xAILAB
EXPLAINABLE MACHINE LEARNING

Deep Learning based Evaluation of Handwriting Legibility using a Sensor Enhanced Ballpoint Pen

Master Thesis

Master of Science in Applied Computer Science

Erik Jonathan Schmidt

February 28, 2025

Supervisor:

1st: Prof. Dr. Christian Ledig

Chair of Explainable Machine Learning
Faculty of Information Systems and Applied Computer Sciences
Otto-Friedrich-University Bamberg

Abstract

This work presents an exploration of how machine learning models can be used to determine legibility ratings for handwriting samples from sensor data, which was recorded using the STABILO DigiPen. The new *StabLe* dataset consists of samples written with this pen and was annotated with descriptive meta data and legibility ratings. This revealed that perceived legibility is correlated with characteristics of the handwriting samples such as being written in cursive or print letters. The performance reported for models, which were trained in related work to determine handwriting legibility from movement sensor data, was shown to be compromised by the design of training and evaluation. The agreement between models and individual raters was suggested as a meaningful evaluation considering the subjectivity of the ratings. Different ways of mapping varying ratings per sample to a single consensus label were examined, as well as training a rater-specific model. In general, trained models overfitted the training data and achieved low agreement with raters on unseen samples. The best-performing model was shown to depend mainly on discriminating between cursive and print-letter writing styles. Failing to train models that accurately determine legibility ratings, this work highlights the challenges of using machine learning methods for an automated assessment of legibility based on time-series sensor data.

Acknowledgements

I am deeply grateful to Prof. Dr. Christian Ledig for his dedicated supervision and constructive feedback, which guided me through the research process.

I would like to express my sincere gratitude to Dr. Jens Barth, Dipl. Ing. Peter Kämpf, and M. Sc. Tim Hamann at STABILO international GmbH for their invaluable support and insightful feedback throughout working on and writing this thesis.

A special thank you to Aaron Lukas Pieger for the excellent teamwork - we collaborated closely on parts of this work, and his contributions were essential to its success.

I also want to acknowledge Francesco Di Salvo for the thoughtful advice on academic writing, which significantly improved the clarity and presentation of this thesis.

Finally, I extend my thanks to PD Dr. Tal Hoffmann and Dipl. OT Susanne Salata at Schreibmotorik Institut e.V. for their guidance regarding the scientific background on legibility, which strengthened the foundation of this research.

Contents

List of Figures

List of Tables

1 Introduction

This work contributes to the long-term goal of automating the assessment of handwriting legibility. In collaboration with STABILO International GmbH, the Chair of Explainable Artificial Intelligence at Otto-Friedrich-Universität Bamberg, and the Schreibmotorik Institut e.V., the STABILO Legibility dataset (*StabLe*) was created. It contains subjective ratings of four criteria of legibility for handwriting samples that were recorded with a sensor-enhanced ballpoint pen. The dataset was used to analyze rater agreement and to train convolutional neural networks to predict legibility labels based on the movement data of the pen.

1.1 Background and Motivation

The Role of Handwriting Legibility and its Assessment Studies on handwriting assessment have examined its significance in education and the relations between legibility and various learning outcomes. A detailed review of these studies is provided in Section 2.1. In summary, handwriting on paper constitutes a significant part of daily school life, and the ability to produce legible writing serves as an indicator of pupils' learning achievements. In addition, poor handwriting quality has been linked to conditions such as dyslexia and attention deficit hyperactivity disorder. Research on legibility assessment has led to the development of standardized handwriting scales designed to evaluate handwriting quality - particularly legibility - as reliably as possible (Section 2.1.4). However, the applicability of these approaches is hindered by two main limitations.

- **Subjectivity** Handwriting legibility is primarily assessed by teachers based on personal experience and opinion, leading to inconsistencies both between and within evaluations of individual raters. Although standardized assessment methods can reduce this variability, it cannot be completely eliminated. Firstly, legibility is inherently subjective, as it depends on the reader's perception. Secondly, differences in rater training and the application of rating criteria contribute to persistent variability in legibility assessment.
- **Resources** The evaluation of handwriting requires considerable time and effort, as each student's writing must be individually assessed. Although handwriting scales have been developed to improve consistency, their reliability depends on proper rater training to ensure uniform application, which demands significant resources. Even with standardized tools, ensuring accurate and unbiased handwriting assessment remains challenging due to practical constraints in educational settings.

Au et al. (2012) summarize that a reliable instrument for measuring changes in handwriting ability over time remains elusive. They argue that, in the meantime, individual clinicians can use handwriting scales to diagnose and rate legibility. Barnett et al. (2018) describe the need for and the status quo of handwriting assessment:

"Demands for the production of legible handwriting produced in a timely manner increase as children progress through school. Despite the considerable number of children faced with handwriting difficulties, there is no quick and practical tool to assess legibility in this population."

The Computerized Assessment of Handwriting Legibility Previous research has explored various approaches to automating handwriting assessment. A detailed review of the literature is given in Section 2.2. The approaches focus either on the writing process (fluency, speed, and effort) or on the writing product (letter shapes and forms). Legibility is assessed either on the character level or for longer passages of handwriting. Tablet-based solutions provide rich spatial and temporal handwriting data, but introduce an unnatural writing environment. Sensor-enhanced pens, on the other hand, preserve a natural writing experience and do not require a complex writing setup, but lack the detailed spatial information which tablets capture. Despite advances in the use of deep learning models for handwriting assessment, existing approaches face several limitations:

- **User-Dependent Evaluation** Models were evaluated on user-dependent data, which means that they are evaluated on handwriting from known students. How the model performs when assessing the writing of unknown students cannot be said, but defines their applicability in real world educational settings.
- **Unbalanced Test Data** Studies used evaluation metrics that do not adequately reflect the ability of models to distinguish different levels of legibility, particularly when data was unbalanced.
- **Inappropriate Labeling** Labeling strategies in previous work introduced inconsistencies, as legibility annotations were sometimes assigned at the sample level based on broader assessments or ambiguous legibility criteria, which potentially obscured fine-grained variations of legibility.

While high accuracy has been reported for some models, closer analysis reveals that these results are affected by the mentioned limitations, and often stem from dataset biases rather than genuine advances in predicting ratings of legibility.

Motivation for a Pen-Based Solution The main arguments for an automated assessment based on pen sensor data were the internal consistency that a solution based on machine learning could provide and the time-efficient application it promises. The use of a sensor-enhanced ballpoint pen to record input data allows seamless integration into students' daily routines and could enable real-time legibility assessment with immediate feedback. Together, the pen and deep learning models could help identify children with learning difficulties and monitor learning progress as children progress in their school careers. In the future, learning applications could support handwriting practice without requiring direct feedback from

parents, teachers, or therapists, thus saving them time. This feedback could either point out areas of improvement, like paying more attention to spacing between letters, or encourage students by recognizing progress already made. As part of an automated legibility advisor, the automated assessment could further reduce the need for repetitive feedback on basic legibility criteria, allowing human advisors to focus on students' individual needs.

1.2 Problem Statement

Different approaches have been proposed to use deep learning for the assessment of legibility (Section 2.2). As motivated above, an assessment based on data from a sensor-enhanced ballpoint pen is desirable.

CNNs were used for such an assessment before, and the low accuracy reported by Grabmann (2023) for user-independet models trained on balanced data presents the status quo of training CNNs to determine legibility ratings from handwriting sensor data. The expressiveness of this accuracy is diminished by the highly unbalanced test set on which it was obtained. Therefore, it remains unclear how the CNN would perform when applied for automated legibility assessment in practice.

Legibility is understood as a quality of the writing product (Section 2.1.1). The sensor data collected with the DigiPen, an electronic pen developed by STABILO International GmbH, captures the writing movement and pressure, which describe the writing process rather than the product itself. How well features of the writing product can be derived from this data is a subject of ongoing research. Consequently, a careful evaluation is needed to examine whether models actually learn to detect features of legibility in the sensor data (e.g. how consistent the slant of letters is) or if they detect features of the process that correlate with legibility (e.g. the length of pauses during writing).

Legibility itself is a loose concept that is understood and defined differently from person to person. In view of supervised machine learning, which requires a single numeric ground-truth label per sample, this subjectivity presents a challenge. Previous work did not account for this subjectivity. How the variance of legibility ratings can be modeled and how models can be evaluated given the ambiguous ratings require further investigation.

In previous work, legibility was mostly assessed as a whole. To improve the explainability of predictions and to enable specific feedback for improving ones handwriting, a separate assessment of different criteria of legibility is needed.

In summary, imbalances and subjectivity in the data make training and evaluation of models that predict legibility ratings challenging. Approaches to derive ground-truth labels from varying ratings need to be investigated. Whether performances can actually be attributed to models learning features of legibility itself was not examined before. A meaningful way to evaluate models against varying ratings is needed, and a compartmentalized assessment could improve explainability and enable helpful feedback.

1.3 Research Questions

This work investigates the applicability of automated handwriting legibility assessment using time-series data recorded with a sensor-enhanced ballpoint pen. The six main research questions are outlined below.

R1 *Is the rater agreement found in the StabLe dataset comparable to the agreement reported in related research?*

The assembled dataset contained ratings of four legibility criteria. Similar criteria were addressed in handwriting scales used in related research. The rater agreement on the dataset was evaluated to show how reliable the collected ratings were compared to the assessments with the reviewed scales. The measured variance and reliability served as a baseline for model predictions.

R2 *Is the rater agreement higher on the criteria that are assumed to be more specific?*

Compared to previous work, ratings were collected for different criteria of legibility that were assumed to exhibit different levels of subjectivity, depending on how directly they assess specific characteristics of the writing product. Rating the overall legibility of a handwriting sample was assumed to be highly subjective, while an inquiry about the form of specific letters in the writing was assumed to be more objective. To test this assumption, rater agreement was examined on the different criteria.

R3 *Can the results of previous work on automated legibility assessment be reproduced and are the results meaningful?*

The findings of previous work by Grabmann (2023) were reproduced to consolidate the status quo. Here, it was tested whether similar results were obtained when conducting similar experiments and whether the reported results represent a meaningful depiction of how close the existing solutions are to being applied in practice.

R4 *Does reducing the variety of texts in the dataset help the models to find patterns related to legibility?*

It was hypothesized that reducing the variety of texts, which were recorded to build a dataset for legibility assessment, makes the prediction task easier. Having a smaller set of texts means that the sensor data of samples should be more similar because the same letters and words were written. Therefore, differences in the sensor data seemed more likely to be caused by differences in legibility. To check whether this hypothesis holds, models were trained on both data from previous work, which comprised many different texts, and on data collected in this work, which comprised handwriting samples of a small set of reference sentences.

R5 *How can the uncertainty inherent in assessing legibility be addressed when training supervised models?*

A key problem of training models for legibility assessment lies in the subjectivity of ratings. Assessments of the same sample can differ between raters. This uncertainty is inherent in the task of assessing handwriting quality. To explore how this uncertainty can be modeled, different strategies were tested to deduce ground-truth labels from ratings.

R6 *Do models perform better in assessing criteria that are assumed to be more specific?*

The criteria that were assumed to be more objective (R2) were also assumed to be closer related to patterns in the sensor data. The overall perceived legibility is a compound of many factors and, therefore, it seems unlikely that it is correlated to specific patterns in the sensor data. On the other hand, the slant of the letters and the appropriate length of strokes in specific letters were assumed to be deducible from patterns in the sensor data. Writing with consistent slant should result in a regular combination of accelerations measured along the different spatial axes. Writing a letter too short should show in smaller amplitudes compared to writing the letter with a longer stroke. To test this assumption, similar model architectures were trained to predict the different criteria and then compared with respect to their performance.

1.4 Outline

Next, Section 2 reviews the literature on legibility and its assessment. Methods for creating and analyzing the *StabLe* dataset and experiments with CNNs trained on this dataset are described in Section 3. Technical details on the preprocessing of sensor data, the developed annotation tool, and the training of models are given in Section 4. Section 5 provides a statistical analysis of the dataset and the evaluations of the trained models. In Section 6 the results are summarized and interpreted with respect to the stated research questions. In Section 7 observations and thoughts are presented to point out limitations of this work. Finally, Section 8 proposes approaches to handle limitations in future work.

2 Literature Review

2.1 Legibility and its Assessment

2.1.1 Definitions of Legible Handwriting

The goal of this work is to examine whether the legibility of handwriting can be assessed computationally. For this purpose, it is mandatory to specify what is generally meant by legible handwriting. There is no universal definition of what makes handwriting legible; instead, an overview of how it is understood in the literature helps to distinguish legibility from other qualities of handwriting. Rüb (2018) defines legibility as a subset of readability. **Legibility** is determined by the geometric shapes and strokes the writer produced, independent of the meaning that this writing is meant to transport. **Readability** indicates how understandable the text produced is. Therefore, readability comprises legibility, syntax and semantic meaning of the text. Feder and Majnemer (2007) describe legibility as a compound of letter formation, spacing, size, slant, and alignment. These criteria affect the ease with which individual letters can be identified. The authors describe legibility as one of the two main factors of handwriting performance, next to writing speed or fluency. For Rosenblum et al. (2003) readability and legibility are qualities of the **handwriting product**, that is, the shapes and strokes visible on the paper. The authors distinguish between the product and the **handwriting process**, which is the act of writing itself. Together, they determine the **handwriting performance**, that is, how well someone writes overall (Figure 1).

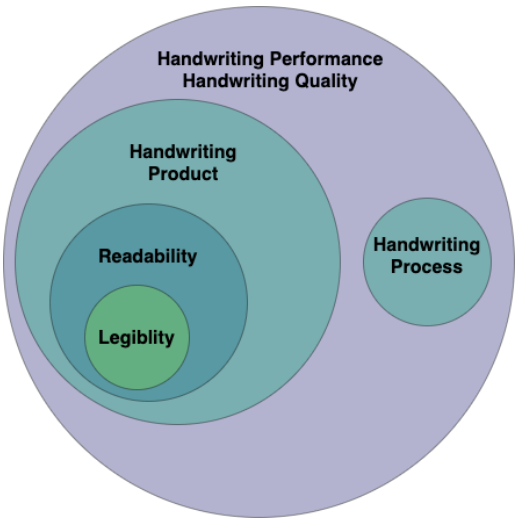


Figure 1: The relation between different terms related to the quality of handwriting.

Stefansson and Karlsdottir (2004) view handwriting as a means of communication. If the writing is not legible to the reader, then the communication between the writer and the reader is negatively affected. The authors state that handwritten text must adhere to "a sufficiently widely accepted standard, specifying the shapes, sizes, and positions of the letters" in order to be perceived as legible by most readers. Following the differentiation of legibility and readability, readable handwriting means that communication between writer and reader works flawlessly. The legibility is then a necessary but not sufficient requirement for flawless communication via handwritten text. Viewing legibility as a requirement for successful communication points out the highly subjective nature that is inherent in assessing legibility. In the end, legibility is perceived by the reader and therefore depends on the reader.

This is reflected in the varying inter-rater reliability found for different standardized assessments, so called handwriting scales (Table 3). The assessor or rater introduces a personal and subjective understanding of what legibility is in the assessment. Furthermore, different assessments address legibility by asking different questions, thus capturing different notions of how the concept of legibility is understood. This is in tune with Harris and Association (1960) who state that terms like legibility and readability "resist precise definition and appear to be complexes with whole and part attributes which may exist in many different combinations in handwriting specimens".

Consequently, legibility should not be viewed as a singular characteristic but rather as a composite of multiple criteria that collectively shape its perception. If all the criteria are met for some handwritten text, then the average reader should be able to read the presented letter sequence without complications. Rosenblum et al. (2003) give an overview of different handwriting scales developed in the past. They find that scales which capture legibility as one holistic feature of handwritten text were more common in the early days of this field of research. In more recent research, the focus has shifted towards analytical scales which calculate legibility from a set of more specific criteria, such as letter size or spacing.

This review of the term legibility informed how it was understood in this work. Legibility is the perceived quality of handwritten text that is determined by the criteria listed in Table 1. In addition to syntax and semantics, legibility is a requirement that must be met for handwriting to be readable. Legibility is subjective because it represents the quality of text that is perceived by the reader. Consequently, there is no hard truth about the legibility of a given piece of handwriting. This implies limitations for any assessment that is meant to measure legibility.

Table 1: Criteria belonging to legibility based on the reviewed literature.

Criteria of Legibility	Other Qualities of Handwriting
shapes and strokes of individual letters, consistency of letter slant, consistent size of letters, adherence to line, appropriate spacing in and between words	spelling, grammar, semantics, writing effort, writing speed, page layout

2.1.2 Correlations with Legibility and its Implications

This section reviews the literature that examines the role of handwriting legibility in school life, the correlations that come with poorly or sufficiently legible handwriting, and its diagnostic significance.

Benefits of Legible Handwriting A field study by McHale and Cermak (1992) examined the fine motor skills required and exercised in American elementary school

classes. This study found that children spent 31% to 60% of their school day performing handwriting tasks. This supports Feder and Majnemer (2007) who consider the development of handwriting ability to be an important requirement for success in school. Steve Graham (2020) states that writing is essential not only for school, but also for work and home life.

Further studies underscore the positive effects of acquiring sufficiently developed handwriting skills.

McCarney et al. (2013) had primary school students participate in cognitive, literacy, and writing tests. Through a group analysis, the authors found that the ability to produce readable handwriting was weakly correlated with performance in a working memory test, a verbal IQ test, and word reading and spelling tests.

Dinehart and Manfra (2013) measured the fine motor skills of preschool children. Preschoolers were tested on writing and object manipulation tasks. As they reached second grade, the children performed a variety of cognitive, reading, and mathematical tests. Modest but statistically significant correlations were observed between these early motor skills and the achievements in the later tests. The authors conclude that fine motor skills, particularly the ability to produce well-formed letters, can indicate the school readiness of a child.

The effect of writing well-formed letters is further examined in neurological studies. James and Engelhardt (2012) measured the brain activity of preliterate children while letters were presented to them. They compared the brain activation of children who participated in free-form handwriting exercises, children who performed tracing exercises, and children who typed letters on a keyboard. The authors documented that a "reading circuit" in the brain was activated only when they had performed free-form handwriting before being presented with the letter. This led to their conclusion that handwriting helps develop the skill to process letters because producing a well-formed letter on blank paper demands more attention to detail than just tracing or typing it, thus forcing the child to build an understanding of what makes up a certain letter.

