A. Graves and N. Jaitly, “Towards end-to-end speech recognition with recurrent neural networks”, in *International Conference on Machine Learning*, 2014, pp. 1764–1772.

[52] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, *et al.*, “Mastering the game of go with deep neural networks and tree search”, *nature*, vol. 529, no. 7587, pp. 484–489, 2016.

[53] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning", *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1285–1298, 2016.

[54] X. Meng, J. Bradley, B. Yavuz, E. Sparks, S. Venkataraman, D. Liu, J. Freeman, D. Tsai, M. Amde, S. Owen, *et al.*, "Mllib: Machine learning in apache spark", *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 1235–1241, 2016.

[55] G. Cheng, P. Zhou, and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in vhr optical remote sensing images", *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 12, pp. 7405–7415, 2016.

[56] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes", in *The Tenth IEEE International Conference on Computer Vision (ICCV'05)*, Beiging, 2005, pp. 1395–1402.

[57] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding", *arXiv preprint arXiv:1408.5093*, 2014.

Appendix A

Article Information

Id	*Title*	*Person responsible*
1	Measuring the Objectness of Image Windows [26]	Odd Cappelen
2	Generalized Correntropy for Robust Adaptive Filtering [27]	Martin Mølnå
3	Development and investigation of efficient artificial bee colony algorithm for numerical function optimization [28]	Martin Mølnå
4	Blind Image Quality Assessment: A Natural Scene Statistics Approach in the DCT Domain [29]	Odd Cappelen
5	Cooperatively Coevolving Particle Swarms for Large Scale Optimization [30]	Odd Cappelen
6	Learning Sparse Representations for Human Action Recognition [31]	Odd Cappelen
7	Visualizing and Understanding Convolutional Networks [32]	Odd Cappelen
8	iSuc-PseOpt: Identifying lysine succinylation sites in proteins by incorporating sequence-coupling effects into pseudo components and optimizing imbalanced training dataset [33]	Odd Cappelen
9	A modified Artificial Bee Colony algorithm for real-parameter optimization [34]	Nicklas Grimstad Nilsen
10	RASL: Robust alignment by sparse and low-rank decomposition for linearly correlated images [35]	Nicklas Grimstad Nilsen
11	Classification with Noisy Labels by Importance Reweighting [36]	Odd Cappelen
12	Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition [37]	Odd Cappelen
13	Context Aware Saliency Detection [38]	Martin Mølnå
14	Distributed representations of sentences and documents [39]	Martin Mølnå
15	XGBoost: A scalable tree boosting system [40]	Martin Mølnå
16	Facial landmark detection by deep multi-task learning [41]	Nicklas Grimstad Nilsen
17	Deep learning-based classification of hyperspectral data [42]	Nicklas Grimstad Nilsen
18	Semi-supervised and unsupervised extreme learning machines [43]	Nicklas Grimstad Nilsen
19	DeepReID: Deep Filter Pairing Neural Network for Person Re-Identification [44]	Nicklas Grimstad Nilsen
20	Deep neural networks: A promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data [45]	Nicklas Grimstad Nilsen
21	Clustering by fast search and find of density peaks [46]	Martin Mølnå
22	DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition [47]	Martin Mølnå
23	Single image super-resolution with non-local means and steering kernel regression [48]	Odd Cappelen
24	Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease [49]	Odd Cappelen
25	Robust text detection in natural scene images [50]	Nicklas Grimstad Nilsen
26	Towards end-to-end speech recognition with recurrent neural networks [51]	Martin Mølnå
27	Mastering the game of Go with deep neural networks and tree search [52]	Martin Mølnå
28	Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning [53]	Odd Cappelen
29	MLlib: Machine learning in Apache Spark [54]	Odd Cappelen
30	Learning Rotation-Invariant Convolutional Neural Networks for Object Detection in VHR Optical Remote Sensing Images [55]	Martin Mølnå

Table A.1: Information about which person was responsible for each reproduction attempt

ID	Experiments in paper	Identical	Consistent	Different	Not Conducted
1	18	0	4	0	14
2	4	1	0	1	2
3	46	17	17	12	0
4	10	0	1	0	9
5	21	0	4	3	14
6	14	0	0	1	13
7	20	0	0	0	20
8	1	0	1	0	0
9	68	0	1	7	60
10	31	0	0	24	7
11	6	0	0	1	5
12	4	0	1	1	2
13	2	0	0	2	0
14	3	0	0	0	3
15	4	0	2	0	2
16	12	0	0	0	12
17	16	3	3	0	10
18	38	0	0	5	33
19	6	0	0	0	6
20	10	0	0	0	10
21	7	6	0	0	1
22	4	0	0	1	3

Table A.2: The total number of experiments per article and the number of each experiment outcome

