Advancing Code Readability: Mined & Modified Code for Dataset Generation

Lukas Krodinger

Master Thesis in M.Sc. Computer Science
Faculty of Computer Science and Mathematics
Chair of Software Engineering II

Matriculation number   89801
Supervisor   Prof. Dr. Gordon Fraser
Advisor   Lisa Griebl

20th February 2024

Abstract

This work presents an innovative method for generating datasets intended for code readability classification in the context of a master's thesis. We offer a comprehensive overview of code readability, delving into existing classifiers that consider both manually crafted and automatically extracted code readability features. Furthermore, we summarize existing datasets of manually annotated Java code snippets.

The core contribution of this work lies in the introduction of an automatic data generation technique, alongside a dataset produced using this methodology. Our approach relies on the extraction and modification of code snippets sourced from public GitHub repositories. Notably, our dataset significantly surpasses the scale of any previously available dataset designed for readability classification.

To evaluate the newly generated dataset, we conducted a user survey and train a state of the art code readability classification model, both with and without the integration of the new dataset. This analysis aims to assess the quality of our new dataset and the effectiveness of the generation approach employed.

Contents

1 INTRODUCTION

In the realm of software development, the significance of code readability cannot be overstated. Together with understandability, it serves as the foundation for efficient collaboration, comprehension, and maintenance of software systems [33, 1]. Maintenance alone will consume over 70% of the total lifecycle cost of a software product and for maintenance, the most time-consuming act is reading

code [9, 13, 35, 6]. Therefore, it is important to ensure a high readability of code. In order to archive this, we need to measure readability.

In the last years, researchers have proposed several metrics and models for assessing code readability with an accuracy of up to 81.8% [9, 33, 14, 12]. In recent years, deep learning based models are able to achieve an accuracy of up to 85.3% [27, 29]. However, these models do not capture what developers think of readability improvements [15]. This suggests that there is room for improvement in readability classification of source code.

In the following sections we will clarify what source code readability is. We will summarize background knowledge for both classical and deep learning based classification approaches. We will have a look at related work.

In section TODO we will explain our new dataset generation approach in detail. We will SUBSECTIONS

After introducing our new dataset we will evaluate it based on a user study and a state of the art readability classification model in section TODO. Based on the dataset, the user study and the model we will propose our research questions.

We will then answer our research questions based on results of the survey and model performance in section TODO.

We will show possible threads to the dataset generation approach and our evaluation and discuss them in section TODO.

In the last section we summarize our work, draw conclusions from it and propose future work.

2 BACKGROUND AND RELATED WORK

In the following subsections you find an overview of the background and related work on code readability and our approach to dataset creation.

2.1 CODE READABILITY

In the domain of software development, the importance of code readability cannot be emphasized enough. Alongside understandability, it forms the basis for effective collaboration, comprehension, and maintenance of software systems [33, 1].

It is a critical aspect of software quality, significantly influencing the maintainability, reusability, portability, and reliability of the source code [2, 38]. Poorly readable code increases the risk of introducing bugs [21, 36] and can lead to

higher costs during subsequent software maintenance and development [17]. On the other hand, readable code allows developers to identify and rectify bugs more easily [24].

Recent studies indicate that developers spend nearly 58% of their time reading and comprehending source code [9, 13, 35, 6, 41, 38, 44]. Therefore, it is important to ensure a high readability of code. In order to archive this, we need to define and measure code readability.

Buse and Weimer provides one of the first definitions: "We define readability as a human judgment of how easy a text is to understand."

Tashtoush et al. combines numerous other aspects from various definitions. According to them code readability can be measured by looking at the following aspects [41]:

- Ratio between lines of code and number of commented lines

- Writing to people not to computers

- Making a code locally understandable without searching for declarations and definitions

- Average number of right answers to a series of questions about a program in a given length of time

Recent definitions of code readability are shorter, trying to focus on the key aspects. Oliveira et al. defines readability as "what makes a program easier or harder to read and apprehend by developers" [31].

Also Mi et al. summarizes code readability as "a human judgment of how easy a piece of source code is to understand" [28]. This comes close to the definition of Buse and Weimer [9].

There are various related terms to readability: Understandability, usability, reusability, complexity, and maintainability [41] [41]. Among those especially complexity and understandability are closely related to readability.

Readability is not the same as complexity. Complexity is an "essential" property of software that arises from system requirements, while readability is an "accidental" property that is not determined by the problem statement [9, 7].

Readability is neither the same as understandability, as the key aspects of understandability are [36, 22, 43, 5]:

- Complexity

- Usage of design concepts

- Formatting

- Source code lexicon

- Visual aspects (e.g., syntax highlighting)

Posnett et al. states that readability is the syntactic aspect of processing code, while understandability is the semantic aspect [33].

Based on Posnett et al., Scalabrino et al. writes about readability: "Readability measures the effort of the developer to access the information contained in the code, while understandability measures the complexity of such information" [36, 33].

For example, a developer can find a piece of code readable but still difficult to understand. Recent research gives evidence that there is no correlation between understandability and readability [37].

Comparing the definitions of code readability in literature we can see, that there are some common aspects in most definitions. These are:

- Ease/complexity of understanding/comprehension/apprehension

- Human judgment/assessment

- Effort of the process of reading (differentiation to understandability)

Based on this, we come up with the following definition:

Definition 2.1 (Code readability): *Code readability is the human assessment of the effort required to read and understand code.*

In the last years, researchers have proposed several metrics and models for assessing code readability with an accuracy of up to 81.8% [9, 33, 14, 12]. In recent years, deep learning based models are able to achieve an accuracy of to 85.3% [27, 29] on available datasets. By using data augmentation the score can be improved even further to 87.3% [28]. Examining these works more closely in the following, we delve into their intricacies.

2.2 CLASSICAL CALCULATION APPROACHES

A first estimation for source code readability was the percentage of comment lines over total code lines [1]. Then researchers proposed several more complex metrics and models for assessing code readability [9, 33, 14, 36]. Those approaches used handcrafted features to calculate how readable a piece of code is. They were able to achieve up to 81.8% accuracy in classification [36].

```
1   /**
2   * Logs the output of the specified process.
3   *
4   * @param p the process
5   * @throws IOException if an I/O problem occurs
6   */
7   private static void logProcessOutput(Process p) throws IOException
8   {
9           try (BufferedReader input = new BufferedReader(new
     ↪   InputStreamReader(p.getInputStream())))
10          {
11                  StrBuilder builder = new StrBuilder();
12                  String line;
13                  while ((line = input.readLine()) != null)
14                  {
15                          builder.appendln(line);
16                  }
17                  logger.info(builder.toString());
18          }
19  }
```

Listing 1: An example for well readable code of the highly rated Cassandra GitHub repository

2.3 DEEP LEARNING BASED APPROACHES

In recent years code readability classification is dominated by machine learning, especially deep learning approaches. As the quality of the models increased, so did their accuracy (see Figure 1).

Mi et al. introduced a deep learning model called IncepCRM for code readability classification. IncepCRM automatically learns multi-scale features from source code with minimal manual intervention [26].

Mi et al. use Convolutional Neural Networks (ConvNets). Their proposed model, DeepCRM, employs three ConvNets with identical architectures, trained on differently preprocessed data [27].

Another study addresses concerns regarding the classification of code script readability by proposing an approach using Generative Adversarial Networks (GANs). The proposed method involves encoding source codes into integer matrices with multiple granularities and utilizing an EGAN (Enhanced GAN) [40].

Mi et al. address the limitation of the previous deep learning-based code readability models, which primarily focus on structural features. Their proposed method aims to enhance code readability classification by extracting features

```
1  private
2          static
3  void
4  debug( Process
5  v1
6  )        throws IOException
7  {
8          // Doo debug
9          try (BufferedReader   b
10         = new
11         BufferedReader(
12         new InputStreamReader(
13         v1.getInputStream()
14         )
15         )
16         )
17         {
18                 StrBuilder b2=new StrBuilder();String v2;while
       ↪  (null!=(v2=input.readLine())){b2.appendln(v2);}
       ↪  // Doo stuff
19                 m.info(  builder.toString()
20                 );
21         }
22  }
```

Listing 2: The same example as in listing 1 but modified to be poorly readable

```
1  /**
2   * Prints a personalized greeting message based on the provided name.
3   *
4   * @param name The name of the person for whom the greeting is
     ↪  intended.
5   *             If {@code null} or an empty string is provided, a
     ↪  generic greeting is displayed.
6   */
7  public static void printGreeting(String name) {
8          // Check if the name is provided
9          if (name != null && !name.isEmpty()) {
10                 System.out.println("Hello, " + name + "! How are you
                 ↪  today?");
11         } else {
12                 System.out.println("Hello! How are you today?");
13         }
14  }
```

Listing 3: A simple method that prints a greeting in Java

from visual, semantic, and structural aspects of source code. Using a hybrid neural network composed of BERT, CNN, and BiLSTM, the model processes RGB matrices, token sequences, and character matrices to capture various features [29].

Mi introduced a novel approach to code readability assessment by framing it as a learning-to-rank task. The proposed model employs siamese neural networks to rank code pairs based on their readability [23].

Mi et al. address the importance of code readability in software development and introduced a graph-based representation method for code readability classification. The proposed method involves parsing source code into a graph with abstract syntax tree (AST), combining control and data flow edges, and converting node information into vectors. The model, comprising Graph Convolutional Network (GCN), DMoNPooling, and K-dimensional Graph Neural Networks (k-GNNs) layers, extracts syntactic and semantic features [24].

You can find an overview over the accuracy scores for the models mentioned in Table 1.