Berninger and Richards (2002) come to a similar conclusion. They state that handwriting is a vehicle for children to develop patience and discipline because learning to write a letter demands the student to maintain sustained focus and acquire fine motor skills. Furthermore, connecting letters with the motor function of writing this letter is believed to foster better memory retention and attention to detail.

Szymczak (2016) examined handwritten samples that were collected in a translation competition. Each sample was scored by a jury to determine a ranking with regard to translation quality. Subsequently, independent raters assessed the legibility of those samples. A significant correlation was found between the legibility ratings and the ranking in the translation quality competition. It is hypothesized that psychological effects caused the jury to associate better translation quality with legible handwriting samples. If this assumption holds, then this suggests that legibility is taken into account subconsciously whenever handwritten text is assessed or graded by humans.

In summary, legible handwriting and the fine motor skills needed to produce such writing are shown to play an important role in education. If a student writes legibly, then this indicates that his or her perception of letters is connected with the skill of persistently writing that letter. This connection is believed to decrease the workload of all tasks related to writing or reading. Although the direction of causality or the interaction with other latent variables is unknown, the literature has shown that the ability to write legibly and precisely is at least weakly correlated with academic success and can benefit other areas of life.

Problems with Illegible Handwriting In contrast, the inability to produce legible handwriting can be an indicator of learning difficulties and has been shown to have undesirable implications.

Berninger and May (2011) summarize the literature on learning disabilities and found that impaired legible automatic letter writing is a common denominator in research on dysgraphia (diagnosed writing difficulties). Automatic writing refers to the ability to write fluently with minimal cognitive effort. Legible automatic letter writing is a compound of legible writing and automatic writing. So diagnosing for dysgraphia involves an assessment of legibility.

Similarly, Martlewm (1992) examined children with and without dyslexia (diagnosed reading difficulties). They compared qualities of the writing process and product. The handwriting of children with dyslexia was perceived as less legible.

Comparing both the writing process and the product of children with and without attention-deficit hyperactivity disorder (ADHD), Sara Rosenblum and Josman (2008) found that those with ADHD exhibit a comparatively poor spatial arrangement of strokes and letters, as well as a higher frequency of unrecognizable letters. This suggests that low legibility could be an indicator of this condition.

Dinehart (2014) and Barnett et al. (2018) summarize that students who failed to acquire the skill of legible handwriting can develop a reluctance to write, which in turn is detrimental to their success in school and affects self-esteem.

2.1.3 Informal Assessments in Practice

The ability to write legibly is a valuable skill. This motivates the intention to assess the legibility of handwriting. According to a teacher survey conducted by Marquardt et al. (2016), 31% of girls and 51% of boys in German schools exhibit some degree of handwriting difficulty. A meta-study on handwriting difficulties and interventions by Feder and Majnemer (2007) found that between 10% and 30% of school-aged children are affected. They observed that these difficulties do not resolve without intervention in most cases, while interventions were shown to improve handwriting, independent of the specific treatment approach. An assessment of legibility as part of handwriting performance as a whole could help to identify and target such difficulties, and could positively impact a significant percentage of schoolchildren.

Stefansson and Karlsdottir (2004) reviewed three studies that investigated how handwriting is evaluated in school. They found that the assessment of handwriting quality is usually based on informal observations by the teacher. Formal handwriting scales are used less frequently.

Rondinella (1962) has shown that individual observations, which teachers base their assessments on, are not reliable because they are determined by the individual standards of the teacher, rather than objective or standardized benchmarks. The individual assessments of teachers were compared to the ratings obtained with a formal handwriting scale, which revealed a wide spread in perceived handwriting quality.

2.1.4 Formal Assessments using Handwriting Scales

A handwriting scale is a standardized and quantifiable measurement of a quality of handwriting. Most scales rely on a subjective assessment conducted by an individual rater, while some aspects of handwriting quality, such as writing speed, can be objectively measured. As discussed above, legibility is a quality perceived by the reader. Handwriting scales related to legibility mainly use questionnaires and are subjective. Rosenblum et al. (2003) reviewed existing handwriting scales and their application to detect handwriting difficulties. She categorized them as global-holistic or analytical handwriting scales.

A global-holistic handwriting scale directly associates a single rating with the quality inquired for the handwriting sample.

An analytical handwriting scale evaluates the quality of handwriting as a complex compound. Samples are rated with respect to different criteria that are assumed to affect quality. An overall score is then calculated from those more specific ratings. The author found that analytical scales are more common in recent research. This can be attributed both to the higher reliability found in analytical scales and to the more detailed insights they provide. Analytical scales can point out which factors make some handwriting illegible, helping to initiate appropriate measures.

2.1.5 Renowned Handwriting Scales

In the following, five analytical handwriting scales are described, which assess the legibility through a questionnaire. For a detailed description of the scales and the corresponding questionnaire items, refer to A.1.

SEMS and (SOS-2) The German Systematische Erfassung motorischer Schreibstörungen (SEMS) is a handwriting scale developed to identify children with handwriting difficulties, which was adopted from the Dutch SOS-2 (Waelvelde et al., 2012). Suspects perform a copy writing task. The scale evaluates both the legibility of the produced writing as well as writing speed. With the corresponding questionnaire, seven criteria related to legibility are rated as being satisfied mostly (0), sometimes (1), or rarely (2). The ratings are summed to obtain a total score.

Franken and Harris (2021) found that this total score can be used to accurately identify children with handwriting problems in the second grade.

HLS Barnett et al. (2018) developed the Handwriting Legibility Scale (HLS) for teachers to quickly assess legibility. The suspect performs a free writing task. A legibility score is calculated as the sum of five criteria that are rated on a five-point Lickert scale following the corresponding questionnaire.

ETCH The Evaluation Tool of Children's Handwriting (ETCH) tests for many aspects of handwriting performance. The suspects participate in several writing tasks. The examiner observes the suspect during the writing tasks because both the writing process and the writing product are assessed. Duff and Goyen (2010) describe it as a criterion-referenced assessment that focuses on the readability of letters, words, and numbers at a glance and out of context. The corresponding examiner's manual by Amundson (2004) provides detailed instructions on the preparation, execution, and interpretation of the proposed assessment. This handwriting scale was designed for use by occupational therapists.

HPSQ Rosenblum (2008) proposed the Handwriting Proficiency Screening Questionnaire (HPSQ) to identify handwriting difficulties among school-age children. It is intended to be used by teachers to assess the handwriting of their students. The questionnaire contains ten questions that are rated from one to five. A principal component factor analysis of these ten criteria revealed two main factors. The first comprises four questions, which the authors summarized as related to legibility.

MHT Following the description by Rosenblum et al. (2003), the Minnesota Handwriting Test (MHT) was developed to assist occupational therapists in identifying school children with writing difficulties. Suspects copy a standardized set of words for a fixed period of time. The examiner checks which of fourteen statements apply to the writing of the suspect and then rates the produced writing according to six criteria, of which five are related to legibility.

To summarize, the described handwriting scales use questionnaires to quantify the judgment of legibility of the raters. The questionnaire items inquire ratings of different aspects of legibility. Although precise definitions differ, there seems to be general agreement on which legibility criteria need to be assessed on an analytical handwriting scale that focuses on legibility. Similar criteria are aggregated in Table 2 to give an overview of how legibility is captured by the described scales. The checkmarks indicate that a scale contains a questionnaire item that at least mentions the given criterion in the rating instruction.

Table 2: An overview of the legibility criteria inquired in handwriting scales.

Handwriting Scale	Global Legibility	Reading Effort	Page Layout	Letter Formation	Alteration Frequency	Consistent Size	Spacing	Line Alignment	Consistent Slant
	Criterion Inquired in Questionnaire								
HLS	✓	✓	✓	✓	✓	✓	✓	✓	✓
SEMS				✓		✓	✓	✓	
ETCH	✓			✓			✓	✓	✓
HPSQ		✓			✓				
MHT	✓			✓		✓	✓	✓	✓

2.1.6 Limitations of Handwriting Scales

Rosenblum et al. (2003) point out the limitations of existing handwriting scales. One drawback of questionnaire-based assessments is the dependence on the rater. For many handwriting scales, there is no specification on who is qualified to administer them. It seems plausible that the perception of legibility differs between therapists, teachers, and laymen. Consequently, the same handwriting might receive varying scores depending on the rater.

Furthermore, the scales differ in how the rater is instructed. The instructions range from short textual descriptions of the writing task and questionnaire items to extensive manuals that provide guidance on the preparation, administration, and interpretation of the assessment. Both Barnett et al. (2018) and Franken and Harris (2021) instructed the raters about the use of the handwriting scales personally. In contrast, the MHT comes with an instruction manual for the rater, and in-person instructions did not take place. Similarly, before administering an ETCH assessment, raters are expected to practice with the corresponding manual. A trial assessment is provided to check that the rater assigns the ratings according to the manual instructions.

These differences in the selection of raters and the provided instructions partially explain the variation of reported inter-rater reliabilities shown in Table 3. The metrics are described in Section 3.4.2. Barnett et al. (2018) instructed two teachers to assess the handwriting according to the HLS. Based on the ratings given, the handwriting samples were grouped into three classes of legibility (low, medium, and high). The ICC indicated excellent agreement, but the authors did not report which variant of the ICC was used. Cohen's Kappa was lower but showed substantial item agreement. Waelvelde et al. (2012) found good agreement on the total SOS-2 scores using ICC (2,1) and varying agreements using Cohen's Kappa agreement on the

different items rated. For the other scales, only the ICC was reported either for total scores or ratings of individual items.

Table 3: Inter-rater reliability of handwriting scales.

Scale	Intraclass Correlation Coefficient	Cohen's Kappa
HLS	.92	.67
SOS-2 (SEMS)	.77	0.39-0.77
ETCH	.85-.92	
HPSQ	.92	
MHT	.77-.99	

The described assessments are time-intensive. For example, the HPSQ can be conducted by a teacher in about five minutes and an assessment using the ETCH is expected to take 20 to 30 minutes. This investment of time per student hinders a wide adoption of handwriting scales for legibility assessment.

2.2 Computerized Assessments of Legibility

Different machine learning based solutions have been proposed before for automated handwriting assessment. They vary in the qualities of handwriting that they aim to assess and in the data they operate on.

- Solutions focus on either the writing process or the writing product (Section 2.1.1).
- Solutions use images of handwriting, trajectories recorded on tablets, or data from sensor-enhanced ballpoint pens.
- Solutions assess single letters or longer passages of writing.

2.2.1 Predicting the Legibility of Single Characters using Tablets

The use of display tablets to record handwriting trajectories provides both temporal and spatial descriptions of the writing, which makes them a popular choice. The recorded data allows for a direct assessment of both writing product and process. Hamdi et al. (2020) developed a system that assessed how well students wrote single letters on a tablet computer. From the recorded trajectories, three representations were calculated using a beta-elliptic model, Fourier transformation, and CNNs. These representations of the student's attempt at writing the given letter were then used to calculate the similarity with selected correctly written reference samples of the given letter. In addition, support vector machines were used to calculate scores that rated different aspects of the written letter. The similarity measurements and

scores were then forwarded to a fusion model that combined inputs to determine scores for the four criteria of interest.

1. Correctness of the overall shape of the letter.
2. Omission of required strokes of the letter.
3. Deformations like an unusual slant of a stroke in the letter.
4. Addition of strokes.

In summary, single letters written on a tablet are automatically rated with respect to four criteria.

2.2.2 Predicting Handwriting Performance using Tablets

Mekyska et al. (2023) had children write paragraphs on a display tablet and used the recorded trajectories to predict the scores of the HPSQ-C handwriting scale (Section 2.1.4), which is used to assess the handwriting performance asking about the writing process and product (Section 2.1.1). Manually engineered features, like the time in air, were used as input to a gradient boosting algorithm to predict the ratings of three items of the HPSQ-C questionnaire. The Mean Absolute Error (see Section 3.4.1) between the model predictions and the ratings ranged from 1.79 to 2.67 for the three individual items and was 5.6 for the overall summed score. With a range of one to five for the individual items and a range of three to fifteen for the total score, these mean errors seem substantial. In a binary classification task between dyslexic and normally developing children, an accuracy of 83.6% was achieved based on the predictions on a balanced test set.

2.2.3 Predicting Legibility with the STABILO DigiPen

The use of sensor-enhanced ballpoint pens allows for an assessment of the natural writing process on paper. The collected sensor data captures features of the writing process and is less descriptive of the writing product than the images and trajectories recorded on tablets. Grabmann (2023) assembled the *Curation Beauty* dataset from handwriting samples recorded with a pen similar to the one described in Section 3.1.2. Ordinal legibility ratings from one to three were assigned to the samples by the author. These ratings represent a count of legibility violations. For each of the five legibility criteria that the annotator saw violated, the ratings increased by one. Samples with more than three violations were assigned the maximum rating of three. Four deep learning network architectures were trained to discriminate between the three legibility ratings. The models were trained and tested on four different categories of data, which dictated how the training and the test were derived from the dataset. The models were compared on the basis of the mean accuracy they obtained on unbalanced test sets in five-fold cross-validation.

- **Category 1** With unbalanced training data and user-independent evaluation, mean accuracies for the different models laid between 41.12% (RNN) and 56.17% (CNN).
- **Category 2** With balanced training data and user-independent evaluation, the MLP achieved the highest accuracy of 45.54%, while both the RNN with 37.77% and the CNN architecture with 37.80% achieved equally poor results.
- **Category 3** Unbalanced training data and user-dependent evaluation led to the highest mean accuracies between 53.06% (RNN) and 63.38% (MLP).
- **Category 4** With balanced training data and user-dependent evaluation, mean accuracies were between 49.74% (RNN) and 62.47% (CNN).

In addition to the ratings by the author, a teacher also provided ratings for entire pages of student handwriting. These pages were then divided into shorter handwriting samples, which were given the same rating as the entire page. Experiments with these ratings used data of category three, so samples of the same student were allowed to appear in the training and test set. An accuracy of 96.5% was reported for the same CNN, which achieved 37.8% with category two data.

In the context of applying automated assessment in practice, the goal was to collect annotated data for a group of students to train models that develop generalizable features related to legibility. These features would then enable the models to assess the legibility of unknown students. To avoid introducing bias into the assessment and to test generalization, the second category holds the greatest significance. However, the results reported for category two are hard to interpret. Possible problems lie in the use of accuracy as metric for evaluation on unbalanced data, the unspecific nature of labels in the dataset, and flaws in model architecture.

Unbalanced Test Data Given the highly unbalanced nature of the test sets, accuracy does not appear to be a suitable or expressive metric. A model that always predicts the most common of the three ratings would have achieved an accuracy greater than 50%, which is higher than the reported accuracy of the four models. Training on unbalanced data makes the model favor the more common ratings, which leads to higher accuracy on an unbalanced test set with a similar distribution of ratings. Consequently, all four models achieved higher test accuracies with category-one data than with category-two data. This increase in accuracy cannot be attributed to an increased ability to determine the legibility of samples. In the dataset, the least frequent of the three legibility ratings marks the least legible samples. From a diagnostic point of view, it is important to reliably identify these samples with low legibility.

Unsuitable Labeling - Ambiguous Ratings As a consequence of the labeling scheme, the same rating can denote different shortcomings in a sample. For example,

a rating of one indicates that one of five criteria was found to be violated in the handwriting sample. A sample with irregular spacing received a rating of one, as well as a sample with irregular letter heights. This makes it harder to relate one rating to specific patterns in sensor data.

Unsuitable Labeling - Unspecific Ratings The accuracy of 37.80% measured when the CNN was trained with individual ratings per sample and data category two was compared to the 96.5% it achieved on samples annotated with ratings that describe entire pages and data category three. This comparison reveals that having identical ratings for all samples from an individual student in combination with a user-dependent evaluation leads to misleadingly high accuracy. It is assumed that samples of most students were contained in both the training set and the test set. Consequently, the model probably learned to identify student-specific patterns in the sensor data rather than finding features related to legibility. A model that is able to tell apart the handwriting samples of the 37 students in the dataset will perform well on the prediction task because all samples of the same student are annotated with the same rating. Consequently, the results show the ability of the model to learn the writing styles of 37 students, but not its ability to assess legibility.

Model Architecture The architecture of the CNN model is shown in Figure 8. As explained in Section 4.3.2 all convolutional layers operate on data of the same size. As a consequence, there is no incentive for the convolutional layers to condense the input data into more abstracted features throughout the convolutional layers. Without reducing the size of the feature map, the convolutional layers seem likely to bring little benefit to the learning ability of the model. The prediction is assumed to be determined mainly by the weights of the classification head.

2.2.4 Predicting Legibility with the SensoGrip Smart Pen

A similar approach to the automated assessment of legibility was described by Bublin et al. (2023). Reviewing related literature, they stated that traditional methods often relied on digital tablets and classical machine learning algorithms. In contrast, the authors used the SensoGrip smart pen, which captures detailed handwriting dynamics similarly to the pen described in Section 3.1.2. Raters assigned scores between zero and twelve according to the SEMS handwriting scale (Section 2.1.5). According to the study on the SEMS handwriting scale (Franken and Harris, 2021), the students were grouped as having handwriting difficulties or not according to the score assigned to their writing.

The models were trained for regression on the overall SEMS scores, rather than predicting ratings for individual questionnaire items, which focus on different aspects of handwriting performance separately. The dataset consisted of the handwriting of 22 students, the corresponding sensor data, and the SEMS score given for each piece of handwriting by a therapist. To increase the amount of training data, the 22

pieces of handwriting were split into 20 shorter handwriting samples. Each sample created this way was annotated with the overall SEMS score that was given for the original piece of handwriting as a whole. The best performance was reported for a long short-term memory network. It achieved a Root Mean Square Error (RMSE) of 0.68, which corresponds to a Mean Squared Error (MSE) of 0.56. Given that the model predicts values between zero and twelve, these errors appear to be low and suggest that the model is able to predict scores similar to the therapists. An accuracy of 99.8% was reported for the binary classification task, between samples of students with or without handwriting difficulties. The limitations of the proposed work put these results in perspective.

Unsuitable Labeling - Unspecific Ratings It seems questionable whether the samples' scores accurately describe their handwriting quality because the score is inherited from the original longer piece of handwriting. For example, such a piece of handwriting could start with a perfectly legible sentence and end with a sentence that is hard to read. To increase the training data, the writing is split. The first and last sentences are individual samples. With the explained annotation scheme, both the perfectly legible and the problematic samples receive the same score, which was assigned to the piece of writing as a whole and probably lies somewhere in between the two scores that would suit the two individual samples.