ID	Research	Method	Pseudocode	Implementation	Phenomenon	Data Description	Data	Partitioned Data	Experiment Description	Experiment	Experiment Result	*Article Model Metric*
1	0.20	1.00	0.00	0.17	1.00	1.00	0.50	0.25	0.00	0.13	0.50	0.43
2	0.20	0.75	0.00	0.00	1.00	0.00	-	-	0.00	0.13	0.50	0.29
3	0.40	1.00	0.50	0.00	1.00	0.00	1.00	-	0.00	0.25	0.50	0.47
4	0.20	1.00	0.00	0.67	1.00	1.00	1.00	0.80	0.00	0.13	0.50	0.57
5	0.00	1.00	0.50	0.00	0.00	0.50	0.50	-	0.50	0.25	0.50	0.38
6	0.60	1.00	0.50	0.00	1.00	1.00	0.50	0.60	0.00	0.25	0.50	0.54
7	0.20	1.00	0.00	0.00	1.00	1.00	0.50	0.80	0.00	0.50	0.50	0.50
8	0.20	0.50	0.00	0.50	1.00	1.00	0.50	0.50	0.00	0.13	0.50	0.44
9	0.20	1.00	0.50	0.50	0.00	0.00	1.00	-	0.00	0.63	1.00	0.48
10	0.40	1.00	0.50	1.00	1.00	0.00	1.00	-	0.00	0.75	0.50	0.61
11	0.20	0.83	0.00	0.00	0.00	0.00	0.42	0.40	0.00	0.13	0.50	0.23
12	0.40	0.83	0.00	0.67	1.00	0.50	0.50	1.00	0.50	0.63	0.50	0.59
13	0.20	1.00	0.00	0.33	1.00	0.50	1.00	0.50	0.00	0.50	0.50	0.50
14	0.20	0.83	0.00	0.00	1.00	1.00	0.33	1.00	0.00	0.13	0.50	0.45
15	0.60	0.83	0.00	1.00	1.00	0.50	0.63	1.00	0.00	0.38	0.50	0.58
16	0.40	1.00	0.00	0.00	1.00	0.00	1.00	1.00	0.00	0.38	0.50	0.48
17	0.20	0.83	0.50	0.50	1.00	1.00	0.50	0.20	0.00	0.50	0.50	0.52
18	0.20	1.00	0.50	0.50	1.00	0.50	1.00	0.00	0.00	0.38	0.50	0.51
19	0.40	0.50	0.00	0.00	1.00	0.00	0.50	0.25	0.00	0.38	0.50	0.32
20	0.20	1.00	0.00	0.00	1.00	0.00	0.33	0.20	0.00	0.25	0.50	0.32
21	0.20	1.00	0.00	0.50	1.00	0.50	0.53	-	0.00	0.13	0.50	0.44
22	0.20	0.83	0.00	0.67	1.00	1.00	0.50	1.00	0.50	0.38	0.50	0.60
23	0.20	1.00	0.50	0.00	0.00	1.00	0.33	0.50	0.00	0.75	0.50	0.43
24	0.20	1.00	0.00	0.33	1.00	1.00	0.50	0.25	0.50	0.25	0.50	0.50
25	0.20	1.00	0.50	0.33	1.00	0.50	0.33	-	0.00	0.50	0.50	0.49
26	0.20	0.83	0.50	0.00	1.00	0.00	0.33	0.75	0.00	0.13	0.50	0.39
27	0.40	0.83	0.00	0.00	1.00	0.00	-	-	0.00	0.38	0.50	0.35
28	0.60	1.00	0.00	0.33	1.00	1.00	0.50	0.25	0.50	0.75	0.50	0.58
29	0.20	0.00	-	1.17	0.00	0.00	0.00	0.00	0.00	0.38	0.50	0.22
30	0.20	0.83	0.00	0.00	1.00	1.00	0.50	0.25	0.00	0.38	0.50	0.42

Table A.3: The *Component Metrics* and *Article Model Metric* for each article

Appendix B

Discrepancies

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
P1	1	0	0	1	0	0	0	0	0	1	0	1	0	0	0	0	1	0	0	0	0	0
P2	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
P3	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
P4	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0
P5	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
P6	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0
P7	1	0	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
P8	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	1	1	0	0	1
P9	0	1	0	0	0	0	1	0	1	0	1	0	1	0	0	0	0	0	1	1	0	0
P10	0	1	1	0	1	0	1	1	1	0	1	0	0	1	0	1	0	1	1	0	0	1
P11	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0
P12	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
P13	0	0	0	0	0	1	0	0	1	0	0	0	0	0	1	0	1	1	0	0	1	0
P14	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
P15	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	1	0	0	1	1	0	0
P16	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
P17	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	1	0	0	1	0	0	0
P18	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	1	1	0	0
P19	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
P20	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table B.1: Table of problem categories encountered for each paper

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
A1	1	0	0	1	0	0	0	0	0	0	0	1	0	0	1	0	1	0	0	0	1	0
A2	1	0	0	1	0	0	0	0	0	0	0	1	0	0	1	0	1	0	0	0	0	0
A3	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0
A4	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
A5	0	0	0	0	1	0	1	0	0	0	0	0	0	1	0	0	0	1	1	0	0	0
A6	0	1	1	0	1	0	0	1	1	0	0	0	0	0	0	0	0	1	1	0	0	1
A7	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0
A8	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0
A9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
A10	1	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
A11	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	1	1	0	0	0	0
A12	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	1	0	0	0
A13	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	1	1	0	0	0
A14	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0
A15	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0

Table B.2: Table of assumption categories encountered for each paper

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
E1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
E2	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0
E3	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
E4	0	0	1	0	1	0	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0
E5	0	1	1	0	1	1	0	1	1	0	0	0	1	0	0	0	0	1	0	0	0	1
E6	1	1	1	1	0	1	0	1	1	0	0	0	1	0	0	0	0	1	0	0	0	0
E7	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	1	1	0	0	0	0
E8	1	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
E9	1	0	0	1	0	1	0	0	0	0	0	0	0	0	1	0	1	1	0	0	0	0
E10	0	0	0	0	0	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0
E11	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	1	0	0	0	0	0
E12	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
E13	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
E14	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
E15	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0

Table B.3: Table of error categories encountered for each paper