Model	Accuracy (%)
Buse [9]	76.5
Possnet [33]	71.7
Dorn [14]	78.6
Scallabrino [**scalabrino2016improving**]	81.8
Mi_IncepCRM [26]	82.2 - 84.2
Mi_DeepCRM [27]	83.8
Sharma [40]	84.8
Mi_Towards [29]	85.3
Mi_Ranking [23]	83.5
Mi_Graph [24]	88.0

Table 1: Accuracy scores of two-class readability classification models

The main contribution of this work is not a model that outperforms a state of the art model but rather a new dataset (generation approach). For evaluation we opted for the Mi_Towards model (hereinafter referred to as towards model) from Mi et al. [29]. We did not choose the best performing one, Mi_Graph, as its main contribution is to use the AST representation of the code, while our dataset generation approach includes features that are not represented in the AST [24].

2.4 DATA AUGMENTATION

All the mentioned models were trained with (a part of) the data from Buse, Dorn and Scalabrino consisting of a total of 421 java code snippets. The data was generated with surveys. They therefore asked developers several questions, including how well readable the proposed source code is [9, 14, 36].

The problem that there is little data in the area of code readability classification for machine learning models has been recognized.

Mi et al. address the challenge of acquiring a sufficient amount of labeled data for training deep learning models. Due to the time-consuming and expensive nature of obtaining manual labels, the researchers propose the use of data augmentation techniques to artificially expand the training set. They employ domain-specific transformations, such as manipulating comments, indentations, and names of classes/methods/variables, and explore the use of Auxiliary Classifier GANs to generate synthetic data [28].

Mi et al. address the challenge of multi-class code readability classification. Due to a scarcity of labeled data, most prior research focused on binary classification. The authors propose an enhanced data augmentation approach, incorporating domain-specific data transformation and Generative Adversarial Networks (GANs) [25].

Vitale et al. introduce a novel approach to automatically identify and suggest readability-improving actions for code snippets. The authors develop a methodology to identify readability-improving commits, creating a dataset of 122k commits from GitHub's revision history. They train the T5 model to emulate developers' actions in improving code readability, achieving a prediction accuracy between 21% and 28%. The empirical evaluation shows that 82-90.8% of the dataset commits aim to improve readability, and the model successfully mimics developers in 21% of cases [42].

2.5 DIVERSE PERSPECTIVES

There is also other important research in the field of readability classification that does not directly affect this work, but could have implications for future work.

Fakhoury et al. showed based on readability improving commit analysis that previous models do not capture what developers think of readability improvements. They therefore analyzed 548 GitHubhttps://github.com/2023-07-25 commits manually. They suggest considering other metrics such as incoming method calls or method name fitting [15].

Oliveira et al. conducted a systematic literature review of 54 relevant studies on code readability and legibility, examining how different factors impact comprehension. The authors analyze tasks and response variables used in studies comparing programming constructs, coding idioms, naming conventions, and formatting guidelines [31].

In a recent study participants demonstrated a consistent perception that Python code with more lines was deemed more comprehensible, irrespective of their level of experience. However, when it came to readability, variations were observed based on code size, with less experienced participants expressing a preference for longer code, while those with more experience favored shorter code. Both novices and experts agreed that long and complete-word identifiers enhanced readability and comprehensibility. Additionally, the inclusion of comments was found to positively impact comprehension, and a consensus emerged in favor of four indentation spaces [34].

Choi, Park et al. introduced an enhanced source code readability metric aimed at quantitatively measuring code readability in the software maintenance phase. The proposed metric achieves a substantial explanatory power of 75.74%. Additionally, the authors developed a tool named Instant R. Gauge, integrated with Eclipse IDE, to provide real-time readability feedback and track readability history, allowing developers to gradually improve their coding habits [11].

Mi et al. aim to understand the causal relationship between code features and readability. To overcome potential spurious correlations, the authors propose a causal theory-based approach, utilizing the PC algorithm and additive noise models to construct a causal graph. Experimental results using human-annotated readability data reveal that the average number of comments positively impacts code readability, while the average number of assignments, identifiers, and periods has a negative impact [30].

Segedinac et al. introduces a novel approach for code readability classification using eye-tracking data from 90 undergraduate students assessing Python code snippets [39].

2.6 DATA GENERATION

In addition to related work on models and datasets, there is also related work that uses some of the ideas that we will employ in our proposed approach to data generation.

Loriot et al. created a model that is able to fix Checkstyle[1] violations using Deep Learning. They inserted formatting violations based on a project specific format checker ruleset into code in a first step. They then used a LSTM neural network that learned how to undo those injections. Their approach is working on abstract token sequences. Their data is generated in a self-supervised manner [20]. A similar idea has been explored by Yasunaga and Liang [45]. We will use the idea of intentional degradation of code for data generation.

Another concept we will employ is from Allamanis et al. They cloned the top open source Java projects on GitHub[2] for training a Deep Learning model. Those top projects were selected by taking the sum of the z-scores of the number of watchers and forks of each project. As the projects have thousands of forks and stars and are widely used among software developers, they can be assumed to be of high quality [3].

3 MINED AND MODIFIED CODE FOR DATASET GENERATION

In the following subsections we will describe our approach.

3.1 WORK ON EXISTING DATASETS

Most of the related work (see Section 2) uses a combination of Buse and Weimer, Dorn and Scalabrino et al. data. The raw data from their surveys can be downloaded [3], but their data is not uniformly formatted, including ratings that are not Java code snippets, as well as the individual ratings rather than the mean of the ratings used for training machine learning models.

We converted and combined the three datasets into one: code-readability-merged. In recent years, Huggingface [4] established as the pioneer in making models and datasets easily available. Therefore we decided to publish the merged dataset on Huggingface [5].

3.2 CLASSIFICATION CONSIDERATIONS

In code readability classification it is still state of the art to make a binary classification into readable and unreadable code [26, 27, 40, 29, 23].

---

[1] https://checkstyle.org/, accessed: 2023-07-25
[2] https://github.com/, accessed: 2023-07-25
[3] https://dibt.unimol.it/report/readability/, accessed: 2024-02-18
[4] https://huggingface.co/, accessed: 2024-02-18
[5] https://huggingface.co/datasets/se2p/code-readability-merged, accessed: 2024-02-18

However, code readability classification is not a binary classification task per se. Mi et al. introduced a third, neutral class to address this problem [24].

When rating code snippets, a Likert scale [18] from 1 to 5 was mostly used. So one could argue, that there are actually 5 classes rather than two or three. However, the actual readability score of a code snippet lies somewhere in between. It can have any value from 1.0 to 5.0 and thus the range is continuous, not discrete. In particular, if you normalize by subtracting 1 and divide then by 4, we get a continuous value between 0.0 to 1.0. At this point, it is trivial to say that it is basically a regression problem.

The question arises as to why research in the field of code readability always views the problem as a (binary) classification problem. The answer to this question is probably worthy of its own research and therefore not covered in this work.

Our evaluation model is the towards model of Mi et al. [29]. Therefore, we want to show how they transformed the rating scores into a binary classification problem. First, the mean values of all scores are calculated. In a second step, the snippets are ranked according to their mean score. Then, the top 25% of the data is labeled as well-readable and the bottom 25% is labeled as badly-readable. The 50% of the data in between is not used at all [29].

While this transformation is fine in principle, especially with the argument that the data in the middle is not clearly either readable or unreadable, it has drawbacks that only 50% of the available data is used for model training and evaluation:

First, the available data is further reduced from 421 evaluated Java code fragments to 210 code fragments for training. Note that a bottleneck in readability classification is the small amount of available data. So this is a significant loss.

Secondly, evaluation is performed with only those 210 snippets as well. This means, that the model was only evaluated on 50% of the available data. We suspect that this might be a thread to validity. It could be that the performance of the model is remarkably lower when the evaluation is performed with random, unseen data that also contains moderately readable code snippets.

However, we will both stick to the binary classification approach as well as to the towards model [29] to make our results comparable to theirs.

3.3 DATASET GENERATION APPROACH

Other than previous datasets for readability classification, our dataset is generated using an automated approach. The aim is to mine modify code from GitHub
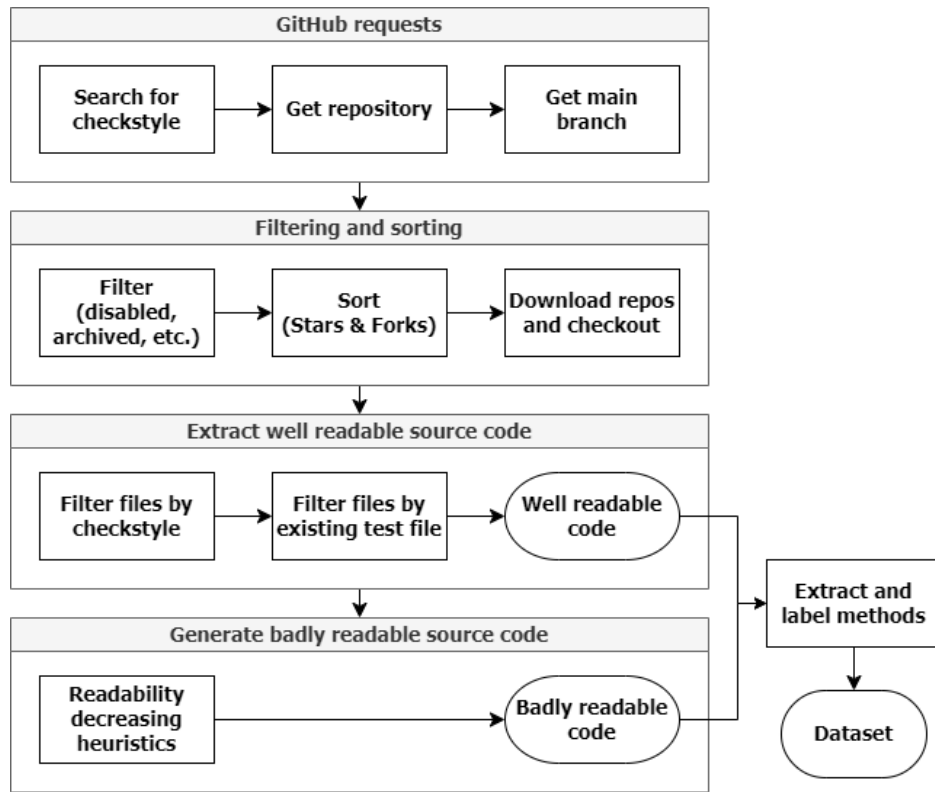
Figure 1: The used dataset generation approach.

to obtain both readable and unreadable code. This approach is novel to the best of our knowledge. You can find a visualization in Figure 1.