User-Dependent Evaluation The issues described above that arise from the approach of annotating samples for training and testing were intensified because samples from the same student were contained in both the training and the test set. As explained in the review of Grabmann (2023), the combination of having the same score assigned to all samples from an individual student and evaluating the model in a user-dependent way leads to high accuracies that are not generalizable for the practical application of automated assessment. The reported performance shows that the model is able to approximate the score by identifying who of the 22 students wrote a given sample.

Use of Additional Data In addition to the sensor signals, the age and gender of a student were given to predict the legibility of a sample. This allows the model to learn how age and gender are related to legibility. This learned bias helps improve the accuracy on the test set, but detracts from learning patterns in the sensor data that determine legibility.

3 Methodology

The idea was to capture the perceived legibility of many raters for a large set of handwriting samples. The ratings in this dataset represent an approximation of the concept of legibility. An assessment system should be created that rates unseen handwriting according to this approximation. Basing the assessment on the judgment of different raters is a step towards more objectivity and robustness, compared to relying on a single rater. Each time the system assesses a handwriting sample, the judgment is based on the opinions of the whole pool of raters who annotated the dataset. If the system gives ratings in accordance to the rater opinions, and the pool of raters constitutes a representative subgroup of all people who assess legibility, then the tool could be used to conduct the assessment in their place.

To train CNNs as such assessment systems and to address research questions, the first step was to settle on the criteria of legibility to investigate. Then, ten sentences were designed as reference text. Handwriting samples were recorded for these sentences with the STBILO DigiPen and then rated according to the criteria to assemble the *StabLe* dataset. The dataset was used to analyze rater agreement and to train models for automated assessment.

3.1 Data Acquisition

3.1.1 Choosing Legibility Criteria

The literature review revealed similar criteria that are addressed to assess the legibility in the handwriting scales discussed. For this work four legibility criteria were chosen. Three factors informed the decision on the criteria that should be investigated.

- The criterion must be grounded in research. There needs to be evidence that the criterion is relevant for the overall perceived legibility of handwriting.
- Expert experiences (engineers involved in developing the sensor pen) and opinions on whether a criterion can be assessed using only the sensor data provided by the pen were taken into consideration.
- Criteria with different degrees of specificity should be investigated. A criterion that asks for a general impression is said to be unspecific compared to one that asks for a specific characteristic of the strokes or letters.

Based on these factors, the common criteria shown in Table 2 were considered and refined to the four criteria shown in Table 4 together with the identifiers used later for reference.

The *global-legibility* was chosen as the first criterion. The raters are asked to give a rating that reflects their overall impression of legibility. An overall impression of

legibility is commonly rated in the reviewed scales. This is the least specific criterion because it is not directly determined by the specific characteristics of the shapes and strokes.

The *slant-consistency* was chosen as the second criterion. The raters are asked to give a rating that reflects how well the vertical strokes of the letters are aligned. This criterion is more specific because it is determined by a characteristic of the strokes, their slant, and it is common in the reviewed scales. It is assumed that information about the slant can be reconstructed from the provided sensor data.

Table 4: The four legibility criteria.

Question Identifier	Criterion
Q1	*global-legibility*
Q2	*slant-consistency*
Q3	*letter-formation_rnh*
Q4	*letter-formation_ad*

Letter-formation-ad and *Letter-formation-rnh* were chosen as the third and fourth criteria. The raters are asked to give a rating that reflects how well these letters are formed and how easily similar letters can be distinguished out of context. Letter formation was assessed in most of the reviewed handwriting scales. The correct form and shape of the individual letters determine legibility. Lewis and Lewis (1965) examined root causes for insufficient letter formation. They found that the incorrect size of parts of letters is the most common letter formation error. A large portion of letter malformations appears in a small subset of letters, which are similar in shape. They explain that the letters 'a' and 'd' are produced by nearly identical hand movements, and their shapes only differ in the extent of one stroke. The same holds for letters 'r', 'n', and 'h'. The question of whether the extent of these strokes is appropriate and whether the letters are distinguishable is specific.

3.1.2 Recording Materials

Different materials were used to record handwriting samples for the *StabLe* dataset.

Reference Sentences The text or content of the handwriting samples of *StabLe* was designed according to the following considerations.

- **Fixed Content** Previous work attempted to predict legibility in a wide variety of written text. According to the hypothesis that reducing the variety of content benefits model performances (**R1** in Section 1.3), this work altered the prediction task to assess legibility in a small number of reference sentences written by all subjects.
- **Dictionary** Because samples were planned to be recorded with students in grades five and six, sentences were not allowed to contain complicated terms. This was meant to ensure that students can read the sentences once and then write them without interrupting the writing flow.

- **Sample Length** To record uninterrupted writing, samples had to be short. In addition, it was hypothesized that one single legibility label for a longer text passage lacks detail. The rating that describes the passage as a whole can differ from the rating that would apply to a shorter excerpt. Rating a long passage requires the rater to summarize observations on different parts of the text into one rating. It is assumed that this makes the ratings more subjective and less reliable than they would be in shorter texts.
- **Letters** As criteria Q3 and Q4 focus on a comparison of letters 'r', 'n', and 'h' or 'a' and 'd', respectively, sentences were designed to contain these groups of letters.

Following these considerations, ten short sentences were designed as writing tasks for the dataset. The process of finding appropriate reference sentences was supported by members of the SMI, who could build on experience in the field of assessing legibility and working with students. The ten sentences are:

1. Der Hahn und der Hund tanzen.
2. Er stand da und lauschte.
3. Hannah hat ein Buch gelesen.
4. Quark ist besonders lecker.
5. David der Kater schnurrt sanft.
6. Die Pförtner lassen dich herein.
7. Sie wandern in Richtung Strand.
8. Kinder spielen draußen.
9. Ein heftiger Blitz leuchtet.
10. Bellende Hunde beißen nicht.

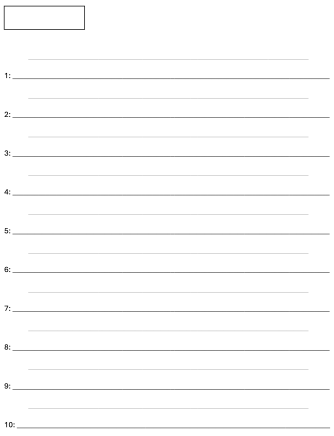
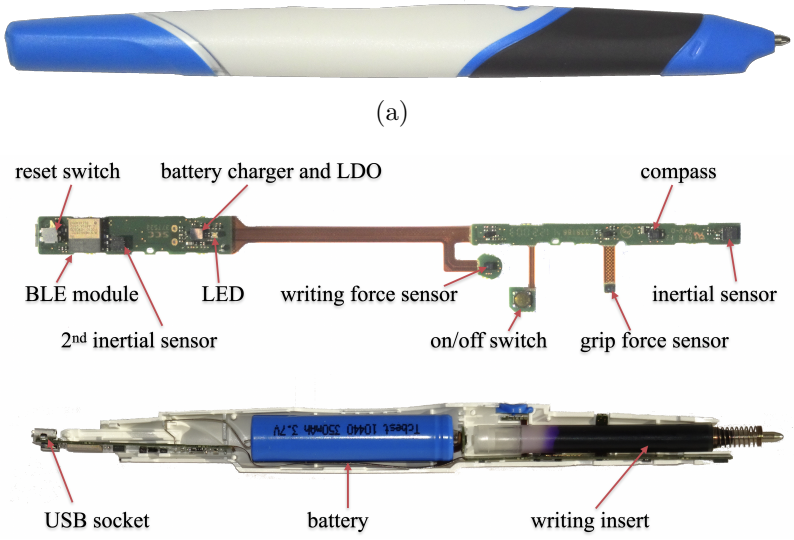


Figure 2: The recording sheet.

Recording Sheet For previous handwriting recordings at STABILO, participants wrote on regular lined, grid, or blank paper to collect samples for text recognition. For this work, the goal was to reduce the variability in the recordings as much as possible to make the task of predicting legibility as easy as possible. Furthermore, single sentences had to be extracted from the scans of the sheets afterward to prepare the labeling process. Therefore, uniform sheets with ten black lines numbered from one to ten were used so that students could write each of the ten sentences on the designated line. In case the students needed to correct themselves, a second gray line was printed for a second attempt at writing the sentence. At the top, the sheets contained a small box to fill in the student's user ID so that the sheets could be associated with the corresponding recorded sensor data afterwards. The sheet is shown in Figure 2

STABILO DigiPen and Recording App Students wrote with the STABILO DigiPen, hardware version 6.3. This pen is a sensor-enhanced ballpoint pen that writes on paper as a normal pen and captures the writing movement with an array of sensors. A force sensor in the tip measures how strongly the pen is pressed onto the paper while writing. In order to capture the movement of the pen, two sensors combining three-axis accelerometers and gyroscopes are used, one in the front of the pen and a second at the rear end of the pen. The pen and its technical components are shown in Figure 3. Sensor signals are recorded with a sampling rate of 400. In order to transmit the recorded sensor signals, the pen has internal data processing capabilities and contains a radio datalink for communication. External devices can be paired with the pen via Bluetooth to receive the stream of sensor data.



(a)

(b)

Figure 3: (a) Picture of the STABILO DigiPen used in this work. (b) The technical components of the pen.

STABILO maintains the HwDon BASIC app, an Android app built to carry out recording sessions with the described DigiPen. The app provides a user interface to pair with an available pen. The pen streams sensor data to the paired device, where it is stored locally and can be uploaded to the centralized Handwriting Donation Server, which hosts a database of handwriting recordings.

In the recording app, users are identified with a unique profile that contains information about the writer. Different projects prompt users with different sequences of samples (texts) to write. Buttons are used to navigate between samples and to restart the recording of a sample. Figure 4 shows the user interface subjects interact with during a recording. For each handwriting an, unique sample id is stored together with start and stop timestamps. After the recording session, the user is asked to take a picture of the produced handwriting with the tablet. These images are stored together with the recorded sensor data.

Figure 4: The Handwriting Donation App displaying a sentence for recording.

3.1.3 Recording Sessions

Handwriting samples were collected from 202 students in two recording sessions. The first recording session was conducted with 86 students in grade six at the Gymnasium Hilpoltstein on 11.06.2024. The second recording involved 116 students in grade five at Friedrich-Alexander-Gymnasium in Neustadt a.d. Aisch on 17.06.2024. The children and their parents were informed beforehand and had agreed to participate in the recording sessions.

Two STABILO employees and two students from the University of Bamberg prepared and supervised the recordings, while the schools provided the rooms and brought in groups of students. A tablet, a sensor-enhanced ballpoint pen, and the described sheet, were placed on twelve desks to record the handwriting of the students. They were briefly instructed about the usage of the recording app and could ask supervisors for help at any time. The students were advised to write in their natural writing style. The handwriting samples or sentences were recorded in the same order for all students and the sheet was subsequently photographed by the supervisors. The sheets were scanned afterwards to obtain higher-resolution pictures without distortions. Due to complications, sensor data was not properly recorded for all handwriting samples.

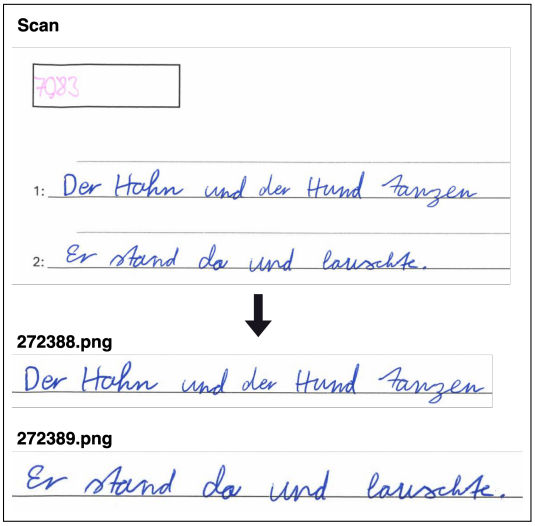


Figure 5: Two images of individual handwriting samples extracted from the scan.

3.1.4 Sentence extraction

Images of individual sentences were extracted from the scanned sheets. The main steps of the sentence extraction process were as follows. The user ID was identified within the scan, and the ID inside was read using pretrained optical character recognition (OCR) models. Bounding boxes were placed around each line that contained handwriting, and the content of each box was extracted and stored to a single image. OCR was used to read the text of each extracted image. For each of the ten sentences, the extracted image with the highest character match was stored.

3.1.5 Manual Curation and Validation

The recording and the sentence extraction produced handwriting samples consisting of a sentence, a sample ID, sensor data and an image of the handwriting.

The samples were all examined manually to verify the sentence extraction, and for later analysis and interpretation of the results. Tags were added to capture information on flawed samples and characteristics of their content.

- **cursive**: The sample was written purely in cursive letters.
- **typo**: The sample contained a spelling error.
- **correction**: The sample contained some correction by the student, for example a crossed out word.
- **image**: The image did not show the sentence as expected. Either parts were cut of or the image showed text from other lines as well.

Tracking whether a sentence was written in print or **cursive** seemed relevant, because the movement of the pen was assumed to be different between the writing styles. Furthermore, the writing style could affect how legible a sample is. The cursive tag was added when a sentence was written in cursive letters only, while many samples showed a mix of both cursive and block letters. If a sentence contained a **typo**, then the recorded sensor data does not correspond to the writing of the exact text of the reference sentence. Furthermore, spelling errors could affect the legibility ratings. The same holds for sentences that contained **correction**s. For samples with a flawed **image** the sentence was extracted manually from the scan. This was the case for twelve images.

3.2 Data Annotation

The recording sessions and manual curation described above resulted in 2017 unlabeled samples, of which sensor data was recorded for 1916. In order to train the supervised models, a legibility label had to be assigned to each sample.

To train models to assess the *global-legibility* the *StabLe*(Q1) dataset was created by collecting ordinal ratings of this criterion for each sample. A *global-legibility* rating of '1' denotes that the sample is perfectly legible and a rating of '5' marks it as not legible. Similarly, ratings were collected to compile datasets for the other criteria (Section 3.1.1).

To analyze and account for the suspected low inter-rater reliability, ratings for each sample were collected from multiple raters. The level of expertise was suspected to affect the ratings. Therefore, at least one rating per sample had to be collected from a rater with experience in assessing handwriting. To examine the intra-rater reliability, a subset of the samples was rated twice by some of the raters. Approximately 22000 ratings were needed for the 2017 samples and the four criteria.

3.2.1 Requirements to the Annotation Tool

To collect the ratings for the four labeled datasets, an annotation tool was needed. The had to provide a seamless annotation process and data integrity.

Simple User Interface: The annotation had to be self-explanatory and the rater had to be provided with all needed instructions. The tool had to provide a single streamlined and linear user experience where raters were first instructed and then rated the samples. To avoid problems, the user interface should not contain any settings that the rater could change.

Web-App: To make the annotation tool easily accessible, a web app was preferred over applications running locally. In this way, participants could annotate from any device. A web app with a designated backend allowed to manage several raters in parallel. Furthermore, the annotation process could be steered and adjusted in the centralized backend without involving the raters themselves.

Rollout Management: To collect the redundant ratings required for a subsequent analysis of inter- and intra-rater reliability, the tool had to provide logic for a planned rollout. Therefore, the tool had to store which rater already rated which sample. Based on this information and an adjustable prioritization, the next criterion and samples to be prompted to a rater were determined.

Batching of Samples: When using the tool, raters were prompted with batches of samples to increase efficiency. All samples in a batch were rated with respect to the same criterion. That way, the rater read the task instructions once before rating a batch of samples correspondingly.

Random Sampling: It was suspected that the order in which sentences were displayed could affect the ratings given. To reduce the impact of such effects, samples had to be displayed in random order within the constraints of the rollout strategy.

3.2.2 Existing Annotation Tools

With the specified requirements, existing annotation tools were examined. For an overview, the annotation tools mentioned in three lists maintained on GitHub were systematically reviewed.

- "awesome-data-labeling" by Tkachenko (2022) is a curated list of annotation tools maintained by *HumanSignal* who develop LabelStudio.
- "awesome-data-annotation" by Pungas (2022) is another curated list of annotation tools maintained by Taivo Pungas.
- "awesome-open-data-annotation" by van Linschoten (2024) is a curated list of open-source data annotation tools maintained by *ZenML GmbH*.

The three mentioned lists overlapped partially. Together, they covered 52 image annotation tools. Each tool was briefly examined by visiting the project GitHub page to answer the following questions.

- Is the project an application that allows to mark images with class labels (ordinal ratings)?
- Is the the application web based with a centralized backend?
- Does the web app offer a pure rater user role for which administrative functionalities are hidden?
- Is the user interface self-explanatory to the extent that in-person instructions are not needed to start annotating?

Ten listed applications supported image classification. Six of those were also web apps with a centralized backend and only the Computer Vision Annotation Tool (CVAT) offered a simplified rater user role. However, the user interface for raters, CVAT's job page, was still loaded with functionalities that were not crucial for the presented scoring task.

The reviewed tools did not allow placing longer task instructions up front on an onboarding page or similar. As raters would later use the tool without being taught how to use it, having detailed instruction texts was crucial and was part of the requirement for a simple user interface. Table 5 shows a subset of the tools reviewed and the checklist for the specified requirements, the full list is given in . None of the reviewed tools met all the defined requirements.

Table 5: Depiction of the review process on a subset of the reviewed annotation tools.

Name	Supports classification	Centralized web-app	Rater user-role	Simple user-interface
COCO Annotator	no	yes	yes	no
VoTT	yes	yes	no	no
LabelStudio (community)	yes	yes	no	no
CVAT	yes	yes	yes	no

3.2.3 Design of the Annotation Web App

A web application was developed to collect ratings for the four criteria. When the raters entered the website, they were first shown a login screen. A successful log-in led to a welcome page with a short description of the project goal. From there, users reached the batch introduction page. Here, an instructional text was given for one of the four criteria to explain how the next ten sentences should be rated. The full description texts are given in A.4. In addition to the text, an example image showed a handful of handwriting samples and suitable ratings for these. After the instruction page, an image of the first sentence was displayed with a five-point scale below to select a rating. Short interpretations of the numerical rating values were given to provide guidance. Translations of these interpretations are listed in Table 6. Next to the image, users could open a pop-up to review the example image or another one to report issues. When the rater had rated the ten sentences in the batch, the batch instruction page was shown again. Depending on the underlying configuration, the raters were then asked to assess one of the other criteria or the same one again. The user interface of the annotation tool is shown in A.4 and technical details are given in Section 4.2.