The approach is divided into four parts. The first three steps are used to mine well readable java code. In a final step, we will modify the well readable code to achieve our second goal, namely poorly readable source code.

We start by querying the GitHub REST API[6] for repositories that use checkstyle (query string: "checkstyle filename:pom.xml"). The repository informations (including the URL) are stored together with the main branches. We remove all repositories that do not fulfill these criteria:

- The repository is not a fork of another repository

- The repository is not archived

- The repository is not disabled

- The repository language is Java

- The repository has at least 20 stars

- The repository has at least 20 forks

The remaining ones are sorted by their star and fork count (equally weighted). The 100 best are cloned and their main branch is checked out.

In a third step we run checkstyle ([7]) against the project own checkstyle configuration to get all java class files, that pass the own checkstyle test. From the java classes that passed this filter we extract all methods that have a comment of any kind at the beginning of the method. This results in 36077 code snippets which we assume to be well readable.

The fourth and final step is to generate badly readable code from the well readable one. Therefore we use the proposed Readability Decreasing Heuristics (RDH, see section 3.4). Afterwards we again extract all methods with a comment at the beginning of the method. Initially we planned to not require comments for the badly readable dataset part. However, it turns out that in this case all well-readable methods have a comment while most of the badly-readble do not have one. This lead to shortcut learning, whether a method has a comment or not instead of learning to distinguishing the methods by all other criteria as well.

3.4 READABILITY DECREASING HEURISTICS

We already mentioned, that Readability Decreasing Heuristics (RDH) are used to generate badly readable code from well readable code. In this section we will explain how we achieved this and what is meant by RDH. The RDH are a set of code manipulation heuristics that are applied to Java files. One part is performed on the abstract syntax tree (AST) representation of the Java files using the spoon library ([8]) [32]. Another part is performed when converting the AST back into Java files. From now on, we will use the abbreviation RDH or RDH tool when referring to the entire program. If a specific manipulation or a subset of all manipulations is meant, we will use [name] rdh instead.

The RDH tool converts the Java code of each well readable Java class file into an AST. In the end the AST is parsed back to Java code using an pretty printer. If nothing else is done, this results in the "none" rdh. Note that the code produced by the tool in this way will be slightly different from the original input code, as

---

[7]`https://checkstyle.sourceforge.io/`, accessed: 2024-02-15
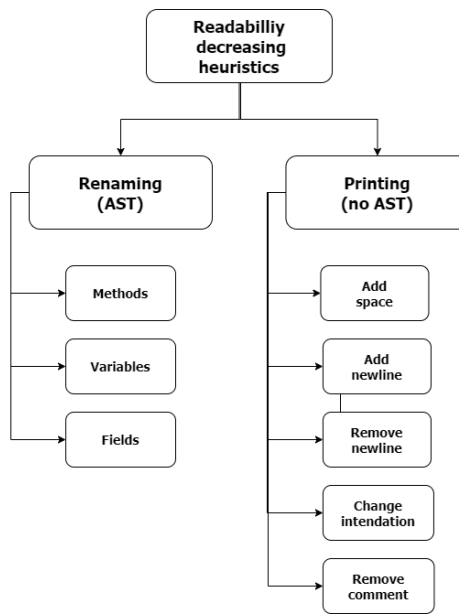[8]`https://spoon.gforge.inria.fr/`, accessed: 2024-15-02

Figure 2: The available Readability Decreasing Heuristics (RDH)

the styling and formatting of the original code will be overwritten by the default formatting of the Java Pretty Printer of the spoon library.

Various code changes can be made between the two steps and during printing (see Figure 2). The renaming is done while the code is in its AST representation to ensure that the declarations and usages of variables, fields and methods are all renamed to the same new name. The other transformations are performed when the AST is converted back to source code. This includes adding additional spaces, adding or removing newlines and changing the indentation of code in term of tabs.

The individual methods are then extracted from the class files. As already mentioned, we need a method comment for all methods. We therefore use the RDH "remove-comment" after completing the method extraction.

The RDH works with a configuration file in which one can specify a probability for each heuristic that can be applied. We have chosen the probabilities so that the generated code snippets are still realistic in the sense that they could also be written by humans. You can find the configuration file for the none rdh in Appendix ??.

Some of the heuristics were not included in the final version of the new dataset. You can find them in table 2 together with the reason why they were not included.

Heuristic	Reason
`inlineMethod`	Makes methods to long
`add0`	Limited survey capacity
`insertBraces`	Limited survey capacity
`starImport`	No effect after comment extraction
`inlineField`	Limited survey capacity
`partiallyEvaluate`	Might increase readability

Table 2: Readability Decreasing Heuristics that are not included in the final version of the new dataset and why

We have also added Code2Vec [4] to the tool. This makes it possible to rename methods not only to iterating or arbitrary strings or numbers, but also to other realistic method names. The idea was to use worse method names predicted by Code2Vec and rename the methods to these. However, due to time and resource constraints regarding the survey, we did not pursue this approach. However, there is a corresponding mode supported by the tool. This can be used for further research.

3.5 SURVEY

We evaluate the dataset generated with the new approach with a survey. Therefore we had to carefully sample suitable code snippets from the dataset. You can find an overview over the approach in Figure 3.

The first step was to find realistic configurations for the readability decreasing tool. After creating an initial dataset using the heuristics a pilot survey was conducted. Afterwards the heuristics that seemed to be too strong were weekend and adjusted according to the pilot survey result. The result of this were 9 different rdhs, which can be found in Table 3.

Together with the original methods this resulted in 10 different configurations.

The configurations are based on probabilities for different heuristics. A heuristic is applied with the given probability one each occurrence of the object to apply it to. For example, comment_remove is applied with a probability of 10% on each comment that occurs within the code snippet. Thus, the exact amount of changes for one method is uncertain. It might even happen (especially for short methods, such as getters and setters) that a method is not changed at all. For example, if a method has only a single comment and we apply comment_remove, it is very likely, that a method is not changed at all.
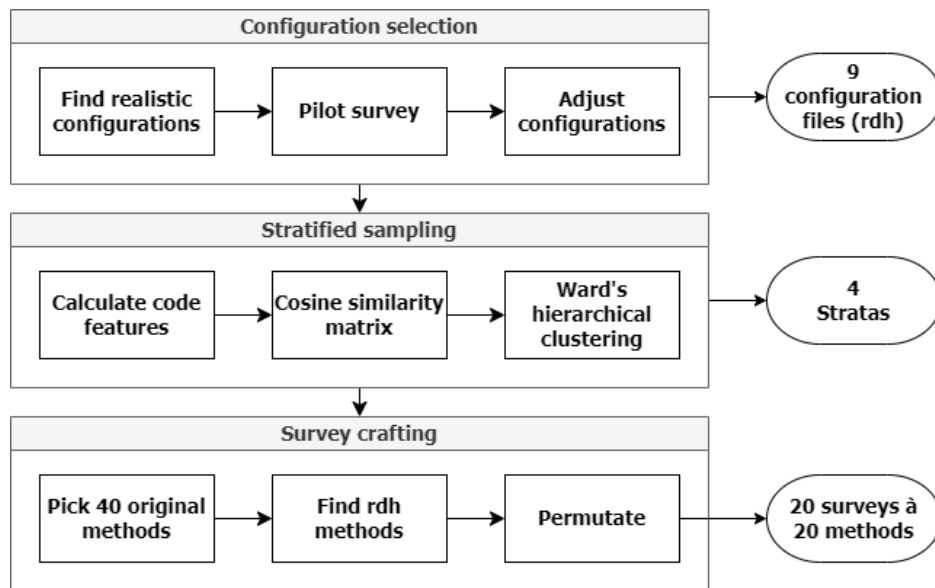
Figure 3: Overview over the steps performed to craft survey sheets from the dataset.

In a second step we applied stratified sampling to those to be able to distinguish between very simple methods such as getters and setters and more complex methods. In order to be able to compare original methods with their modified variants, we performed the sampling only on the original methods and matched the rdh methods to those in a later step.

Thus we first had to calculate features for the original code snippets. This was done using TODOs tool. Therefore a TODO-dimensional features vector was calculated for each original code snippet. Next we calculate the cosine similarity metrix between all feature vectors and finally, using Ward's hierarchical clustering we were able to cluster the methods into an arbitrary amount of clusters.

By comparing the merge distances in each step (see Figure 4) we evaluated that a cluster size of 4 makes most sense: The merge distance from 5 to 4 is small, so we should still do this merge, but the merge distance from 4 to 3 is large, so one should rather not do this merge. Additionally, 4 is the size with the last possibility for a small merge distance. We assigned each cluster/strata manually a name, which can be found in Figure 4.