Table 6: Interpretations for the five numeric ratings for each question.

Rating	Q1	Q2	Q3	Q4
'1'	very legible	slant is consistent	very easy to distinguish	very easy to distinguish
'2'	legible	minor variation of slant	easy to distinguish	easy to distinguish
'3'	rather legible	varying slant	rather distinguishable	rather distinguishable
'4'	rather not legible	strong variation in slant	hard to distinguish	hard to distinguish
'5'	not legible	very strong variation in slant	very hard to distinguish	very hard to distinguish

3.2.4 Analysis of Rater Agreement

With the collected ratings of the four criteria, the rater agreement was measures using the ICC and Cohen's Kappa (Section 3.4). The obtained values for inter- and intra-rater reliability were compared to the agreement reported for the renowned handwriting scales (Section 2.1.5) to examine how reliable the ratings in the *StabLe* dataset are (**R1** in Section 1.3). Furthermore, agreement between the different criteria was compared to examine whether the criteria that were assumed to be more specific were rated more reliably (**R2** in Section 1.3).

3.3 Machine Learning

Several models were trained to examine how the legibility ratings of the *StabLe* dataset can be modeled and to investigate the research questions. A list of all models is given in Section A.7.

3.3.1 Setup of Training Runs

The models were trained to predict the legibility labels derived from the ratings in *StabLe*. The preprocessing of sensor data described in Section 4.1 produced 1907 samples for model training and evaluation. Aspects of the training setup and model evaluations are described in the following

Classification and Regression Models The task of predicting legibility was initially framed as a classification problem because this resembles the act of rating the sample with discrete legibility classes. Furthermore, this allowed to compare results on the new dataset to results of classification models trained in previous work. Subsequently, regression models were trained. Regression models were hypothesized to make better use of ground-truth labels derived from the ratings because the ratings represent an ordered sequence rather than distinct and unrelated classes. A false prediction of the rating '4' while the ground truth is '1' leads to stronger weight adoption than predicting '1' using a regression model. In a classification model, both errors weigh equally.

Question-Specific Models and Datasets All models were trained to be specific to a single question, which means that each model is trained to assess samples according to the ratings collected for one of the legibility criteria (Section 3.1.1). The dataset with ratings of *global-legibility* is denoted as *StabLe*(Q1) and the datasets for the other criteria accordingly.

User-Independent Data Splits Models were trained on user-independent data, so they learn to assess handwriting legibility solely based on sensor data recorded during writing and without prior knowledge of the subject.

Label Merging Strategies For each criterion, the according *StabLe*(QX) dataset contains several ratings per sample. For example, a sample may have been rated with respect to question Q1 by six raters. When this sample is fed into a network for training, then the supervised models used in this work demand a single ground-truth label. The mapping from all ratings of a sample to a single ground-truth label at run-time is called label merging in the following. Figure 6 shows an example of how the ground-truth label is derived from six ratings for a single sample with different labeling strategies. The five label merging strategies are examined in this work are:

- **rounded-mean-label** The ratings are averaged and then rounded to obtain a discrete ground-truth label.
- **majority-label** The most common rating is chosen as the discrete ground-truth label. In case of a tie, the lower rating is chosen.
- **random-label** At runtime, one of the ratings is randomly chosen as the discrete ground-truth label. Ratings of all raters have the same probability of being picked.
- **mean-label** The ratings are averaged to obtain a continuous ground-truth label.
- **rater-specific-label** Ratings from a specific rater are used directly as discrete ground-truth labels, ignoring the other ratings of the sample.

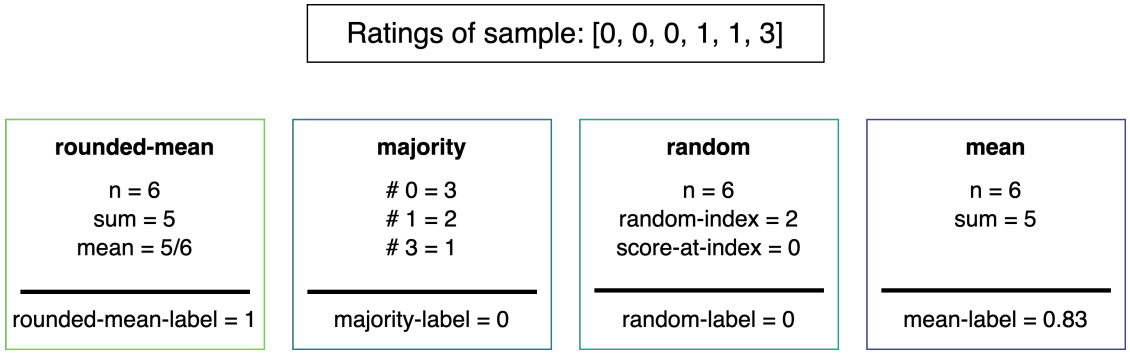


Figure 6: Exemplary depiction of four label merging strategies.

Data Balancing In all datasets *StabLe*(Q1 - Q4) ratings were not evenly distributed. The ratings '1', '2', and '3' were assigned frequently while '3' and '4' were rather rare. For the purpose of training machine learning models, unbalanced data is undesirable because it can introduce bias into the predictions, which detracts from learning features in the data that are decisive of each label. This problem was addressed by oversampling samples with rare ratings during training. That is, a handwriting sample with rare ratings was drawn several times per epoch. Consequently, the model was exposed to samples with rare ratings more often than it would be when sampling randomly (unweighted) or without replacement. Technical details on balancing the training data are given in Section 4.3.4.

Validation and Testing of Models The validation set is used for model comparison, that is, for experiments that examined how architecture, label merging, machine learning tasks, and other parameters affect model performance. The validation set was also for model selection within the training runs to determine a winning epoch and the corresponding model (Section 4.3.1). The test set was kept for final evaluations. That way, the selection of parameters was tuned towards the validation set, but not the test set. Once the best models were found, their performance was evaluated on the test set. These results serve as a point of comparison with other work.

3.3.2 Comparative Experiments

To explore how the ratings of *StabLe*(Q1) can be predicted using CNNs and to address research questions **R3** and **R4** (Section 1.3), a range of models were trained and compared changing different parameters of the training setup. The models were trained as part of the following experiments. For these experiments, models were evaluated on the validation set only.

Reproducing Results of Previous Work (Repr) As described in Section 2.2.3, Grabmann (2023) trained several machine learning models on the *Curation Beauty* dataset of 3290 handwriting samples. The samples are labeled with one of the three legibility classes. Both the code and the dataset were available for this work. To check if the results were reproducible (**R3** in Section 1.3), the described CNN was recreated and trained with the given dataset. The model A (Table 7) consists of the *CNNbroad* convolutional layers followed by the *1l class* head (Section 4.3.2). Here, the described approach to balancing the training data was applied while data balancing was done by removing samples with more frequent ratings before.

Table 7: The models trained to compare the different CNN architecture (CnnArc).

Model	CNN	Head	Dataset	Label	Task
A	broad	1l	*Curation Beauty*	-	3 class
B	cone	1l	*Curation Beauty*	-	3 class
C	broad	1l	*StabLe*(Q1)	mean-rounded	3 class
D	cone	1l	*StabLe*(Q1)	mean-rounded	3 class
E	broad	1l	*StabLe*(Q1)	mean-rounded	5 class
F	cone	1l	*StabLe*(Q1)	mean-rounded	5 class

Comparing CNN architectures (CnnArc) Potential shortcomings were identified within the *CNNbroad* architecture previously used. These issues were addressed in the *CNNcone* architecture (Section 4.3.2. To test whether this change in architecture benefits performance, three pairs of models, which differ only in the architecture of convolutional layers, were trained and compared. All models use the

single-layer classification head *1l class*. Models A and B were trained on the *Curation Beauty* dataset. Models C and D were trained on *StabLe*(Q1) using *rounded-mean* labels but with the five legibility labels mapped to just three legibility classes. For that, labels '1' and '2' (very legible, legible) make class 1, '3' (rather legible) is class 2 and '3' and '4' (rather not legible, not legible) make class 3. Models E and F were trained on the original *StabLe*(Q1) dataset using *rounded-mean* labels with original legibility ratings ranging from one to five. The models are summarized in Table 7

Comparing Datasets (CompDs) To test whether models perform better in predicting legibility labels when the variety of texts is reduced (**R4** in Section 1.3), models trained on the *Curation Beauty* dataset were compared to models trained on the *StabLe* dataset. Handwriting samples of the *Curation Beauty* dataset were recorded for a wide variety of texts. The *StabLe* dataset contains samples of just ten fixed sentences. In the *Curation Beauty* dataset, a rating resembles a count of legibility violations, so the same rating can result from different combinations of violations. As a consequence, the curation beauty ratings were assumed to be coupled rather loosely with qualities within the handwriting samples. The legibility ratings of the *StabLe* dataset refer to individual legibility criteria, which could be related more directly to patterns in the data. Models A and B were compared to models C and D to assess if the ratings and reduced variety of texts of the *StabLe* dataset allow for better generalization. The models are summarized in Table 8

Table 8: The models trained to compare different datasets (CompDs).

Model	CNN	Head	Dataset	Label	Task
A	broad	1l	*Curation Beauty*	-	3 class
B	cone	1l	*Curation Beauty*	-	3 class
C	broad	1l	*StabLe*(Q1)	mean-rounded	3 class
D	cone	1l	*StabLe*(Q1)	mean-rounded	3 class

Comparing Prediction Heads (HeadArc) This experiment tested different prediction head architectures. The models discussed so far used a single-layer classification head. This design was compared to a three-layer head that contains nonlinear activation functions. This was believed to increase the capability of the model to map embedding vectors to wanted labels because the head is able to learn more complex functions that map the features from the CNN layers to final outputs. Furthermore, two similar regression heads were tested. The four prediction heads are described in Section 4.3.2. Models F and H were compared to G and I to see if more complex prediction heads improve performance. Models F and G were compared to H and I to see if a classification or regression task is better suited for the prediction of legibility labels. The models are summarized in Table 9

Table 9: The models trained to compare different prediction heads (HeadArc).

Model	CNN	Head	Dataset	Label	Task
F	cone	1l	*StabLe*(Q1)	mean-rounded	5 class
G	cone	3l	*StabLe*(Q1)	mean-rounded	5 class
H	cone	1l	*StabLe*(Q1)	mean-rounded	reg
I	cone	3l	*StabLe*(Q1)	mean-rounded	reg

Comparing Label Merging Strategies (LabMerg) The *StabLe* dataset contains ratings of several raters per sample. The experiments described so far used the *rounded-mean* strategy to derive the single ground-truth label of a sample at runtime. Different label merging strategies were examined to handle the variety of ratings per sample for supervised training (**R5** in Section 1.3). Model H (*rounded-mean-labels*) was compared to models J (*majority-label*), K (*random-label*), L (*mean-label*), and M (*rater-specific-label*). Model M was trained on all samples with Q1 ratings of rater '4' (the author). The models are summarized in Table 10.

Table 10: The models trained to compare label merging strategies (LabMerg).

Model	CNN	Head	Dataset	Label	Task
H	cone	1l	*StabLe*(Q1)	mean-rounded	reg
J	cone	1l	*StabLe*(Q1)	majority	reg
K	cone	1l	*StabLe*(Q1)	random	reg
L	cone	1l	*StabLe*(Q1)	mean	reg
M	cone	1l	*StabLe*(Q1)	rater-specific	reg

Hyperparameter Tuning (HypPar) Based on previous experiments, the models K and L appeared to be best suited to predict legibility labels. It was examined how adjusting classical hyperparameters affected the performance of these models. The models were trained with different learning rates, kernel sizes, and regularization parameters.

3.3.3 Evaluating and Verifying Experiments

Previous experiments explored the performances of models with respect to predicting labels of the *StabLe*(Q1) dataset. These models were evaluated only on the validation set. The two models K and L, which achieved the highest agreement with the raters, were further evaluated, as well as the rater-specific model M. *StabLe*(Q1) was filtered to train variations of model L to investigate how different uncontrolled independent variables in the samples contributed to model performance. Furthermore, variations of model L were trained with ratings of questions Q2, Q3, and Q4 to examine whether more specific criteria can be predicted more accurately (**R6** in Section 1.3). Performances on the held-out test set were examined to assess the models based on data that they had not seen and toward which they had not been tuned.

Evaluation of Models on the Test Set For previously trained models K, L, and M, the test set was used to obtain generalizable measurements on unseen samples.

Ratings were used to derive ground-truth labels *and* to evaluate model agreement, introducing data leakage. Rater '4' rated all samples with respect to Q1, so these ratings directly influenced the ground-truth labels. The model, trained to approximate these labels, was then evaluated against rater '4' and the other raters, making the agreement value a reflection of how well the model aligns with a rater whose ratings influenced the training data. To evaluate the model in an application-oriented way, the ratings of a benchmark rater were excluded from the training and only used for evaluation. Here, rater '4' is the benchmark rater, while raters '3', '16', '25', '27', '29', and '30' form the training rater group. The ground truth-labels were derived from *StabLe*(Q1)*, which contained only the ratings of the training raters. With this adopted dataset, a variation of model L, termed L*, was trained accordingly. The model was then compared to the benchmark rater, which is similar to measuring the agreement of two independent raters. If a model achieves an agreement with independent raters (excluded from training data) as high as the agreement between such raters, then this would suggest that the model could be used to complement or even perform the legibility assessment in their place. The four models that were evaluated are listed in Table 11.

Table 11: The models evaluated on the test set.

Model	CNN	Head	Dataset	Label	Task
K	cone	1l	*StabLe*(Q1)	random	reg
L	cone	1l	*StabLe*(Q1)	mean	reg
L*	cone	1l	*StabLe*(Q1)*	mean	reg
M	cone	1l	*StabLe*(Q1)	rater-specific	reg

Evaluation of Models for the Different Legibility Criteria The four models L, N, O and P were each trained identically with *mean-labels* derived from the ratings given for the four questions Q1, Q2, Q3 and Q4, respectively. The models' performance was analyzed to determine whether the specific criteria - *slant-consistency*, *letter-formation_rnh* and *letter-formation_ad* - could be more accurately derived from sensor data compared to the broader criterion of *global-legibility* (**R6** in Section 1.3). The four models that were evaluated are listed in Table 12.

Table 12: The models evaluated on the test set to compare the four legibility criteria.

Model	CNN	Head	Dataset	Label	Task
L	cone	1l	*StabLe*(Q1)	mean	reg
N	cone	1l	*StabLe*(Q2)	mean	reg
O	cone	1l	*StabLe*(Q3)	mean	reg
P	cone	1l	*StabLe*(Q4)	mean	reg

Evaluation of Models Considering Uncontrolled Variables So far, models were trained and evaluated on all samples of the *StabLe* dataset. Based on the tags described in Section 3.1.5 three filtered datasets were created and then used to train variants of model L to examine the effect of so far uncontrolled variables in the samples on model performance. At first, the 369 samples tagged as being written purely in cursive letters were excluded from *StabLe*(Q1) to obtain *StabLe*(Q1\cursive). The 70 samples that were tagged as containing a spelling error were removed for *StabLe*(Q1\typo). The third dataset *StabLe*(Q1\correction) excluded the 98 samples with corrections. The corresponding models are listed in Table 13.

Table 13: The models evaluated on the test set to examine uncontrolled variables.

Model	CNN	Head	Dataset	Label	Task
L	cone	1l	*StabLe*(Q1)	mean	reg
$L^{\setminus cursive}$	cone	1l	*StabLe*(Q1\cursive)	mean	reg
$L^{\setminus typo}$	cone	1l	*StabLe*(Q1\typo)	mean	reg
$L^{\setminus correction}$	cone	1l	*StabLe*(Q1\correction)	mean	reg

3.4 Evaluation Metrics

Previous work on predicting legibility ratings from sensor data was discussed in Section 2.2.3 and Section 2.2.4. As noted, the chosen metrics in combination with an unbalanced test set and user-dependent evaluation weakened the informative value of the reported results. Although the test set was unbalanced in this work as well, all evaluations were user-independent. Furthermore, classical machine learning metrics were complemented with agreement evaluations to assess the reliability between predictions and individual raters.

3.4.1 Machine Learning Performance Metrics

Classical machine learning metrics were recorded for each epoch with the training and validation set to track the training process of the models and to compare their learning capabilities. These metrics compare predictions with the corresponding ground-truth labels, which were derived using different label merging strategies (Section 3.3), so that they do not account for the uncertainty of the ratings when derived deterministically.

Cross Entropy Loss (CE) was used as loss to train classification models and to select the best performing model within a training run. It measures the difference between the predicted probability distribution and ground-truth labels. The negative log-likelihood of the correct class gives the training loss, encouraging the model to assign higher probabilities to the correct class.

The Mean Squared Error (MSE) measures the average squared difference between predictions and ground-truth labels. It was used as loss to train regression

models and for model selection. Used as training loss it encourages the model to predict values as close to the ground-truth label as possible while penalizing greater errors more heavily.

The Mean Absolute Error (MAE) measures absolute differences. It is less sensitive to outliers and more straightforward to interpret.

Accuracy (Acc) is the ratio of correctly predicted labels to the total number of samples. It was used to state the performance of classification models, noting that it is likely to produce misleading values due to unbalanced data.

Precision measures the proportion of true positive predictions out of all positive predictions made by the model. It evaluates how accurate the model is when predicting a positive class. It can only be calculated for binary classification tasks. Here, it was first calculated for each class against the remaining classes and then averaged to be applicable to the three- and five-class settings ($\overline{Prec}$).

Recall measures the proportion of true positive predictions out of all actual positive samples. It evaluates the model's ability to correctly identify all positive samples. Similarly to precision, it is only applicable to binary classification tasks and was averaged to be applied with three and five classes ($\overline{Rec}$).

3.4.2 Rater Agreement Metrics

To put the performance of the models in perspective with the uncertainty of the legibility ratings, trained models were evaluated with respect to their agreement with the raters. The agreement assesses the association between two attempts to measure the same construct (Liu et al., 2016), here the concept of legibility. An assessment of agreement comprises intra- and inter-rater reliability. Both were assessed when the reliability of the *StabLe* dataset was examined (Section 3.2.4). As models are deterministic, they exhibit perfect intra-rater reliability. Consequently, only inter-rater reliability between the model and the raters was evaluated.

Cohen's Kappa measures the inter-rater reliability with respect to class-level agreement. It is calculated for pairs of raters based on the samples they both rated. In this work, Kappa was used to measure the intra-rater reliability between the discrete ratings of different raters and between models and raters. To measure the agreement between regression models and raters, the continuous predictions were rounded to obtain discrete class labels. To state the agreement of models with multiple raters, Kappa is calculated between the model and each individual rater pairwise and the averaged ($\overline{Kappa}$)). Interpretations of Cohen's Kappa values according to Sim and Wright (2005) are given in Table 14.

Table 14: Interpretations of Cohen's Kappa.

Kappa	Strength of Agreement
< 0	Poor
0.01 - 0.20	Slight
0.21- 0.40	Fair
0.41 - 0.60	Moderate
0.61 - 0.80	Substantial
0.81 - 1.00	Almost Perfect

The **Intraclass Correlation Coefficent** (ICC) measures the agreement between continuous ratings of the same concept. It is calculated for samples with multiple ratings. Its applicability in this work was reduced because the raters provided ordinal ratings that were then compared to discrete outputs of the classification models or continuous outputs of the regression models. There are variations of the ICC, each suitable for a different study setup as described by Shrout and Fleiss (1979). ICC(3,1) was chosen following the guidelines of Koo T. K. (2016), who also provide the interpretations for the ICC shown in Table 15. The ICC was assessed for different fixed pairs of raters. The raters were not randomly chosen to participate. So, the observed reliability cannot be generalized to be the reliability found in the total population of raters. Therefore, Two-Way Mixed-Effects Model=ICC(3,-) was chosen. The reliability of individual rater pairs was assessed instead of measuring agreement for groups of raters. Therefore, ICC(-,1) was chosen.

Table 15: Interpretations of ICC.

ICC	Strength of Agreement
< 0.50	poor
0.51 - 0.75	Moderate
0.76 - 0.90	Good
0.91 - 1.00	Excellent

4 Implementation

This section covers technical details of the sensor data preprocessing, the developed annotation tool and implementations related to machine learning, like the model architectures or the data balancing approach.

4.1 Preprocessing of Sensor Data

For each sentence that a student had written, the pen streamed ten sensor channels to the connected tablet. After the recording sessions, this sensor data was uploaded to the Handwriting Donation Server. Figure 7 displays a subset of the files stored for the recording of each student.

The "meta" file contains information such as the user ID, profile data the student entered, the version number of the pen, and the sampling rate. The "calibration" file stores data for calibrating a compass. The table stored in the "labels" file contains a row for each recorded handwriting sample. For each sample the 'Label', that is, the text that was copied, is stored alongside two timestamps of when the text was starting to be displayed in the app and when the user pressed 'continue'. Lastly, each row contains a unique sample ID. The "StreamData" file contains one line per time step in which the signals were read from the pen sensors. Each line contains a timestamp and the ten signals of the four sensors. The two accelerometers and the gyroscope record three signals each because they register movement along three axes. The force sensor produces a single signal. Due to complications, not all sentences were recorded properly. The data consisted of 1916 handwriting samples with properly recorded sensor data.

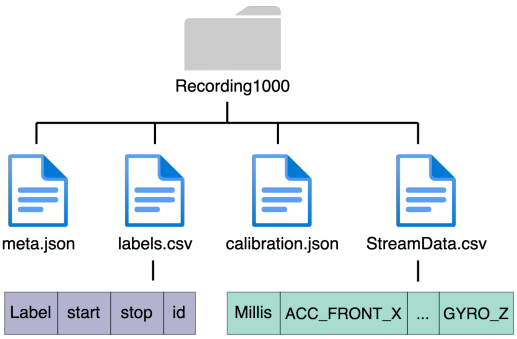


Figure 7: Sketch of the data format of the Handwriting Donation Server.

A preprocessing pipeline was developed to extract the sensor data of each handwriting sample from the stream data files and to prepare the sensor data for further use. For each preprocessing step, a corresponding preprocessor class was created to carry out the necessary manipulations. The pipeline consisted of the following six steps. Table 16 shows how the sensor data changed during preprocessing.

1. `ReadInXaiDatasetPreprocessor` The processor extracted the stream data corresponding to individual samples and stored it to a separate file for direct access.

2. `DownSamplingDatasetPreprocessor` The sensor data was down-sampled from 400 values per second to 100 by replacing each subsequent four values by their mean.

3. `TrimDatasetPreprocessor` The force sensor values were used to detect the timestamps that were recorded before and after writing. Data was cut to include only the time steps between the first and last contact with the paper.

4. `RemoveLongSamplesDatasetPreprocessor` Samples that were longer than 60 seconds after trimming are removed.

5. `AddPaddingDatasetPreprocessor` All samples shorter than the longest remaining sample were padded with zero values in the end to have matching length.

	# samples	mean	min	max	rate	storage
1	1916	29.5	8.0	729.4	400	3.04
2	1916	29.5	8.0	729.4	100	0.87
3	1916	21.6	3.9	685.6	100	0.65
4	1907	21.1	3.9	59.4	100	0.63
5	1907	59.0	59.0	59.0	100	0.93

Table 16: Descriptive statistics of the sensor data after each preprocessing step. The number of samples contained in the dataset, mean, minimum and maximum length of samples in seconds, sampling rate in times per second and storage in GB.

4.2 Data Annotation Web Application

The backend application with the REST API shown in Table 17 was developed in Kotlin using the Spring Boot framework. The frontend shown in A.4 was developed with ReactJs and provides functionalities only for the raters. All administrative functionalities were implemented in the backend and can be steered using the admin endpoints. The rollout configuration provided a way to prioritize the criteria. By updating this configuration, ratings for specified questions were collected first until the data was complete, before raters were asked to rate other criteria. The backend and the frontend web application were combined in a wrapping GitHub project and

Endpoint	Type	Role	Description
/users/login	POST	User	Authenticate to use the application
/batch	GET	User	Prepare a batch of ten handwriting samples
/files/image	GET	User	Load the image of a sample from the server
/batch/score	POST	User	Store the selected rating for a sample
/users	POST	Admin	Create a new user account
/config	POST	Admin	Update the configuration that steers the rollout
/answers	GET	Admin	Export answers from the raters and meta data
/reports	Get	Admin	Export all reports

Table 17: Short descriptions for the endpoints of the annotation tool backend.

dockerized. The applications resided on a privately hosted server and the web app was accessible under labeling.stabilodigital.de.

4.3 Machine Learning

4.3.1 Setup of Training Runs

All models were trained on an NVIDIA GeForce RTX 2080 SUPER GPU provided by STABILO. With the compact network size of the CNNs used, the training runs took about five to ten minutes. The computational effort was not a focus in this work and was not tracked further.

The datasets contain multiple discrete ratings per sample and criterion. Based on labels derived from these ratings (Section 3.3), the supervised classification and regression models were trained. The classification models were trained using CE with three or five classes and regression models were trained using MSE as loss (Section 3.4.1).

Maxabs normalization was applied in the training, validation and test sets separately. This avoids data leakage that would be introduced when the *StabLe* dataset would be normalized as a whole before splitting it.

Xavier uniform initialization (Bengio and Glorot, 2010) was used for the linear layers of all models. For convolutional layers, Kaiming uniform initialization (He et al., 2015) was used.

`AdamW` (Kingma and Ba, 2014) was used as the optimizer in all training runs. Unless stated differently, a learning rate of 0.001 was chosen and no regularization was introduced as weight decay was set to 0.

All models were trained for 100 epochs. For all training runs, improvements with respect to the validation set stalled before the 100th epoch, so that this fixed number of epochs seemed appropriate.

Model selection within single training runs was done based on the validation set. For both classification and regression, the model at the epoch with the lowest loss on the validation set was chosen as the best-performing model. As some models performed best within the first epochs, where the training error was still high, a minimum of 50 was introduced for the winning epoch. The model of the winning epoch was then evaluated on the validation set. The evaluation of the winning model on the test set was only performed for the final models after architecture, label merging, machine learning tasks, and other parameters were selected based on previous experiments.

4.3.2 Network Architectures

Two CNN architectures were compared. The first architecture was adopted from Grabmann (2023) and is called *CNNbroad* in the following. As discussed in Section 2.2.3, this architecture does not make perfect use of its convolutional layers,

because channel lengths are not scaled down as data is fed through the network. This is illustrated in Figure 8. After each convolutional layer (and the trailing batch normalization and ReLU activation), channels have the same length. After the last convolutional layer, the resulting 64 channels are each reduced to a single scalar value by calculating the mean per channel. This is referred to as `GlobalAvgPool1d`, although it was implemented using the `AvgPool1d` PyTorch class with the kernel size equal to the channel length. The resulting vector of 64 scalar values is the embedding of the input data. A classification or regression head takes it as input to produce the corresponding prediction. Reducing each channel to a single value in this one single pooling step was assumed to lead to a loss of information. The convolutional layers before do not reduce the channel size, as is often done with CNNs. It was assumed that this CNN fails its main purpose of condensing large input data into few information-rich features step by step.

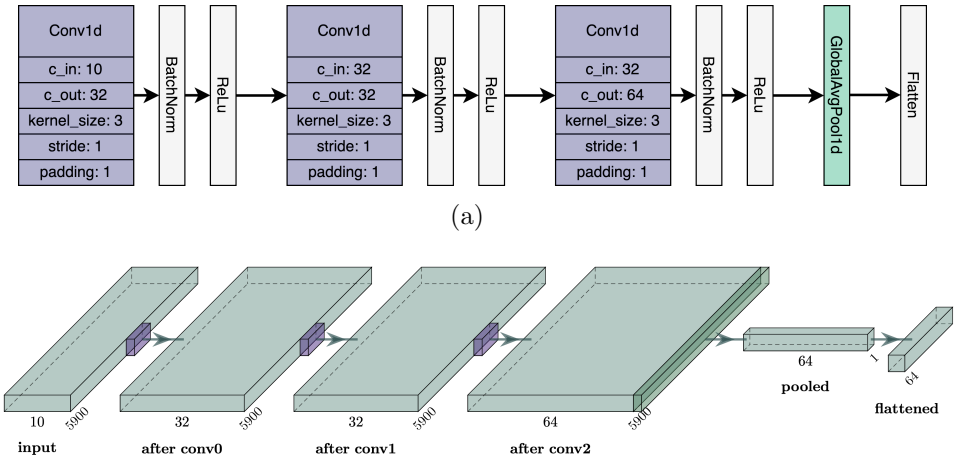
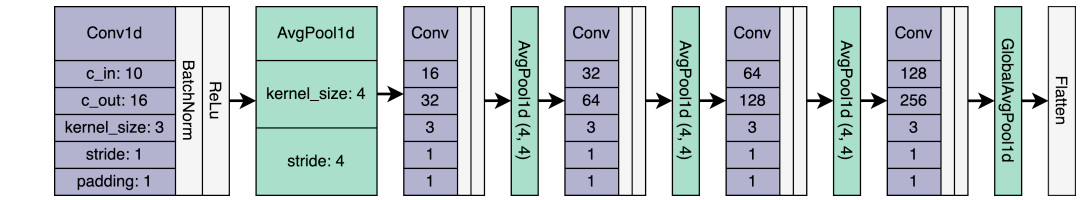


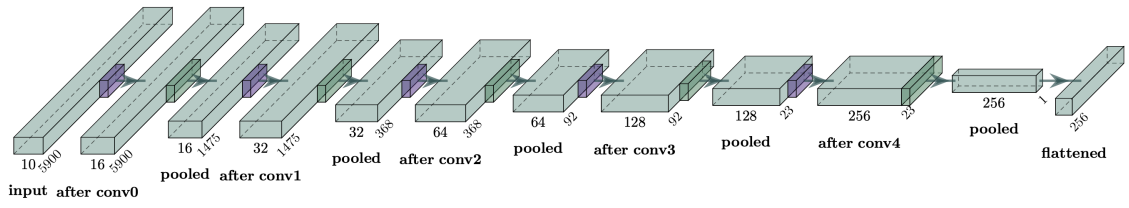
Figure 8: Architecture of *CNNbroad*: (a) Displays the layers the model is composed of and (b) illustrates feature map sizes as a sample of ten sensor channels with 5900 time steps each is fed through the different layers. Layer names are adopted from the corresponding `PyTorch` class names. The convolutional and pooling layers in (a) are represented by boxes of similar color in (b), while other layers are left out. The number of input and output channels of convolutional layers are denoted as c_in and c_out.

To address these issues, a CNN called *CNNcone* was constructed from the same building blocks as the *CNNbroad*. Average pooling is applied after each convolution, so that the length of data is reduced every time the number of channels is increased. This results in the desired progressive reduction in spatial dimension, which is indicated by the cone shape visible in Figure 9. After each block consisting of a convolutional and a pooling layer, less data points are available per channel. The number of channels increases with each such block. This is an incentive to dissect the input data into different features relevant to predicting the target label because not all information can be held on to (as it is possible with *CNNbroad*). The

input length of 5900 time steps (for sample of 59 seconds Table 16) leads to channels of length 23 after the last convolution. A final average pooling layer with a kernel width of 23 (referred to as `GlobalAvgPool1d`) reduces each of the 256 channels into a single scalar, yielding an embedding vector of length 256.



(a)



(b)

Figure 9: Architecture of *CNNcone*: (a) Displays the layers the model is composed of while (b) illustrates feature map sizes as a sample is fed through the network.

To summarize the two CNN architectures, the *CNNcone* reduces the input data to an embedding vector in incremental steps. *CNNbroad* does this by a single pooling layer without any learnable parameters. Subsequent to the described layers, a classification or regression head maps the embedding to a single prediction. Grabmann (2023) used a single fully connected layer together with the *CNNbroad* architecture. The layer takes in the embedding and outputs three scalar values corresponding to the three classes in the dataset used. The last `Softmax` layer rescales the values to resemble a probability distribution. In addition to the *CNNcone* architecture, three new prediction heads were created. The four prediction heads are depicted in Figure 10. *1l class* is the previously used single-layer classification head, and *1l reg* is a single-layer regression head. *3l class* and *3l reg* both consist of three fully connected layers that reduce the embedding vector to class outputs and a single scalar. The classification heads were adjusted to the three or five classes in the datasets.

4.3.3 Data Splits

After the preprocessing described in Section 4.1, the *StabLe* dataset contains 1907 handwriting samples, which were written by 202 students. All experiments in this work were conducted with the same data splits. To train user-independent models, all samples written by the same student were assigned to the same split. Approximately 70% of the students were randomly assigned to the training set, 20% to the validation set, and 10% to the test set. Both the validation and test sets were

checked to contain the least frequent ratings '3' and '4' as answers to Question Q1. Table 18 shows the number of samples, the number of students, and the number of ratings for question Q1 of each split. For the other questions, refer to A.5. The displayed distribution shows the total counts of ratings. The ground-truth label distributions drawn and derived during training deviated from the displayed distribution depending on the question and the label merging strategy (Section 3.3).

Table 18: Composition of training, validation, and test splits for *StabLe*(Q1).

Split	# samples	# students	# rating 1	# rating 2	# rating 3	# rating 4	# rating 5
Train	1398	140	2758	2133	1535	719	191
Validation	410	41	836	672	417	174	31
Test	209	21	270	282	272	202	45

4.3.4 Data Balancing

The balancing was implemented with the PyTorch `WeightedRandomSampler`. This sampler associates a weight with each sample. When samples are drawn to assemble a batch, they are chosen randomly and with replacement, while the probability of drawing each individual sample is dictated by its weight.

Let $\boldsymbol{x}_1, ..., \boldsymbol{x}_n \in \mathbb{R}^{c \times l}$ be the samples of the data set with the number of channels c and the length l. Then $\boldsymbol{y}_1, ..., \boldsymbol{y}_n$ with $\boldsymbol{y}_i \in \mathbb{R}^{m_i}$ are the corresponding lists of labels available per sample with varying lengths $m_1, ..., m_n$. For each scalar y that

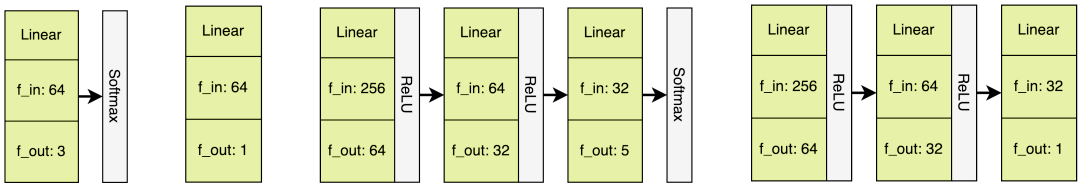


Figure 10: Architectures of prediction heads: (a) Displays the single-layer classification head *1l class*, (b) is the single-layer regression head *1l reg*, (c) shows the classification head with three layers and *3l class* (d) shows the corresponding regression head *3l reg*. The number of input and output features of the fully connected linear layers are denoted as f_in and f_out.

appears as a label of a sample in some $\boldsymbol{y}_i$ a label weight w_y is calculated as the inverse frequency of this label score:

$$w_y = \frac{\#\text{ all labels}}{\#\ y}$$

For each sample x_i its intermediate sample weight is then given as the mean of its label weights:

$$w_i^* = \frac{1}{m_i} \sum_{j=0}^{m_i} w_{\boldsymbol{y}_{i,j}}$$

A trade-off between balancing the training dataset and overfitting to single rare samples was taken by limiting the weight of each sample to be at most ten times the weight of the sample with the lowest weight $w^*_{\text{min}} = \min(w_1^*, ..., w_n^*)$. So the final weight of each sample is given as:

$$w_i = \min(w_i^*, 10w^*_{\text{min}}) \tag{1}$$

As explained, the `WeightedRandomSampler` randomly draws samples from the entire dataset with replacement. Therefore, the samples drawn per epoch vary, which affects the exact distribution of samples and labels fed to the model. Figure 11 showcases how this procedure produced a more balanced distribution of labels in an exemplary drawing. Without balancing, labels '4' and '5' ("rather not legible" and "not legible") would make up a negligible portion of samples fed to the models.