In the third step we crafted the surveys from the stratas. We decided to provide for each original method all 10 configurations mentioned earlier as we want to compare the original methods to their rdh-variants. As we have a capacity of

Configuration	RDH probabilities
none	-
comments_remove	removeComment: 1.0
newline_instead_of_space	newLineInsteadOfSpace: 0.15
newlines_few	removeNewline: 0.3
	spaceInsteadOfNewline: 0.05
newlines_many	add1Newline: 0.15
	add2Newlines: 0.05
rename	renameVariable: 0.3
	renameField: 0.3
	renameMethod: 0.3
spaces_many	Add1Space: 0.2
	Add2Spaces: 0.1
	spaceInsteadOfNewline: 0.05
tabs	remove1IncTab: 0.2
	add1IncTab: 0.1
	remove1DecTab: 0.1
	add1DecTab: 0.1
	incTabInsteadOfDecTab: 0.05
	decTabInsteadOfIncTab: 0.05
all7	all probabilites/7

Table 3: Chosen configurations and their probabilities for the RDH tool

rating 400 code snippets, we require to sample 40 original code snippets (and than add all their rdh-variants).

We chose to sample randomly from within the stratas. However, we distributed the 40 snippets among the stratas as can be seen in table 5.

This decision was made due to the relative high frequency of methods not being different from their original methods (see Figure 5). Another factor for this decision is, that especially simple methods are rather uninteresting for readability classification, as they are often generated and mostly follow a straight forward pattern.

After sampling the 40 original methods, we next sampled all 9*40 rdh-variants that belong to the original methods. This was done mostly automatically using the sample name of the original methods and the sample names of the rdh-variant-methods. However, if the method was renamed earlier due to the rename-method-heuristic, the new method name did not match the one of the original method anymore and therefore we had to match those manually.
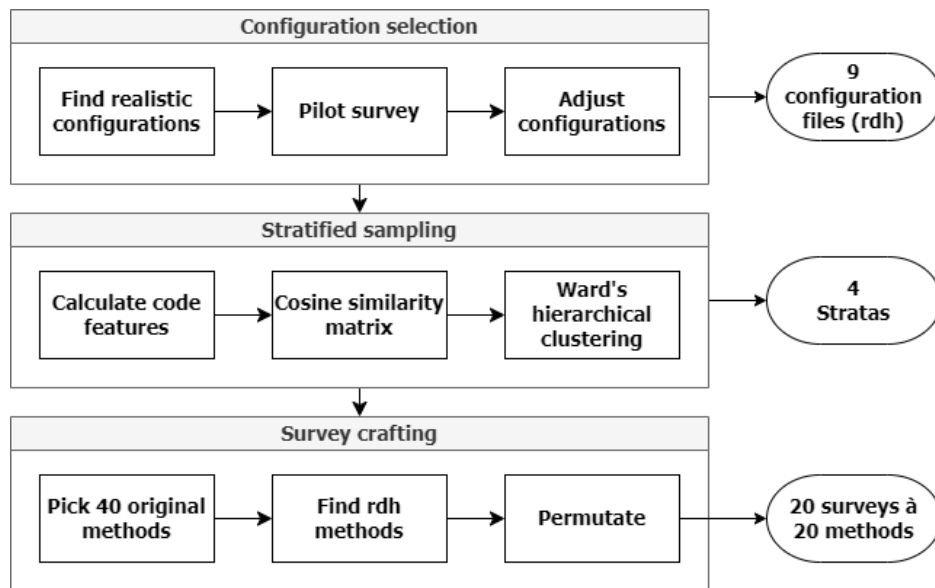
Figure 4: TODO: Replace. The merge difference and the local derivative for amount of clusters between 2 and 10.

Stratum	Method Type
Stratum 0	Simple methods
Stratum 1	Complex methods
Stratum 2	Magic number methods
Stratum 3	Medium complex methods

Table 4: The created stratas and a manually assigned name based on the methods within.

After gathering all 400 methods we than had to distribute those among the 20 survey sheets, 20 snippets each. To not manipulate the raters we decided, that a variant of each method can only be once in each survey sheet. For example, if the original method is in survey sheet 1, the comment_remove-variant (or any other variant) must not be in the same sheet.

To achieve this, we created four permutation matrices with 10 snippets each. 10 was chosen, as it is possible to distribute 10 snippets with 10 variants each among at least 10 survey sheets while not violating our condition. By combining two 10-permutation-matrices we could create 10 survey sheets with 20 code snippets each. One implication of this approach is, that each survey sheet has each variant exactly twice. By doing this twice, we ended up with a desired

Stratum	Percentage	Count
Stratum 0	10%	4
Stratum 1	40%	16
Stratum 2	10%	4
Stratum 3	40%	16
Total	**100%**	**40**

Table 5: Distribution of Sampled Methods

distribution of 20 survey sheets à 20 methods. Additionally, our condition holds: In each survey sheet there is only one variant of the same method.

Finally the methods for each survey sheet where randomly shuffled within each survey sheet so that the positions of the variants within each sheet are randomly distributed. This was done to minimize any effect of the position of a snippet/variant within a survey on the rating.

3.6 READABILITY CLASSIFICATION MODEL

We will investigate whether it is possible to score a higher accuracy as current models in classifying code readability for Java using Deep Learning. Therefore, we will train the model from Mi et al. [29] with more data. We will consider augmenting the model with a method name classifier and incorporating semantic encoding for tabs and spaces. The training data will be generated in a novel way for classification of readability, inspired by Loriot et al. [20]. The method name classifier is similar to Code2Vec [4]. The combination of all components is novel to the best of our knowledge. You can find a visualization of the planned modifications of Mi et al.'s model in figure 6. We will focus on generating training data, as the approach will be usable for further research in the field of source code readability.

Deep Learning based models perform better the more training data they get [16]. Therefore, one approach in order to further improve existing models is to gather more training data. This requires, as it was done previously, a lot of effort and persons willing to rate code based on their readability. We present another approach for gathering training data.

In a first step, GitHub repositories with known high code quality are downloaded and labeled as highly readable. We select repositories using a similar approach as Allamanis et al. [3] and then assume that they contain only well readable code. In a second step, the code is manipulated so that it is subsequently less
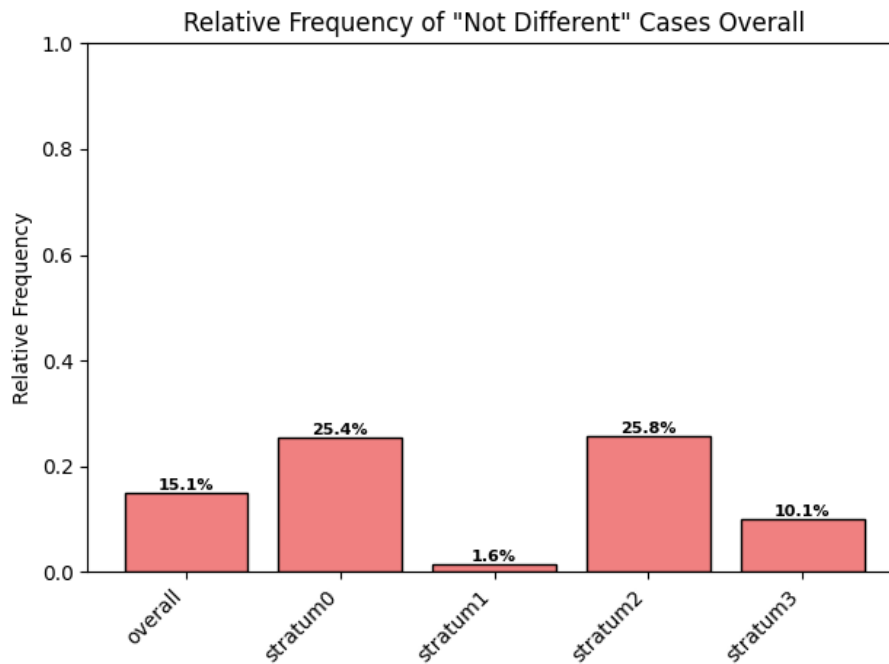
Figure 5: Relative frequency of the case that a rdh-method is not different from it's original method.

readable. This approach is similar to the approach of Loriot et al. [20]. After both steps, we have a new, automatically generated training dataset for source code readability classification.

This brings up the question, how to manipulate code so that it is less readable afterwards. We therefore introduce a tool called Readability Decreasing Heuristics. As the name suggests this is a collection of heuristics that, when applied to source code, lower the readability of it. For example such a heuristic is to replace spaces with newlines. Another example is to increase the indentation of a code block by a tab or multiple spaces. Moreover, with most changes it is also possible to do exactly the opposite (replacing newlines with spaces, decreasing indentation), which in most cases also decreases the readability of source code.

Code snippets in Java are syntactically the same, before and after applying Readability Decreasing Heuristics. Complexity did not change either. However, if various modifications are applied many times, those changes are capable of lowering the readability of source code, as the comparison of listing 1 and listing 2 suggests.
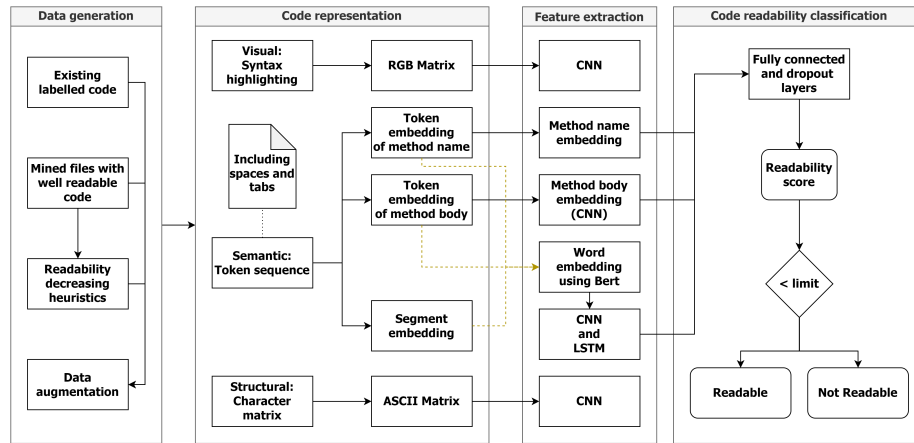
Figure 6: Overview of the planned approach.

Note that we assume two things for the data generation approach:

Assumption 1 **(well-readable-assumption)** The selected repositories contain only well-readable code.

Assumption 2 **(poorly-readable-assumption)** After applying Readability Decreasing Heuristics, the code is poorly readable.

In recent years it was shown that Deep Learning models can be further improved by modifying the structure of the architecture or by introducing new components, parts or layers to existing architectures. We suggest two improvements for the model of Mi et al. [29]. Firstly, we want to embed spaces and tabs as semantic tokens. Secondly, adding a method name fitting classifier as a component of the overall model could be an improvement. If there is time left, we will try to surpass the performance of recent source code readability classifiers with those improvements to data generation and the model.

We will evaluate our suggestions with two methods. Firstly, we conclude a user study. Secondly, we compare code readability models with each other.

Figure 7: Overview over the used scripts and their output

3.7 CODE BASE

Our code is available on GitHub. TODO: Refs

You can find an overview over all programs used to create the merged dataset, the new dataset the model and all our evaluation results in Figure 7.

While re implementing the original model we came across the following possible issue of the model: The token length for the bert encoding (bert-base-cased) used within the model is set to 100. What is a token in a piece of code? Besides special tokens to mark the beginning [CLS] and end [SEP] of the input, each word is represented by a token. However, also each special character (such as /(),;= and many more) is represented by a own token. Java identifiers are split using the camel case convention. Long words are split again into multiple tokens.

Consider the method from Listing 1. With a token limit of 100, the last encoded token is the last closing brace in line 9. Everything from line 10 onward is not encoded anymore, which means, that the information is lost for the semantic

encoder. To put it in other words: The model of Mi et al. considers only the first few lines of code snippets in it's semantic component.

The visual and structural have similar limitations, however, they are less conservative. The structural encoder encodes the first 50 lines of each code snippet and the visual encoder encodes the first 43 lines. While the restrictions for both of those encoders seem to be long enough to capture most of the code snippets fully, the semantic encoder seems to be too limited to capture most of the code snippets fully. One reason for this could be, that the first part of a method is of higher semantic importance.

While we want to remark these limitations, we will stick with them to allow later a fair comparison of the datasets against the model of of Mi et al. trained on the combined dataset.

3.8 USER STUDY

The goal of the user study is to answer the following key questions:

1. Does the well-readable-assumption (assumption 1) hold?
2. Does the poorly-readable-assumption (assumption 2) hold?

We will achieve this by showing programmers code snippets that were generated with the presented approach. Therefore, human annotators give each code snippet a rating of its readability. The annotators are selected by prolific[9]. Particular attention is paid to a high proportion of people from industry. The readability rating is based on a five-point Likert scale [18] ranging from one (i.e., very unreadable) to five (i.e., very readable). We apply the same rating as done previously [9, 14, 36], but, other than before, we will not use the rating for labeling the training data. Instead, we will only use the ratings to validate a few randomly selected code snippets out of many that are automatically labeled.

COMPARING MODELS

Besides the user study we will evaluate our suggestions by comparing machine learning models against each other. The comparisons are based on common metrics such as accuracy, F1-score and MCC [10]. One can distinguish further between the following variants of comparing models:

In one variant we compare models that have the same architecture (same layers, same weight initialization, same components, etc.) while they differ in the data

---

[9] https://www.prolific.com/, accessed: 2023-09-30

they are trained on. For example, we can train a model with the old and new data-sets, separately and combined. If done for multiple model architectures we can evaluate how the differences in training data influence the model performance.

Another variant would be to compare models with different architecture but the same training data. In this way, we can evaluate newly introduced components by measuring and comparing the performance of such models.

A third comparison variant is created by combining the first two. Both of them lead to many options in what to compare, especially if only small changes to training data or model architecture are done. To find out, if our suggestions lead to a better model overall, we will compare our newly created model with all changes at once to the state-of-the-art model of Mi et al. [29].

RESEARCH QUESTIONS

We come up with the following research questions:

Research Question 1: *(select-well) Can automatically selected code be assumed to be well readable?*

In our new approach for generating training data, we assume that the code from repositories is readable under certain conditions (assumption 1). We want to check whether that holds. To answer this question we will use the results of the user study (section 3.8).

Research Question 2: *(generate-poor) Can poorly readable code be generated from well readable code?*

It is not sufficient to have only well readable code for training a classifier. We also need poorly readable code. Therefore, we will try to generate such code from the well readable code. We will investigate whether this is possible in principle, and we will propose an automated approach for archiving this: Readability Decreasing Heuristics.

As the name already suggests, the applied transformations on the source code are only heuristics. To answer, whether the generated code is badly readable (assumption 2) we will utilize the results of the user study (section 3.8).

Research Question 3: *(best-heuristics) Which heuristics are best to generate poorly readable code from well readable code?*

We want to compare the modifications of the proposed heuristics for generating poorly readable code to each other. Therefore we will train the same classifier

model with badly readable code generated by different Readability Decreasing Heuristics. We will then evaluate the model variations against each other (section 3.8) to answer the research question.

Research Question 4: *(new-data) To what extent can the new data improve existing readability models?*

It was shown that Deep Learning models get better the more training data is available [16]. This holds under the assumption that the quality of the data is the same or at least similar. We want to check if the quality of our new data is sufficient for improving the Deep Learning based readability classifier of Mi et al. [29]. Therefore we will train their proposed model with and without the new data and then evaluate the models against each other (section 3.8).

Research Question 5: *(embedding-spaces) Optional: To what extend does the embedding of spaces and tabs in semantic code representations improve readability classification?*

The state-of-the-art model of Mi et al. [29] does consider spaces and tabs only in its visual component. We want to investigate if it can improve the quality of a Deep Learning based model if spaces and tabs are encoded as semantic tokens. We also want to investigate if this makes the visual component superfluous. We will evaluate the proposed improvement as described earlier (section 3.8).

Research Question 6: *(name-classifier) Optional: To what extend does the usage of a method name classifier improve readability classification?*

Correct naming of identifiers is crucial for ensuring readability of software programs. It is of outstanding importance for readability of code that the name of methods fit the method bodies [19]. We want to introduce a new component to the model of Mi et al. [29] that is built similar to Code2Vec [4]. We want to investigate if the newly introduced component improves the quality of the resulting model. We will evaluate the proposed improvement as previously described (section 3.8).

4 EVALUATION

In this section we will present our evaluation results.

4.1 SURVEY RESULTS

The results from our survey are separated into two parts: The results from the pilot survey used to improve the main survey pre start and the main survey results used to answer our research questions.

PILOT SURVEY

The pilot survey was used to adjust the rdh-probability settings and the survey itself before start.

The pilot survey gave an insight into how much time one would need to rate 20 code snippets regarding it's readability. You can find an overview over the times needed in Figure 8. Based on this we estimated the time requirement for our survey with 10 minutes.

Most of the problems that came up where due to the survey tool (for example: "I also thought that I should use the drop-down menu on the upper left."). You can find all feedback in appendix TODO.

Additionally some manual evaluation of the different rdhs, especially for stratum 1 suggested a few things:

First of all, original methods where rated comparably well, which suggests the well-readable-assumption to hold.

Secondly, the rdh tool did always fully declare each imported method or class with its fully specified classifier. For example, instead of "InputStream", "java.io.InputStream" was declared. This gave the raters the feeling, that the code was actually not written by a human and drastically decreased it's readability. Afterwards, we adjusted the rdh-tool to not print the fully qualified name anymore. This has however the drawback, that it can not be guaranteed anymore that the generated code compiles and does the same as it's original. However, this property was lost anywhere upon extracting the individual methods from the class files.

Third, some rdhs where configured to strong so that some methods were again not presumed to be written by a human anymore. These rdhs were adjusted and their probabilities were decreased accordingly (for example newlines_few). For all results see Table 6.

After the feedback regarding the survey tool was implemented, the actual study was conducted.

PROLIFIC STUDY RESULTS

In this section we will summarize the results of the main study conducted via prolific.

You can find an overview over the time required by the participants in the Figures 9 and 10. Additionally you can find the amount of participants that required less than 3 minutes for one survey sheet in Figure 11. The results of those participants might be a thread to validity to which we will come back later.

You can find an overview over the familiarity of the participants with Java in Figure 12.

You can find an overview of the ratings for each RDH for all stratas combined in Figure 13.