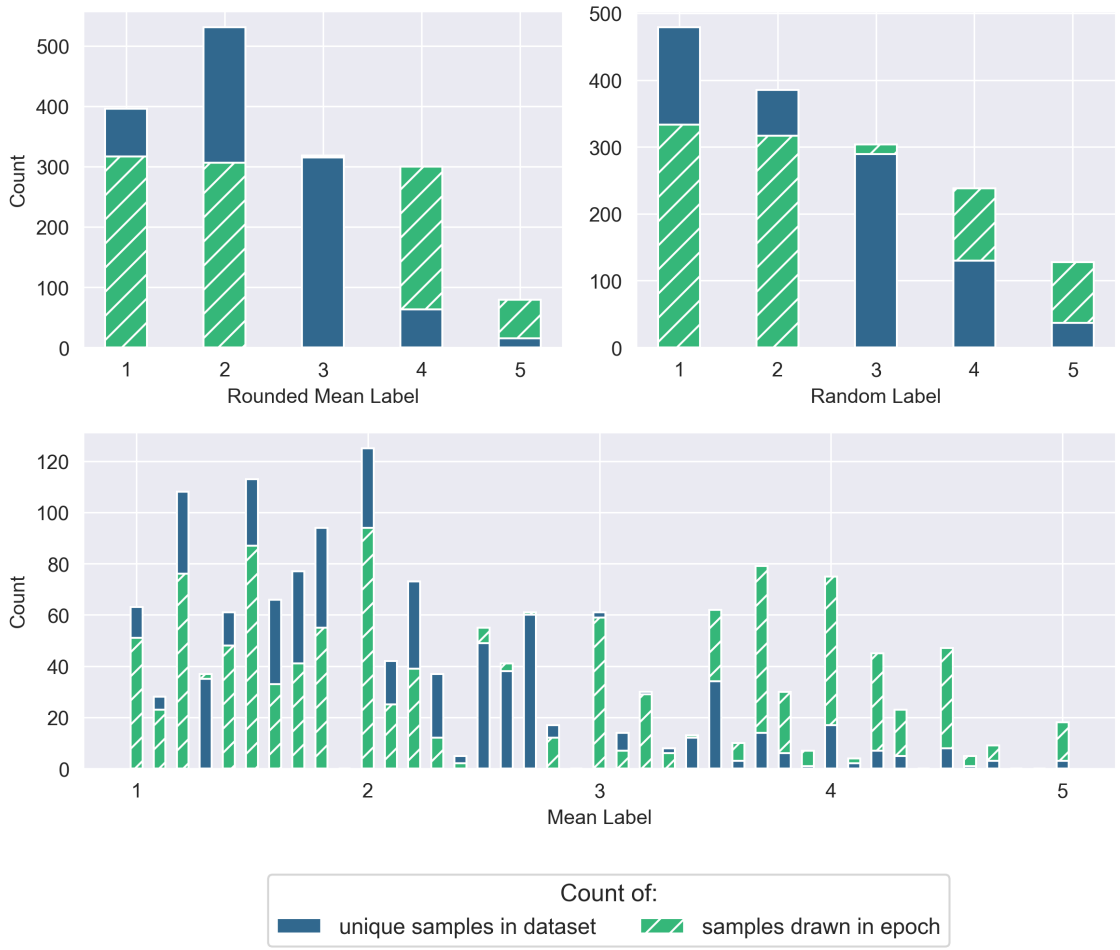


Figure 11: The distributions of ground-truth labels derived with differen label merging strategies from *StabLe*(Q1) for all 1320 samples in the training split. The distribution when each sample in the training split is drawn exactly one time is displayed in blue, and the distribution of one epoch where the `WeightedRandomSampler` drew the same number of samples is given in green (hatched).

5 Evaluations and Results

This section reports descriptive statistics on *StabLe* dataset described in Section 3.1 and Section 3.2, and the performance of the models described in Section 3.3.

5.1 Dataset Statistics

This section provides descriptive statistics on the assembled *StabLe* dataset.

5.1.1 Annotation Process

With the annotation web app (Section 3.2.3), user accounts for 38 raters were created. Ten of the raters were experts whom the SMI had reached out to, the rest were laymen. The labeling process started in August 2024 and ratings were collected until the end of December. Question Q1 was prioritized. Once the *StabLe*(Q1) dataset was completed, the ratings were analyzed and used to train the models. The exact progress of collecting ratings is shown in Figure 12. Figure 13 shows the contribution of individual raters to the total number of ratings in the dataset.

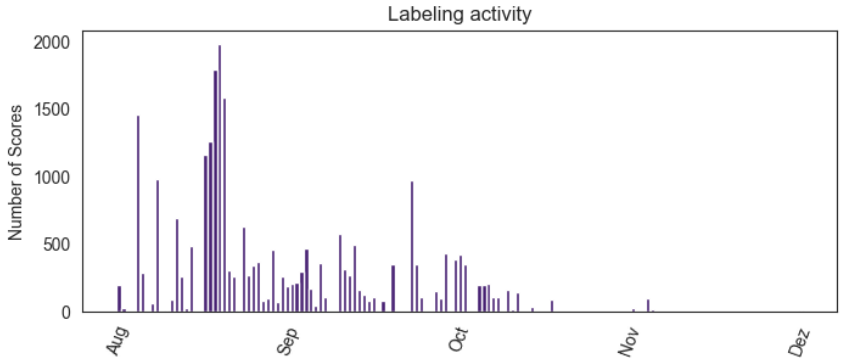


Figure 12: The number of ratings collected per day.

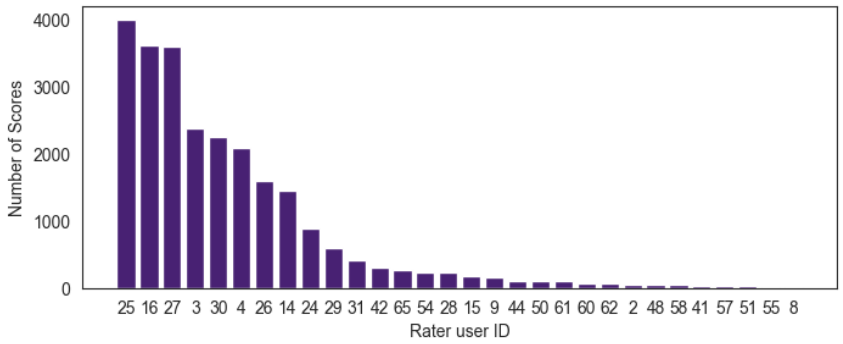


Figure 13: Contribution of individual raters to the total number of ratings collected.

5.1.2 Tagged Samples

As described in Section 3.1.5, all handwriting samples have been validated manually and tagged with additional information. The numbers of samples with each tag are shown in Table 19. Images of the sentences written by the students were also extracted and shown in the labeling app when the sensor data was not recorded properly. Therefore, the total number of tagged samples is shown as well as the number of tagged samples with recorded sensor data. The tags were used in the further analysis of the *StabLe* dataset and to verify model performance.

Table 19: Number of samples tagged according to the criteria described above.

Tag	Cursive	Typo	Correction
# All Samples	403	76	104
# Sensor Samples	369	70	98

5.1.3 Distribution of Ratings

Table 20 shows how often each rating was assigned to samples for each question. The distribution leans heavily towards the better ratings '1', '2' and '3' which corresponded to ratings of criteria as being fully, mostly, or sufficiently fulfilled. The ratings '4' and '5', which marked samples that exhibit deficiencies, make up only 5.5% to 12.9% of the ratings. This relative distribution of the ratings is shown in Figure 14, where the ratings were counted for each question and sentence. Not all sentences contained the letters 'a' and 'd', so question Q4 was not applicable to all ten sentences. The proportion of ratings provided by raters with experience in assessing handwriting is depicted. Expert ratings were collected for almost all questions and sentences.

Table 20: The distribution of ratings for the different questions.

Question	Rating '1'	Rating '2'	Rating '3'	Rating '4'	Rating '5'
Q1	3675	2971	2123	1043	254
Q2	2117	2181	1062	315	30
Q3	2390	1654	839	456	179
Q4	1541	674	284	111	35

The experts' ratings were marginally lower for Q1 and Q2 than the ratings of the non-experts, and marginally higher on Q3 and Q4. As described above, all handwriting samples have been validated manually and tagged with additional information. The samples written in cursive letters were perceived as less legible on average, indicated by a mean of 2.93 for the Q1 ratings compared to the mean of 2.13 for all samples. Spelling errors and corrections also significantly influenced Q1 ratings. This was also observed for the other questions, but there the effect was not as strong. The averages of the filtered ratings on subsets of *StabLe* are given in Table 21.

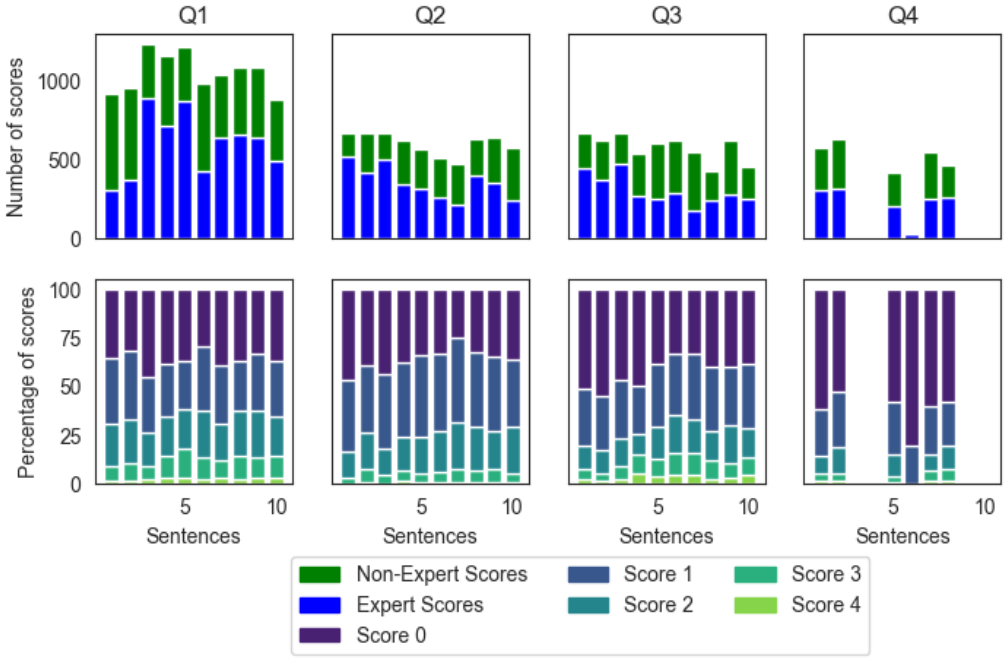


Figure 14: (top) The total number of ratings collected per question and sentence. The share of ratings given by experts and non-experts. (bottom) The relative distribution of ratings per question and sentence.

Table 21: Average ratings of different groups of raters and different data subsets.

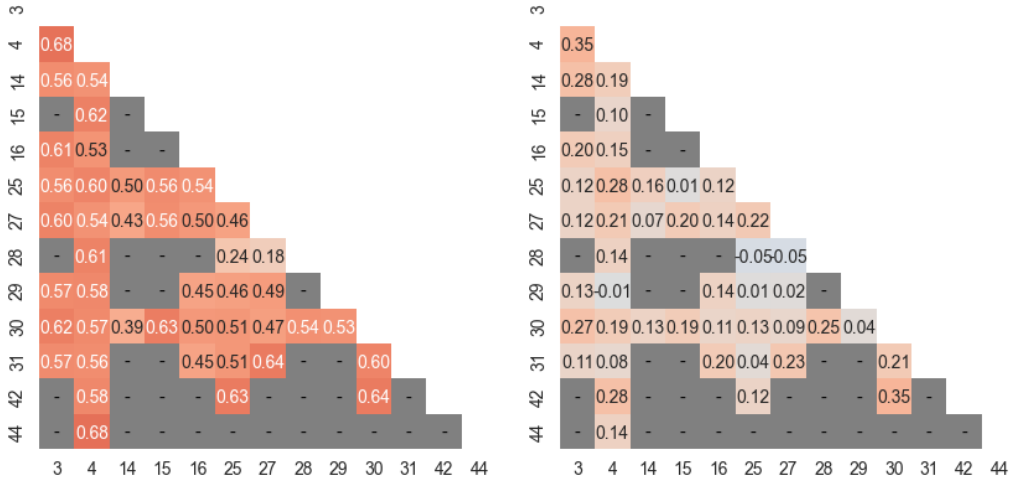
Question	All	Experts	Non-Experts	Cursive	Typo	Correction
Q1	2.13	2.11	2.16	2.93	2.65	2.83
Q2	1.95	1.83	2.11	2.00	2.10	1.99
Q3	1.99	2.01	1.95	2.26	2.09	2.31
Q4	1.65	1.67	1.63	1.72	1.75	1.64

5.1.4 Inter-Rater Reliability

To assess the inter-rater reliability, the variance of the ratings given to the same samples was examined. Furthermore, the ICC and Cohen's Kappa were calculated for pairs of raters who had rated the same samples. For each question, the variance of the ratings was calculated per sample. The averages of these variances per sample are shown in Table 22. For question Q1, the variance was calculated from five ratings per sample on average, while fewer ratings were collected for the other questions. The averages of variances were between 0.53 and 0.79, so ratings of the same sample and criterion tend to differ. This indicates low inter-rater reliability of the ratings in the *StabLe* dataset.

Table 22: The average number of ratings per sample and the average of variances calculated per sample.

Question	$\overline{Count}$	$\overline{Var}$
Q1	5.22	0.65
Q2	2.98	0.59
Q3	2.86	0.79
Q4	2.56	0.53



(a) (b)

Figure 15: (a) The pairwise ICC and (b) the pairwise Cohen's Kappa in *StabLe*(Q1).

The agreement was calculated between pairs of raters. A minimum of 50 samples had to be rated by both raters to examine their agreement on the ratings for those samples. Figure 15 shows all pairwise agreements of ratings in *StabLe*(Q1). According to the ICC, poor reliability (below 0.50) was found for 33 pairs and moderate agreement (0.50 to 0.75) was found for 11. Cohen's Kappa values indicate slight agreement (0 to 0.2) for 11 pairs and fair agreement (0.2 to 0.4) for 33. For the other questions, the reliability measured by the ICC is about 0.2 lower and is interpreted as poor. The values for Cohen's Kappa were lower as well, and indicated slight agreement. The averages of the pairwise agreements are given in Table 23. Following these measurements, the inter-rater reliability of the ratings collected with the annotations app is far below the renowned handwriting scales discussed in Section 2.1.5. All but one of the reviewed scales reported good reliability based on the ICC (above 0.75).

Table 23: The number of rater pairs who rated at least 50 samples for the same question and the mean agreement of those pairs.

Questions	# Pairs	$\overline{ICC}$	$\overline{Kappa}$
Q1	44	$.54 \pm .10$	$.15 \pm .09$
Q2	24	$.35 \pm .10$	$.12 \pm .06$
Q3	21	$.36 \pm .11$	$.11 \pm .06$
Q4	12	$.36 \pm .21$	$.16 \pm .10$

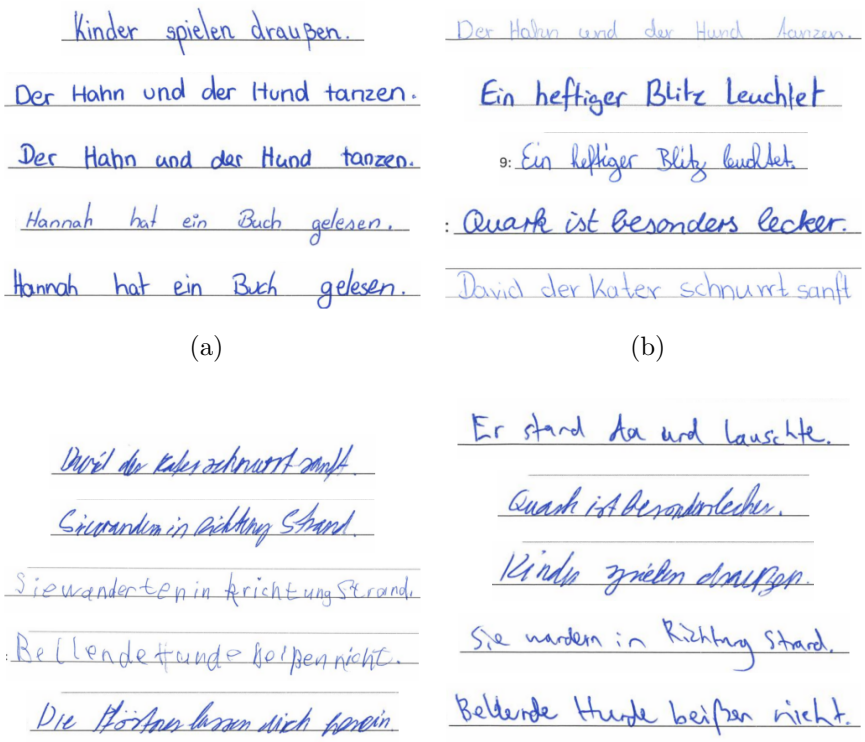
5.1.5 Intra-Rater Reliability

A subset of handwriting samples was displayed to the same raters twice. The ratings from the same rater for the same question and sample were compared to evaluate the intra-rater reliability of the collected ratings. For each question, the mean variance of all samples that were rated twice was calculated. This analysis included about 200 samples per question. For question Q1 this resulted in a mean variance of 0.65, for

Q2 it was 0.59, 0.79 for Q3, and 0.53 for Q4. These mean per-sample variances were comparable to the mean per-sample variances between different raters. This suggests that the low reliability of the ratings cannot be explained by different opinions on legibility by different raters alone because the ratings by the same rater tend to vary as much. The high variances showcase how hard it is to assess differences in legibility reliably.

5.1.6 Qualitative Inspection of Ratings

Figure 16 shows the samples that received the best and worst average ratings in *StabLe*(Q1) and *StabLe*(Q3). Although differentiating between ratings of the criteria was not done reliably (as shown above), the perception of very good or very poor samples seems to be reasonably accurate (a completely subjective observation by the author). The extremes for all four criteria are given in Section A.6.



(a) (b)

(c) (d)

Figure 16: (a) (c) The five best and worst rated handwriting samples in *StabLe*(Q1). (b) (d) The five best and worst rated handwriting samples in *StabLe*(Q3).