We evaluated whether the deviation in the ratings between the various RDHs has statistical significance. We therefore used the Mann-Whitney-U-Test comparing

Stratum	Feature	File	Score
stratum_3	methods	hbase_PreviousBlockCompressionRatePredicator.java_updateLatestBlockSizes.java	4,6
stratum_0	tabs_few	hadoop_DataNodeFaultInjector.java_delayWriteToDisk.java	4,3
stratum_2	tabs_few	framework_WindowMoveListener.java_getTicketNumber.java	3,8
stratum_1	methods	flink_HiveParserSemanticAnalyzer.java_processPTFSource.java	3,7
stratum_2	methods	hibernate-validator_PESELValidator.java_year.java	3,7
stratum_3	newlines_many	hadoop_CompressionCodec.java_createInputStreamWithCodecPool.java	3,3
stratum_1	comments_remove	flink_MemorySegment.java_get.java	3,1
stratum_0	spaces_few	hadoop_FilePosition.java_bufferFullyRead.java	3
stratum_1	all_weak_3	hadoop_SingleFilePerBlockCache.java_getIntList.java	3
stratum_1	newlines_many	pulsar_LoadSimulationController.java_writeProducerOptions.java	2,9
stratum_1	spaces_few	hbase_HBaseZKTestingUtility.java_startMiniZKCluster.java	2,6
stratum_1	misc	hbase_Encryption.java_decryptWithSubjectKey.java	2,4
stratum_2	newlines_few	zxing_ModulusPoly.java_isZero.java	2,4
stratum_1	tabs_few	hbase_ZKMainServer.java_main.java	2,2
stratum_1	tabs_many	flink_HiveParserSemanticAnalyzer.java_findCTEFromName.java	2,2
stratum_1	spaces_many	hadoop_ActiveAuditManagerS3A.java_createExecutionInterceptors.java	2,1
stratum_1	newlines_few	hudi_Pipelines.java_hoodieStreamWrite.java	1,7
stratum_3	tabs_many	hadoop_S3AInputPolicy.java_getPolicy.java	1,7
stratum_1	all_weak	flink_SSLUtils.java_createInternalNettySSLContext.java	1,3
stratum_1	all	hudi_HoodieMultiTableStreamer.java_populateTableExecutionContextList.java	1,2

Table 6: Overview over the pilot survey results.

the ratings for all snippets for a RDH against the corresponding none snippets. You can find the results of this in Table 7.

Our results suggest, that doing no modification (none) is not different from the original methods. The difference here might as well come from a random variation with a probability of 92%. When we compare the RDHs against the none methods we can be sure, that the ratings of all but newlines many and rename are actually statistically different from the ratings of none. This confirms that the readability decreasing heuristics are actually working in the sense that they decrease the readability of the given input methods.

Condition	P-value	Reject
None - Methods	0.9224	False
None - Newlines Few	5.23×10^{-6}	True
None - Spaces Many	0.0407	True
None - Newlines Many	0.3006	False
None - Comments Remove	0.00364	True
None - Rename	0.099	False
None - Newline Instead Of Space	4.57×10^{-6}	True
None - Tabs	3.06×10^{-8}	True
None - All7	1.80×10^{-7}	True

Table 7: Mann-Whitney U Test Results of none against each RDH. If the p-value is smaller or equal to 0.05 the null-hypothesis is rejected confirming, that the ratings are actually different and not just a random variation. TODO: Update data

4.2 MODEL TRAINING RESULTS

You can find the training, evaluation and fine-tuning results in table 8.

Train	Eval	Acc	Prec	Rec	AUC	F1	MCC
Krod	Krod	0.918	0.923	0.913	0.918	0.917	0.836
Krod	All	0.619	0.636	0.636	0.636	0.636	0.236
All	Krod	0.538	0.526	0.778	0.652	0.628	0.087
All	All	0.847	0.877	0.823	0.850	0.837	0.704
Krod-FT-All	All	0.804	0.840	0.738	0.789	0.772	0.600
Krod210	Krod210	0.809	0.827	0.776	0.802	0.789	0.609

Table 8: Performance of different dataset configurations for the same model. FT = Fine-Tuning

When training the towards model on the Krod dataset and evaluating it on it using 10-fold cross validation, we get an average accuracy score of 91.8%. However, when we evaluate the trained model on the combined dataset used in the towards paper, we get an accuracy of only 61.9%. We can conclude a couple of things from this:

The towards model works well for our new dataset. However, the readability retrieved using the all dataset is different from the readability using our approach. Otherwise, the scores for all-krod and krod-all would be similar to the all-all score. This suggests, that our dataset is not suitable for general readability classification, but we might have found a subproblem. However, adding more readability decreasing features and adding well-designed data augmentation might overcome this limitation.

When we compare the accuracy scores for all-krod and krod-all we get another interesting finding: While the model trained on the krod dataset is able to classify readabilty to some extend, this does not hold the other way around, as 53.8% is almost a random classifier. This suggests, that fine-tuning a model trained on the krod dataset using the all dataset might score better results than the original model proposed in the towards paper.

To verify that the model works correctly and as a baseline comparison score we also added the all-all case. Here we achieve the same accuracy as Mi et al., which suggests that our model is implemented in the same way as theirs.

We tried to do fine-tuning using the all dataset by freezing different layers of the model trained on the krod dataset. The best we could come up with was when freezing the input layers as well as the first convolution and pooling layer of all encoders. However, the performance of this fine-tuned model was still worse than the baseline model. This is in direct contrast to our previous suggestion, that fine-tuning in this way might score better results. One explanation for this might be, that the model is to small to deal with the now larger amount of data. Introducing more or bigger layers, so that the model can memorize more features internally might overcome this problem. However, this is not part of this work, where we soely deal with creating a new dataset.

We also trained the model with a random sample of 210 datapoints to get a initial insight of what a change in the training size might accomplish. As we can see, the model has similar metrics as the all-all model. When we now compare the stats, for example the accuracy of the krod-krod against the all-all model, we can see, that an improvement of about 8% might be possible with a larger dataset. This suggests the importance of finding new ways of data generation for readability classification.

The readability ratings of code snippets mined from Github are not very accurate. Therefore the well readable assumption TODO only holds for certain clusters of code snippets: TODO. For clusters that can be labeled with X or Y this assumption does not hold. Therefore we labeled the mined code depending on the cluster they where grouped in as you can see in Table TODO. While the rating does not hold for each and every snippet within such a cluster it is a good estimation on average. This should be sufficient to train a readability classifier.

The badly readable assumption holds. Especially the heuristics X Y and Z decrease the readability by a significant extend. We estimate the readability decrease for a certain probability of a certain type as can be seen in table TODO. We therefore calculate the readability of a new snippet by taking the original readability score and decreasing the readability percent depending on the probability of such a refactoring beeing applied. In Table TODO you can see how can to which extend we combined certain manually selected probabilities and how we then calculated the new readability score.

When we compare the model of TOWARDS trained with the old dataset we can reproduce the results described in their paper. Once we add our own dataset we achieve an accuracy improvement of TODO. You can find a detailed comparison in Figure TODO.

When we scale up the architecture by increasing XY we can achieve an even higher accuracy of TODO when using the new training data. However, with the old dataset only the results are worse. This suggests that the new fits the new larger dataset better while the architecture of TOWARDS was built for small datasets.

By combining the results of the study and model evaluation we can answer the third research question: The rating of our automatic dataset generation approach is not as accurate as letting multiple users rate a certain code snippet. However, being able to fully automate the dataset generation approach and the vast amount of data generated that way makes up for this. The larger dataset improves the performance, especially of our new adjusted model by a significant extent of TODO. Therefore we conclude, that our new dataset and it's generation approach is suitable for training deep learning based code readability classifiers.

5 DISCUSSION

The biggest thread to our approach is the reliance on heuristics for the dataset generation. We can not show, that the labeled code snippets of our dataset actually fit the score we assigned them. This would require an manual evaluation

of X code snippets by human annotators and is therefore not feasible. We can reduce the extent of this with our model results.

TODO: Copy from study paper

TODO: Copy from towards and newer model

6 CONCLUSIONS

TODO: Add conclusion

The new dataset has another advantage that is not yet utilized in this work: For the first time there is a dataset with one well readable and a second one less readable code snippet that is functionally equivalent. This could be used to train a transformer on source code readability improvement. Such a transformer could take code as input and improve it's readability. Such a tool would probably be of high usability among programmers.

A current restriction of the dataset is that it only works for java code. Another proposal for future work is therefore to overcome this restriction by extending the tool for other languages. This is not trivial as one has to adjust the readability decreasing heuristics to work with a different language. Furthermore a general tool that works for all languages will be very hard, if possible at all.

In order to further improve the readability estimations for both, the well and the badly readable code, one could conduct more s. By separating the code snippets into more fine granular clusters and by getting a more accurate average score by asking more persons about readability of code within the same cluster one can increase the accuracy of label estimation. Such an improvement can then again improve the model predictions as it is learning from this data.

As XY suggested another useful representation for code readability studies is the syntax tree representation of code. One could improve the performance of this model by adding another representation encoding extractor for java code that automatically extracts the abstract syntax tree of code.

A crucial aspect of code readability is naming. For the scope of methods, the most crucial part are methods names. Therefore one could improve this tool by adding a component that explicitly considers how well a method name fits its body.

Further research could also be to come up with another encoding that represents code in a different way.

Another way to improve existing code readability classifiers could be to come up with a different structure for some layers or entire different models.