5.2 Results of the Comparative Experiments

This section reports performance on the training and validation set for models described in Section 3.3.2. The winning epoch and the corresponding best-performing model were selected based on performance on the validation set as described in Section 4.3.1.

5.2.1 Reproducing Results of Previous Work (Repr)

Model A replicated the CNN architecture used by Grabmann (2023) and was trained on the same *Curation Beauty* dataset with a changed approach to balancing. Figure 17 shows the learning curves of the model for accuracy and CE loss. The training accuracy increased steadily and reached 75% around epoch 100, similarly to the referenced model in previous work. The validation accuracy oscillated strongly from one epoch to the next, reaching peaks at 57% (refer to Table 24). Similar behavior was reported for the referenced model, where the validation accuracy reached a peak of about 55% but oscillated more strongly between epochs. So model A behaved similarly, but the results could not be reproduced exactly.

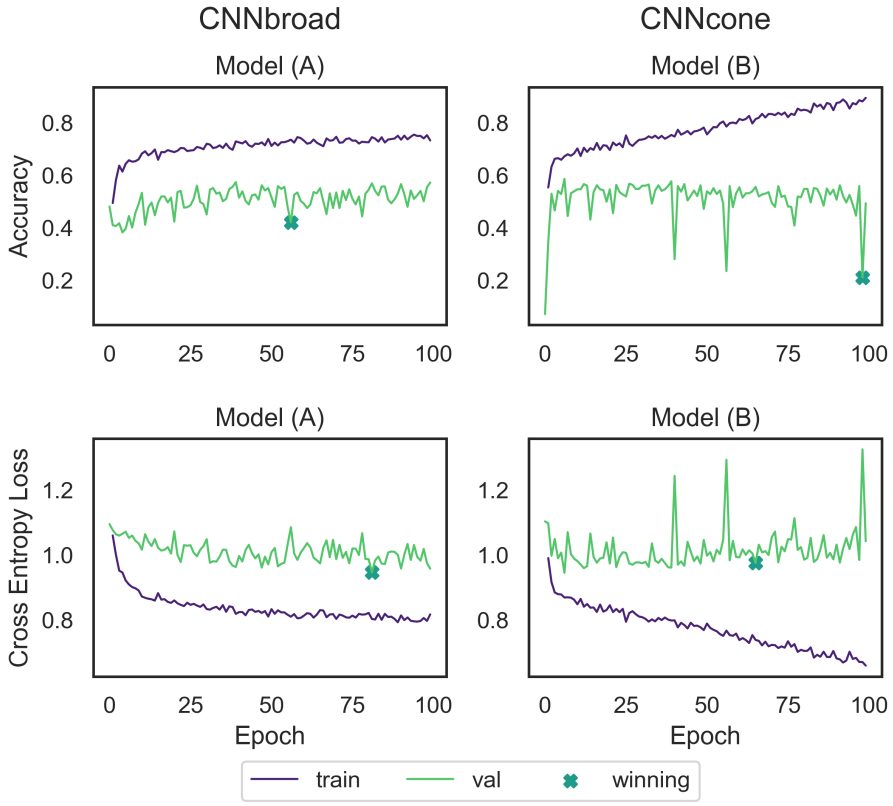


Figure 17: The learning curves of models A and B, which differ in CNN architecture.

5.2.2 Comparing CNN architectures (CnnArc)

The learning curves of models A and B are shown in Figure 17. The *CNNcone* architecture was able to better fit the training data. Both training accuracy and loss improved more steadily with model B. This indicates an improved learning capability with the adopted architecture, but the validation metrics revealed that this did not carry over to unseen samples. Model B overfits to the training data. The evaluation of the winning epoch models revealed only minor differences in performance. Model A beat model B with respect to accuracy on the validation set, while model B achieved higher mean precision and recall (Table 24).

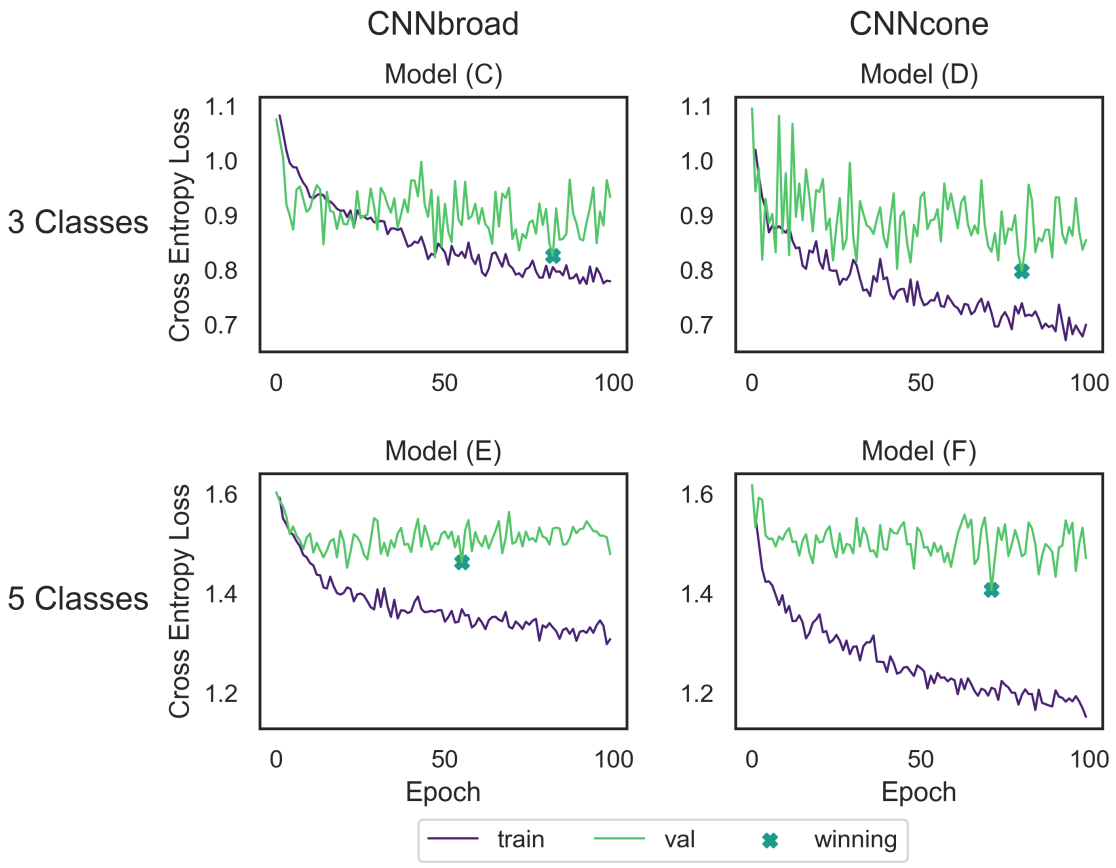


Figure 18: The learning curves of models C, D, E, and F, which differ in their CNN architecture and the number of legibility classes they were trained to predict.

The learning curves of models C and D are depicted in Figure 18. Again, the *CNNcone* architecture led to lower training losses. The validation loss did not follow the training loss and oscillated rather than exhibiting a clear downward trend for both models. An evaluation of the winning epoch models did not show a clear result. Model D performed slightly better with respect to CE, MSE, accuracy, and mean precision, while model C achieved slightly better mean recall and ICC (Table 24).

The learning curves of models E and F are shown in Figure 18. As expected, discriminating between five classes was harder than three classes, so the losses were

higher for the five-class models. This harder task seemed to highlight the improved learning capability of the *CNNcone* as model F performed better than model E with respect to all metrics (Table 24). Although most of the improvements were marginal, MSE and ICC improved substantially. These are the two metrics that do take the degree of error into account, so a false classification where the difference between the prediction and ground-truth label is higher weighs more.

It should be noted that the improvements were only marginal and do not allow a reliable conclusion. The fluctuation of the validation loss between subsequent epochs was greater than the improvement from one architecture to another. *CNNcone* was chosen as the CNN architecture for subsequent experiments due to its improvements in the five-class setting.

Table 24: Performances of winning epoch models on the validation set. Model performances showcase the effect of using different datasets and CNN architectures.

Model	CE	MSE	Acc	$\overline{Prec}$	$\overline{Rec}$	$\overline{ICC}$	$\overline{Kappa}$
A	.95	.57	.57	.40	.40		
B	.97	.57	.56	.44	.41		
C	.83	.42	.71	.48	.49	.24	.04
D	.80	.38	.75	.50	.47	.20	.04
E	1.46	1.26	.43	.39	.42	.21	.13
F	1.41	.75	.49	.43	.44	.41	.14

5.2.3 Comparing Datasets (CompDs)

Models A and C used the same *CNNbroad* architecture, but they were trained on different datasets. Model C was trained on *StabLe*(Q1) and performed better than model C, which was trained on the *Curation Beauty* dataset, with respect to all metrics. The same was found for the two *CNNcone* models B and D. The improvements were substantial. Between models A and C, the CE loss decreased by 0.12, the MSE by 0.15, the accuracy improved by 14%, the mean precision by 8% and the mean recall by 9%. Between models B and D, the CE loss decreased by 0.17, the MSE by 0.19, the accuracy improved by 19%, the mean precision by 6% and the mean recall by 6%. So, fitting to the training data generalizes better to unseen samples with *StabLe*(Q1) than with the *Curation Beauty* dataset.

5.2.4 Comparing Prediction Heads (HeadArc)

Model F used the single-layer classification head *1l class*. Its performance was compared with model G, which used the three-layer classification head *3l class* to examine how a more complex prediction head affects performance. Similarly, the two models H and I with regression heads *1l reg* and *3l reg* were trained and compared. The two classification models F and G were compared with the regression models H and I to examine which prediction task was better suited.

Table 25: Performances of winning epoch models on the validation set. Models were compared to assess how different prediction heads affect the performance.

Model	CE	MSE	$\overline{ICC}$	$\overline{Kappa}$
F	1.41	.75	.41	.14
G	1.43	.83	.37	.13
H		.48	.48	.15
I		.51	.43	.12

The learning curves of the four training runs are shown in Figure 19. The addition of linear layers and ReLU activations did not lead to lower losses. Table 25 displays the recorded winning losses and the agreement metrics. For classification and regression, the three-layer head performed slightly worse with respect to all metrics. The regression models outperformed classification with respect to MSE and agreement metrics.

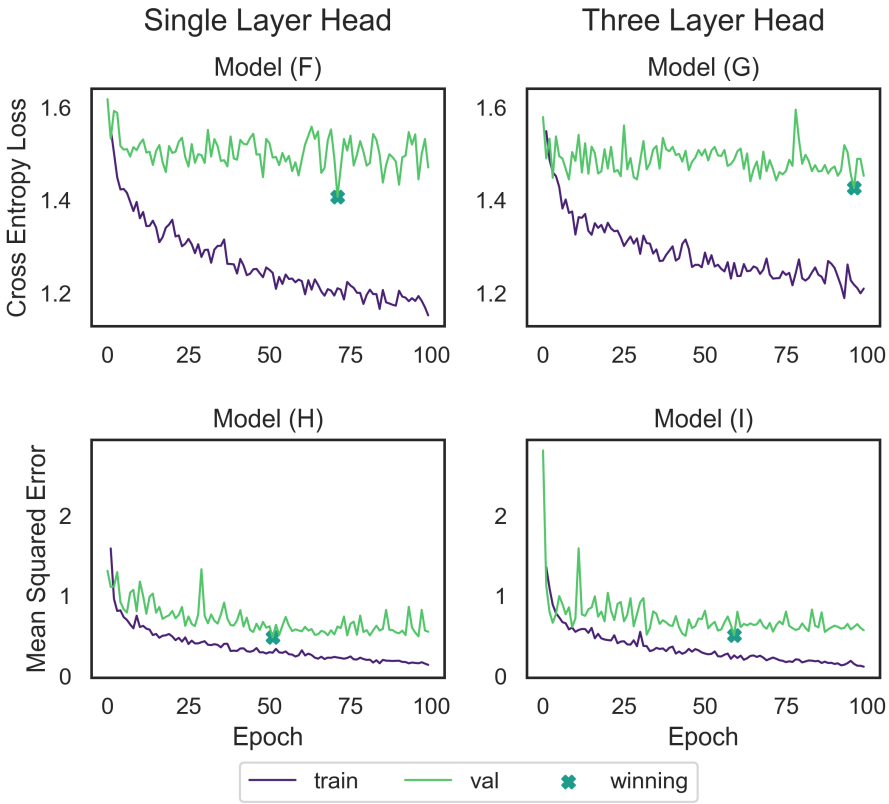


Figure 19: The learning curves of models F, G, H, and I. Models F and G were trained for classification while H and I were trained for regression. Models F and H used a single fully connected layer as the prediction head, while G and I used a head of three layers with ReLU activation functions in between (Section 4.3.2).

5.2.5 Comparing Label Merging Strategies (LabMerg)

The five models H, J, K, L, and M were trained with different label merging strategies. The corresponding learning curves are shown in Figure 20.

As shown in Table 26, the model trained on *majority-labels* (J) performed worse than model trained with *rounded-mean-labels* (H) with respect to all metrics. Model K, which was trained on *random-labels*, achieved the worst MSE of 0.85. This approach introduced the uncertainty found in the ratings into the training process, whereas this uncertainty is hidden behind deterministically derived labels for the other models. This makes the prediction task inherently harder than the other four, as indicated by the training loss, which stayed higher with random labels. Compared to the other label merging strategies, the training and validation loss did not diverge in the later epochs. Despite the higher MSE, model K outperformed other models with respect to Cohen's Kappa and shared the first place with the *mean-label* model (L) with respect to the ICC. Here, it should be noted that the performance of model K depends on random selection at runtime. For experiments to be reproducible, random seeds were set to be the same for all training runs.

Table 26: Performances of winning epoch models on the validation set. For models H, J, K, and M. The models were compared to assess how different strategies of merging the ratings of samples affects performance.

Model	MSE	$\overline{ICC}$	$\overline{Kappa}$
H	.48	.48	.15
J	.70	.41	.09
K	.85	.49	.18
L	.35	.49	.16
M	.76	.43	.10

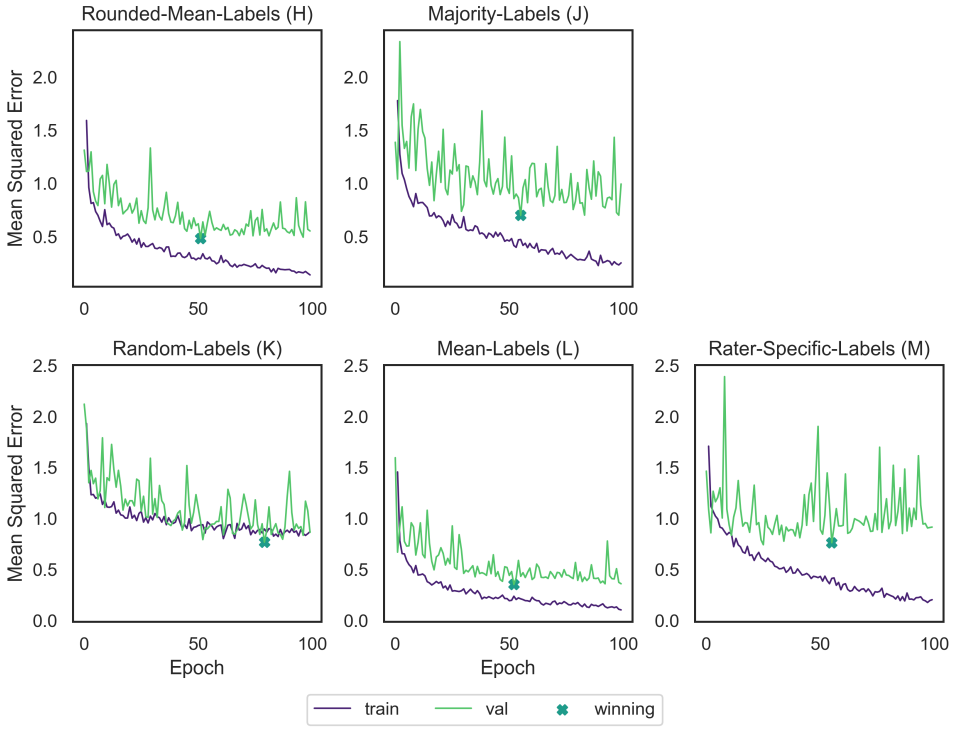


Figure 20: The learning curves of regression models H, J, K, L and M, which were trained with the different label merging strategies described in Section 3.3.

Model M, which was trained with *rater-specific-labels* (the ratings of rater '4'), achieved similar results as training on the majority labels. The model overfitted the training data, as indicated by the high validation loss. The agreement on the validation set between the model and rater '4' measured with the ICC was 0.42, which was lower than the average ICC of 0.43 between the model and all raters.

Model L, which was trained with the floating point mean of ratings for each sample, performed best with respect to all metrics. Training on randomly chosen ratings as labels seemed to counteract overfitting. Therefore, models L and K were chosen for further experimentation.

5.2.6 Hyperparameter Tuning (HypPar)

Models K and L from previous experiments were trained with different learning rates, kernel sizes, and regularization parameters to examine how tuning these hyperparameters affected performance.

Learning Rates Previous models were all trained with a learning rate of 0.001. In this experiment, the seven learning rates 0.0001, 0.0003, 0.001 to 0.1 were tested for both models K and L. Similar performances were found for both label merging strategies. The learning rates of 0.001 and 0.003 were among the best when testing for low MSE and performed the best with respect to the ICC and Cohen's Kappa. For the following experiments, a learning rate of 0.001 was chosen.

Kernel Sizes Previous models all used a one-dimensional kernel with a size of three. Six different kernel sizes from three to 41 were tested. The padding of the convolutional layers was increased accordingly to keep the data sizes consistent. For model L a kernel size of seven achieved the best MSE of 0.33 on the validation set, beating the MSE of 0.35 with kernel size three. For model K the kernel size of 21 achieved the best MSE of 0.79 on the validation set, beating the MSE of 0.85 with kernel size three. With respect to ICC and Cohen's Kappa the models with kernel size three beat all other variants. A kernel size of three was chosen for the following experiments.