The heuristics described in this work is only a part of the possible heuristics one could come up with. One could come up with more heuristics and evaluate them using user studies in order to further improve the variety of badly readable code. This might increase the number of internal features the model could learn which might again increase the tools accuracy.
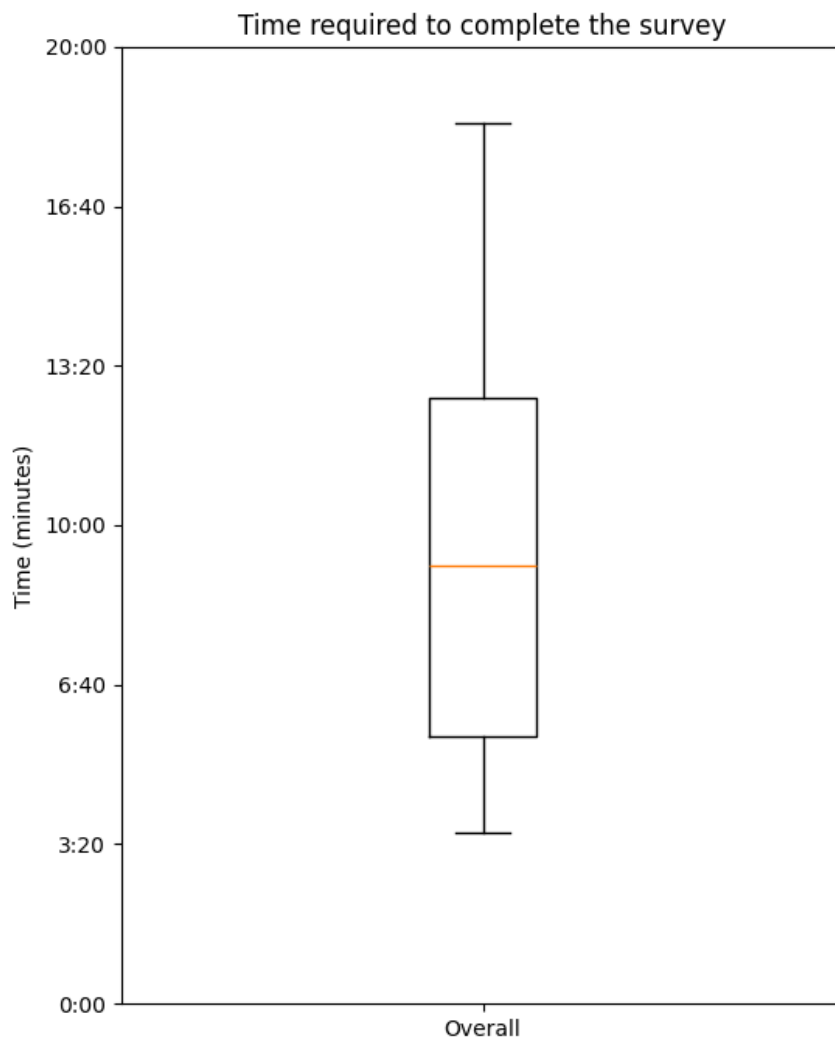
Figure 8: TODO: Replace with actual one.
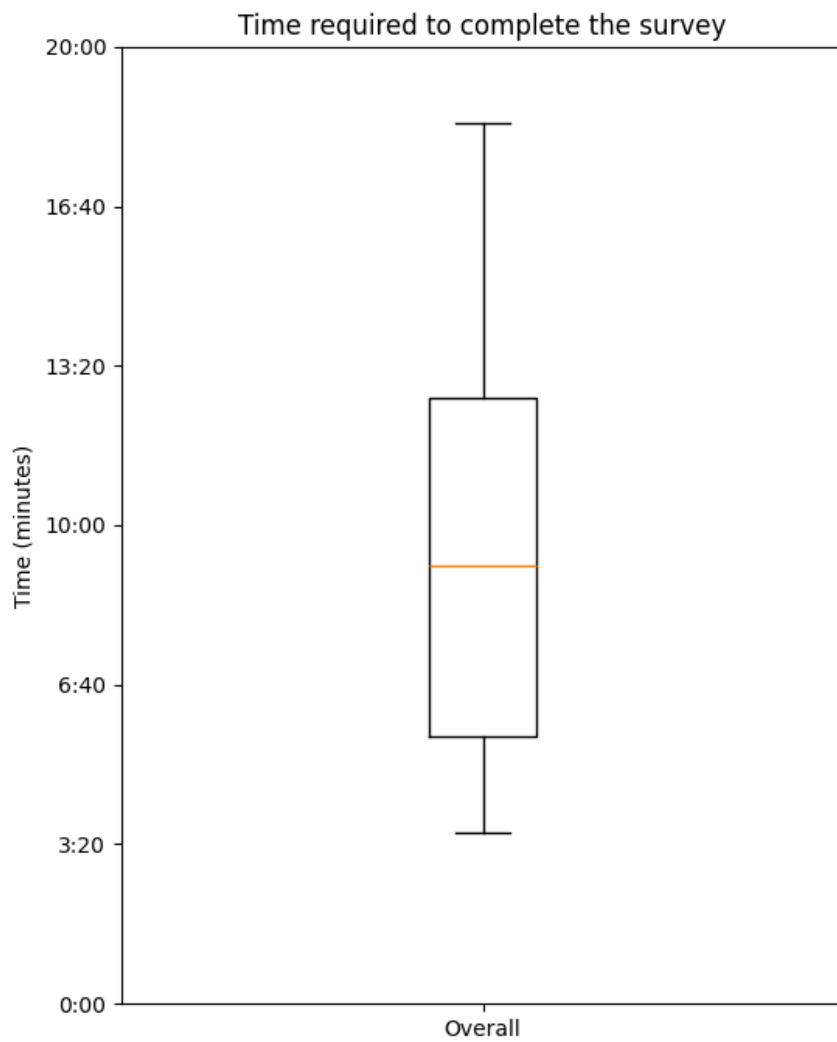
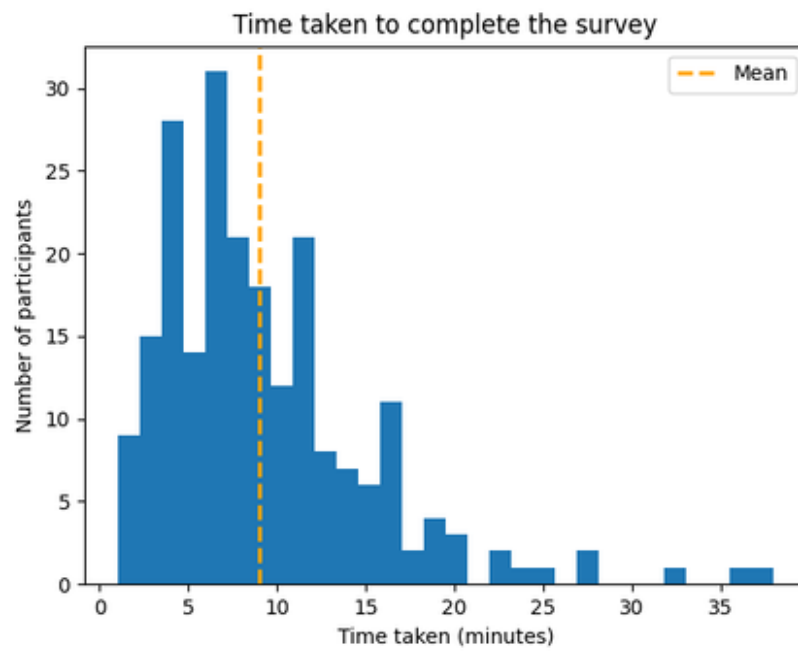Figure 9: Time required by the participants to complete the survey

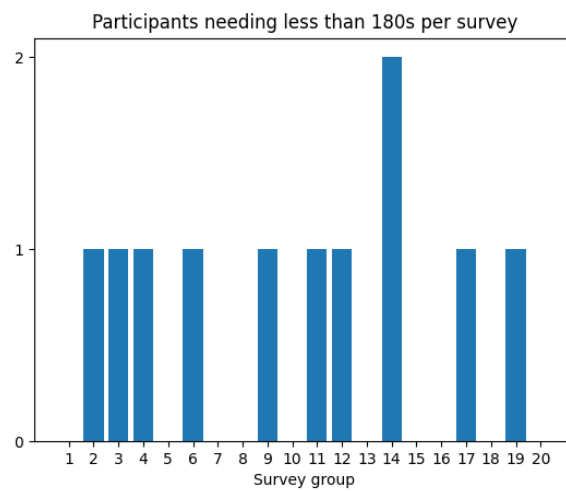Figure 10: Time required by the participants to complete the survey

Figure 11: Amount of participants per survey sheet that required less than 3 minutes to complete the survey
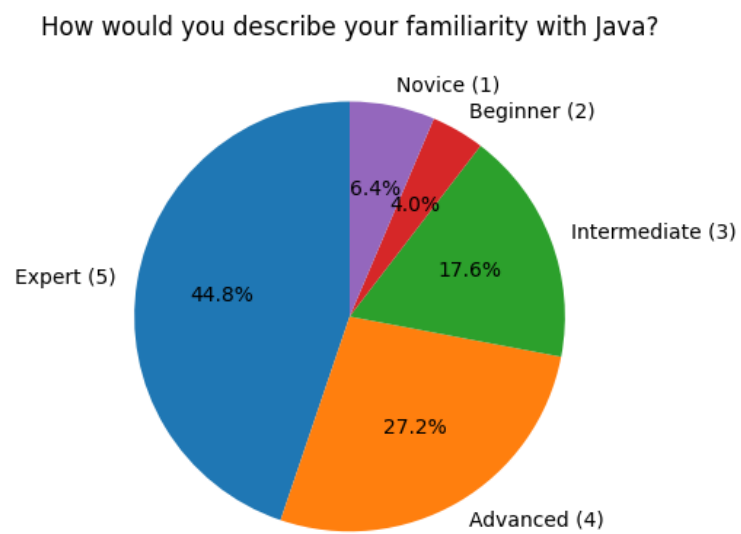
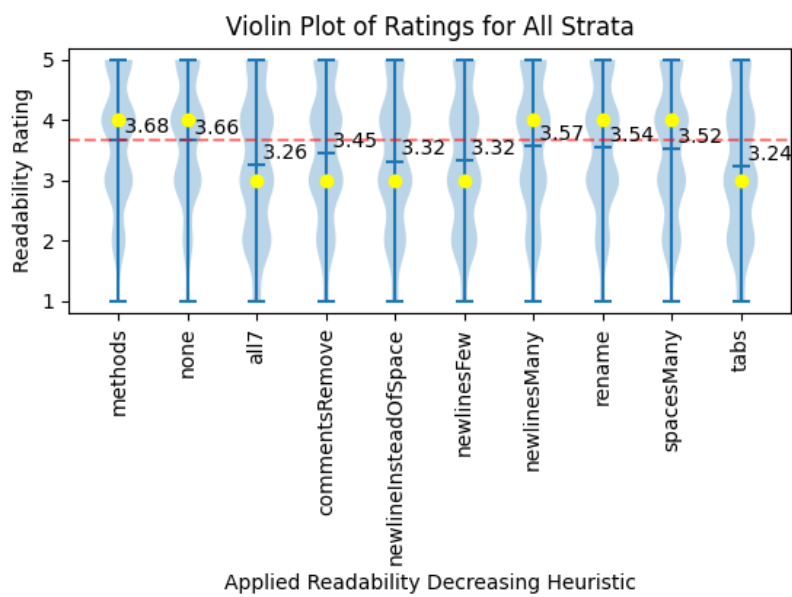Figure 12: Familiarity of survey participants with Java

Figure 13: Violin plot of the survey ratings for each RDH and all stratas

Bibliography

[1]   Krishan K Aggarwal, Yogesh Singh and Jitender Kumar Chhabra. 'An integrated measure of software maintainability'. In: *Annual Reliability and Maintainability Symposium. 2002 Proceedings (Cat. No. 02CH37318).* IEEE. 2002, pp. 235–241.

[2]   Duaa Alawad et al. 'An empirical study of the relationships between code readability and software complexity'. In: *ArXiv preprint* abs/1909.01760 (2019). URL: https://arxiv.org/abs/1909.01760.

[3]   Miltiadis Allamanis, Hao Peng and Charles Sutton. 'A Convolutional Attention Network for Extreme Summarization of Source Code'. In: *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016.* Ed. by Maria-Florina Balcan and Kilian Q. Weinberger. Vol. 48. JMLR Workshop and Conference Proceedings. JMLR.org, 2016, pp. 2091–2100. URL: http://proceedings.mlr.press/v48/allamanis16.html.

[4]   Uri Alon et al. 'code2vec: Learning distributed representations of code'. In: *Proceedings of the ACM on Programming Languages* 3.POPL (2019), pp. 1–29.

[5]   Kent Beck. *Implementation patterns.* Pearson Education, 2007.

[6]   Barry Boehm and Victor R Basili. 'Defect reduction top 10 list'. In: *Computer* 34.1 (2001), pp. 135–137.

[7]   Frederick Brooks and H Kugler. *No silver bullet.* April, 1987.

[8]   Raymond PL Buse and Westley R Weimer. 'A metric for software readability'. In: *Proceedings of the 2008 international symposium on Software testing and analysis.* 2008, pp. 121–130.

[9]   Raymond PL Buse and Westley R Weimer. 'Learning a metric for code readability'. In: *IEEE Transactions on software engineering* 36.4 (2009), pp. 546–558.

[10]   Davide Chicco and Giuseppe Jurman. 'The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation'. In: *BMC genomics* 21.1 (2020), pp. 1–13.

[11]   Sangchul Choi, Sooyong Park et al. 'Metric and tool support for instant feedback of source code readability'. In: *Tehnički vjesnik* 27.1 (2020), pp. 221–228.

[12]   Ermira Daka et al. 'Modeling readability to improve unit tests'. In: *Proceedings of the 2015 10th Joint Meeting on Foundations of Software Engineering*. 2015, pp. 107–118.

[13]   Lionel E Deimel Jr. 'The uses of program reading'. In: *ACM SIGCSE Bulletin* 17.2 (1985), pp. 5–14.

[14]   Jonathan Dorn. 'A General Software Readability Model'. In: 2012.

[15]   Sarah Fakhoury et al. 'Improving source code readability: Theory and practice'. In: *2019 IEEE/ACM 27th International Conference on Program Comprehension (ICPC)*. IEEE. 2019, pp. 2–12.

[16]   Joel Hestness et al. 'Deep learning scaling is predictable, empirically'. In: *ArXiv preprint* abs/1712.00409 (2017). URL: https://arxiv.org/abs/1712.00409.

[17]   John Johnson et al. 'An empirical study assessing source code readability in comprehension'. In: *2019 IEEE International conference on software maintenance and evolution (ICSME)*. IEEE. 2019, pp. 513–523.

[18]   Rensis Likert. 'A technique for the measurement of attitudes.' In: *Archives of psychology* (1932).

[19]   Kui Liu et al. 'Learning to spot and refactor inconsistent method names'. In: *2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE)*. IEEE. 2019, pp. 1–12.

[20]   Benjamin Loriot, Fernanda Madeiral and Martin Monperrus. 'Styler: learning formatting conventions to repair Checkstyle violations'. In: *Empirical Software Engineering* 27.6 (2022), p. 149.

[21]   Umme Ayda Mannan, Iftekhar Ahmed and Anita Sarma. 'Towards understanding code readability and its impact on design quality'. In: *Proceedings of the 4th ACM SIGSOFT International Workshop on NLP for Software Engineering*. 2018, pp. 18–21.

[22]   Robert C Martin. *Clean code: a handbook of agile software craftsmanship*. Pearson Education, 2009.

[23]   Qing Mi. 'Rank Learning-Based Code Readability Assessment with Siamese Neural Networks'. In: *Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering*. 2022, pp. 1–2.

[24] Qing Mi et al. 'A graph-based code representation method to improve code readability classification'. In: *Empirical Software Engineering* 28.4 (2023), p. 87.

[25] Qing Mi et al. 'An enhanced data augmentation approach to support multi-class code readability classification'. In: *International conference on software engineering and knowledge engineering*. 2022.

[26] Qing Mi et al. 'An inception architecture-based model for improving code readability classification'. In: *Proceedings of the 22nd International Conference on Evaluation and Assessment in Software Engineering 2018*. 2018, pp. 139–144.

[27] Qing Mi et al. 'Improving code readability classification using convolutional neural networks'. In: *Information and Software Technology* 104 (2018), pp. 60–71.

[28] Qing Mi et al. 'The effectiveness of data augmentation in code readability classification'. In: *Information and Software Technology* 129 (2021), p. 106378.

[29] Qing Mi et al. 'Towards using visual, semantic and structural features to improve code readability classification'. In: *Journal of Systems and Software* 193 (2022), p. 111454.

[30] Qing Mi et al. 'What makes a readable code? A causal analysis method'. In: *Software: Practice and Experience* 53.6 (2023), pp. 1391–1409.

[31] Delano Oliveira et al. 'Evaluating code readability and legibility: An examination of human-centric studies'. In: *2020 IEEE International Conference on Software Maintenance and Evolution (ICSME)*. IEEE. 2020, pp. 348–359.

[32] Renaud Pawlak et al. 'Spoon: A library for implementing analyses and transformations of java source code'. In: *Software: Practice and Experience* 46.9 (2016), pp. 1155–1179.

[33] Daryl Posnett, Abram Hindle and Premkumar Devanbu. 'A simpler model of software readability'. In: *Proceedings of the 8th working conference on mining software repositories*. 2011, pp. 73–82.

[34] Talita Vieira Ribeiro and Guilherme Horta Travassos. 'Attributes influencing the reading and comprehension of source code–discussing contradictory evidence'. In: *CLEI Electronic Journal* 21.1 (2018), pp. 5–1.

[35] Spencer Rugaber. 'The use of domain knowledge in program understanding'. In: *Annals of Software Engineering* 9.1-4 (2000), pp. 143–192.

[36] Simone Scalabrino et al. 'A comprehensive model for code readability'. In: *Journal of Software: Evolution and Process* 30.6 (2018), e1958.

[37] Simone Scalabrino et al. 'Automatically assessing code understandability: How far are we?' In: *2017 32nd IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE. 2017, pp. 417–427.

[38] Todd Sedano. 'Code readability testing, an empirical study'. In: *2016 IEEE 29th International conference on software engineering education and training (CSEET)*. IEEE. 2016, pp. 111–117.

[39] Milan Segedinac et al. 'Assessing code readability in Python programming courses using eye-tracking'. In: *Computer Applications in Engineering Education* 32.1 (2024), e22685.

[40] Shashank Sharma and Sumit Srivastava. 'EGAN: An Effective Code Readability Classification using Ensemble Generative Adversarial Networks'. In: *2020 International Conference on Computation, Automation and Knowledge Management (ICCAKM)*. IEEE. 2020, pp. 312–316.

[41] Yahya Tashtoush et al. 'Impact of programming features on code readability'. In: (2013).

[42] Antonio Vitale et al. 'Using Deep Learning to Automatically Improve Code Readability'. In: *2023 38th IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE. 2023, pp. 573–584.

[43] Greg Wilson and Andy Oram. *Beautiful code: Leading programmers explain how they think*. " O'Reilly Media, Inc.", 2007.

[44] Xin Xia et al. 'Measuring program comprehension: A large-scale field study with professionals'. In: *IEEE Transactions on Software Engineering* 44.10 (2017), pp. 951–976.

[45] Michihiro Yasunaga and Percy Liang. 'Graph-based, Self-Supervised Program Repair from Diagnostic Feedback'. In: *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*. Vol. 119. Proceedings of Machine Learning Research. PMLR, 2020, pp. 10799–10808. URL: http://proceedings.mlr.press/v119/yasunaga20a.html.