Regularization Parameters Except the random labels model K, all models overfitted the training data. Five variants of model L were trained with the six values of 0.0, 0.0001, 0.001 to 0.1 for the regularization term λ, which determines how strongly L2-Regularization is applied. The learning curves were inspected to see if the regularization prevented overfitting. Both training and validation loss behaved similarly for all tested parameters. The best MSE on the validation set was 0.33 and just below the worst MSE of 0.37. ICC and Cohen's Kappa were only marginally affected by regularization as well. Since the experiment did not reveal a clear benefit in introducing regularization, the following experiments did not apply L2-Regularization.

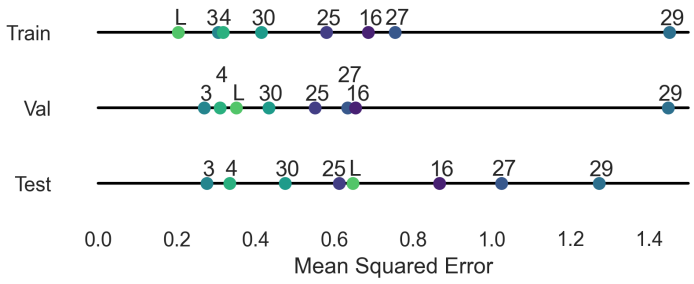
5.3 Results of the Evaluating and Verifying Experiments

This section reports performance on the training, validation and test set for models described in Section 3.3.3. The winning epoch of a training run was selected based on the validation set (Section 4.3.1). The corresponding model was then evaluated.

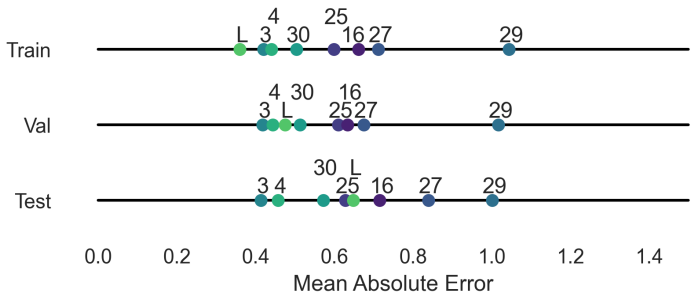
5.3.1 Evaluation of Model L

The best performance on the validation set was achieved by model L, which was trained for regression on *mean-labels* derived from *StabLe*(Q1). Here its performance was evaluated further.

Figure 21 shows the MSE and the MAE the model achieved on the three data splits. To put model performance into perspective, the mean errors between individual raters and ground-truth labels are given. Model L achieved the lowest mean errors on the training data. As it was trained with MSE as the loss function, this reflects the low training error observed in the learning curves. On the validation set, it achieved the third lowest errors. With the test set the model lies in the middle. The MAE on the test set shows that the model on average predicts labels that are 0.65 off the target *mean-label* for the 196 samples in the test set.



(a)



(b)

Figure 21: (a) The MSE and (b) MAE between individual raters (or the model) and the *mean-labels* of the training, validation and test split of *StabLe*(Q1). Includes model L and raters who rated at least 50 samples in each of the three data splits.

The raters contributed to the ground-truth labels of the samples, so that the raters who rated most of the samples tend to have low mean errors. For example, rater '4' rated all samples for question Q1 and is among the lowest mean errors.

In Figure 22 model predictions are plotted against the corresponding ground-truth labels for the different data splits. A line was fitted to the data points by minimizing the sum of squared residuals (SSR) to visualize the linear relationship between targets and predictions. The model fit the training data well, as indicated by the almost diagonal line (slope = 0.93). In the validation set, this relation was weaker, as indicated by the less steep line (slope=0.60). In the test set, the slope dropped to 0.37. This shallow slope of the best-fit line indicates the low discriminative power of the model for unseen samples. The mean errors on the test set show that the model was closer to the average rating of unseen samples than three raters. On the test set, model L achieved poor agreement with respect to a $\overline{ICC}$ of 0.34 and a slight agreement indicated by a $\overline{Kappa}$ of 0.08.

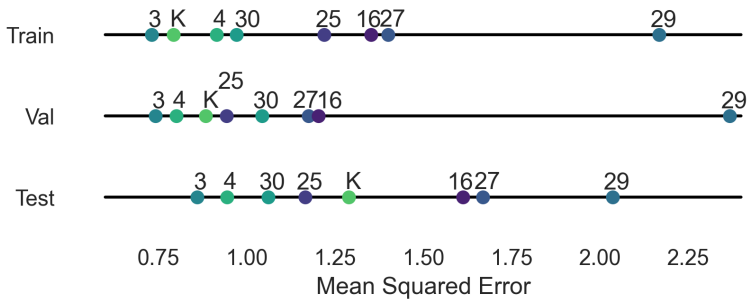
Figure 22: Predictions of model L plotted against the ground-truth *mean-labels*. The best-fit lines were fitted to minimize SSR.

5.3.2 Evaluation of Model K

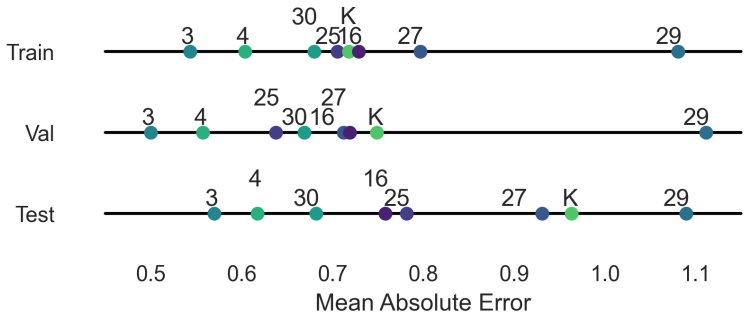
Model K was trained on *random-labels* picked from ratings in *StabLe*(Q1) at runtime. Having varying labels per sample in different draws made this training objective more difficult than training on deterministic *mean-labels* as done for model L. This was reflected in the higher MSE of 0.82 on the the training set compared to model L with a MSE of 0.20. On the validation set, the MSE increased minimally to 0.85, and then jumped to 1.35 on the test set.

Figure 23 shows the mean errors between model K and the ground-truth labels drawn from the data splits, as well as the mean error between the raters and the ground-truth labels. The MSE of model K behaved similarly to model L with respect to its ranking among raters. Based on the MAE model K ranked worse than model L with the second highest error on the test set. Plotting the predictions against the ground-truth labels indicated a weak linear relationship (Figure 24). The slope

of the best-fit line is 0.52 on the training set and drops to 0.30 and 0.15 on the validation and test sets. On the test set, model K achieved poor agreement with respect to a $\overline{ICC}$ of 0.29 and slight agreement indicated by a $\overline{Kappa}$ of 0.04.



(a)



(b)

Figure 23: The MSE (a) and MAE (b) between individual raters (or the model) and the *random-labels* of the training, validation, and test split of *StabLe*(Q1). Includes raters who rated at least 50 samples in each of the three data splits and model K.

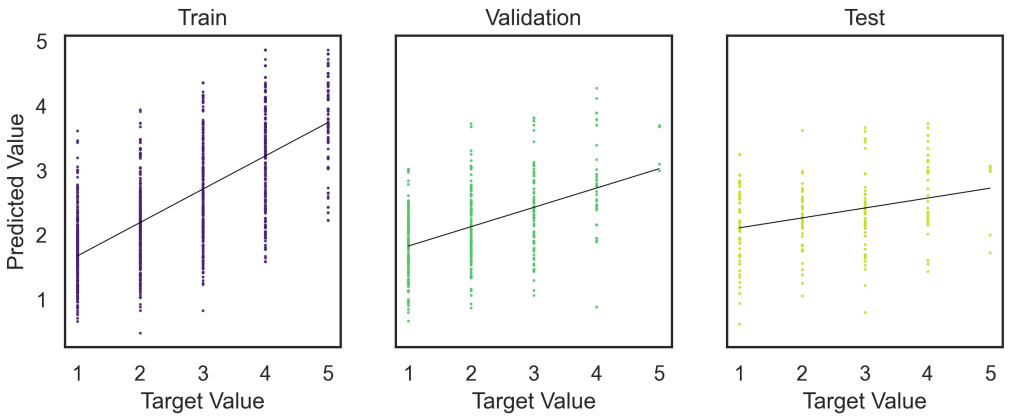


Figure 24: Predictions of model K plotted against the ground-truth *random-labels*.

5.3.3 Evaluation of Model M

Model M was trained with *rater-specific-labels* corresponding to the ratings of the rater with user ID '4' in *StabLe*(Q1).

On the validation set, it performed worse the models L and K, which used ground-truth labels derived from several raters. Figure 25 opposes the ratings, which are the targets, and the predictions of the model. Compared to model L, the best-fit lines indicate a weaker linear relationship between targets and model predictions, with a slope of 0.76 on the training, 0.36 on the validation, and 0.17 on the test split. The agreement with rater '4' was poor as indicated by the ICC of 0.25 and slight with a Cohen's Kappa of 0.03.

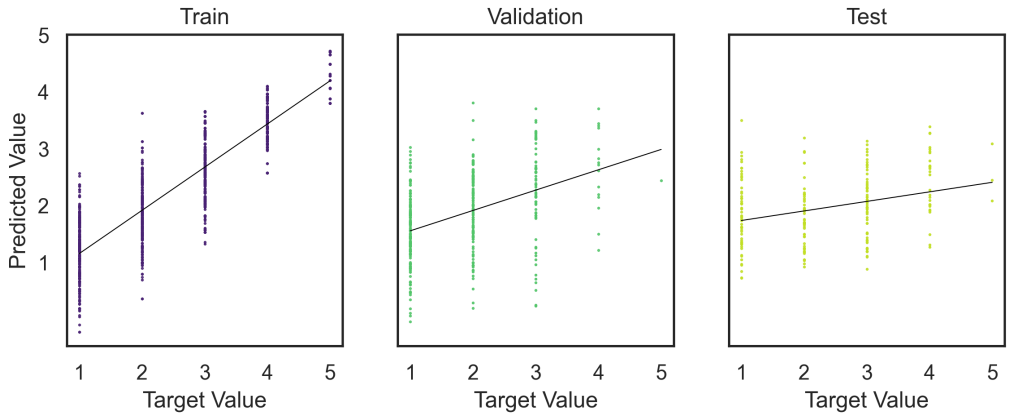


Figure 25: Predictions of model M plotted against ground-truth *rater-specific-labels*.

5.3.4 Evaluation of Model L*

Rater '4' is the benchmark rater for the evaluation of model L*. The benchmark rater gave ratings for all samples of *StabLe*(Q1), so these ratings co-determined the *mean-labels* used to train model L. Model L* was trained on *mean-labels* of *StabLe*(Q1)*, which contained ratings of the six training raters and excluded ratings of the benchmark rater.

Consequently, the MAEs between the ground-truth labels and the ratings of the benchmark rater were lower (0.44 - 0.46 in Figure 21) on *StabLe*(Q1) and higher (0.55 - 0.57 in Figure 26) on *StabLe*(Q1)*.

Similarly, the agreement between the benchmark rater and model L was higher than with model L* as shown in Table 27. The agreement between the benchmark rater and model L* shows how well the model predictions align with the ratings of an unknown rater. The average of the agreement between the benchmark rater and the training raters served as baseline. On the training set, model L* was in greater agreement with the benchmark rater (ICC of 0.67), than the individual training raters were on average (mean of pairwise ICCs of 0.59). On the validation set, the

agreement of model L* and the benchmark rater decreased to 0.38, which was lower than the mean agreement of the training raters and the benchmark rater (0.52). This went hand in hand with the mean errors of the model increasing from the training to the validation set (see Figure 26). This discrepancy was the greatest on the test set, where the benchmark rater agreed more with the training raters (average of pairwise ICCs of 0.64) than with the labels predicted by model L* (ICC of 0.31). So, the agreement between the benchmark rater and the training raters was significantly higher than the agreement between model L* and the benchmark rater. This finding was in tune with the weak linear relationship of predicted and target labels on the test set observed above.

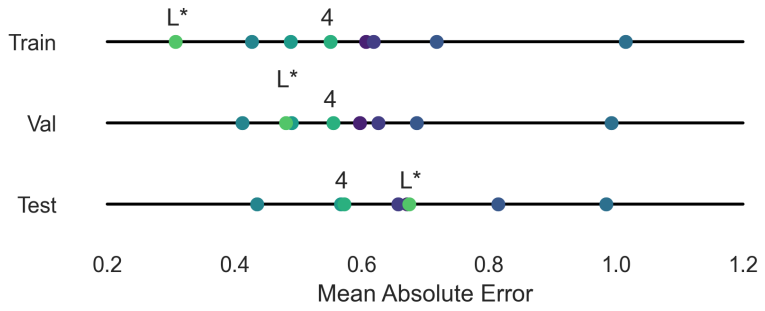


Figure 26: Mean absolute errors between raters (and model L*) and the ground-truth labels of *StabLe*(Q1)*, which excludes ratings from rater '4'.

Table 27: The agreement (ICC) with the benchmark rater '4' on samples of the three data splits. Ratings from the benchmark rater were compared to predictions from model L and L*, and to ratings from the six training raters. The mean of pairwise agreements between the benchmark rater and each of the training raters is given as $\overline{\text{Raters}}$.

	L	L*	3	16	25	27	29	30	$\overline{\text{Raters}}$
ICC (Train)	.71	.67	.68	.53	.60	.54	.58	.58	.59
ICC (Val)	.50	.38	.68	.59	.48	.52	.42	.46	.52
ICC (Test)	.40	.31	.76	.70	.68	.47	.64	.58	.64

5.3.5 Evaluation of the Legibility Criteria

The four models L, N, O, and P were each trained with *mean-labels* of the four datasets *StabLe*(Q1) to *StabLe*(Q4) respectively. The corresponding learning curves are shown in Figure 27. All models were able to fit to the training data, reducing the MSE to between 0.10 and 0.21 on the training set. For Q1 the best MSE increased to 0.35 on the validation set. For the questions Q2, Q3, and Q4, the validation MSE was higher and oscillated between values already achieved in the first few epochs. Similarly, on the test set the lowest MSE was achieved for Q1.

For Q1 the $\overline{ICC}$ between the model and the raters decreased from 0.49 on the validation set to 0.34 on the test set. For the other questions, the $\overline{ICC}$ was substantially lower on the validation set and decreased further in the test set. A similar picture emerged for the values of Cohen's Kappa. For Q2, Q3, and Q4, the rater's agreement was lower than for Q1 as shown in Table 23. Similarly, the agreement between the model and the raters was the highest for Q1 on all data splits, and low for the other questions (Table 28).

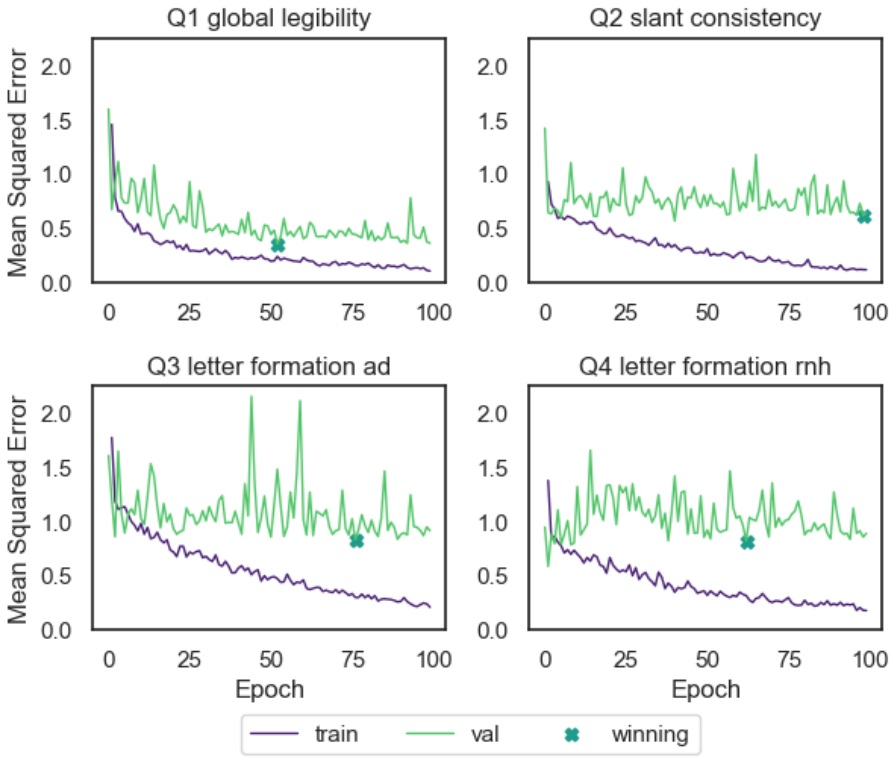


Figure 27: The learning curves of the models L, N, O, and P, which were trained on ratings corresponding to the four questions Q1, Q2, Q3, and Q4.

Table 28: Lowest MSE achieved on the data splits and the agreement between the best-performing model and raters on validation and test set for models L, N, O, and P, which were trained on ratings from questions Q1, Q2, Q3, and Q4, respectively. Mean agreement of raters within the corresponding data sets and splits is given in brackets where enough data was available.

Question, Model	MSE Train	MSE Val	MSE Test	$\overline{ICC}$ Val	$\overline{Kappa}$ Val	$\overline{ICC}$ Test	$\overline{Kappa}$ Test
Q1, L	.10	.35	.65	.49(.50)	.16(.14)	.34(.58)	.08(.14)
Q2, N	.11	.61	.75	.10(.32)	.02(.09)	.02	-0.04
Q3, O	.21	.83	1.18	.11(.28)	-0.01(.09)	.09(.36)	-0.02(.12)
Q4, P	.17	.82	.88	08	.03	.05	-0.03

5.3.6 Evaluation of Uncontrolled Variables

Three variations of model L were trained, where samples with cursive writing, spelling errors, or with corrections were excluded from *StabLe*(Q1). Table 29 shows that the removal of the cursive samples strongly decreased the performance of the model. The $\overline{ICC}$ on the test set fell from 0.34 with all samples to 0.17 without the cursive samples. Removing the the few samples with spelling errors did not affect the agreement and removing the corrections resulted in a decreased $\overline{ICC}$.

Table 29: Agreement on the test set of four models trained on different filtered variants of *StabLe*(Q1) and the number of samples in those subsets.

Model	# Samples	$\overline{ICC}$	$\overline{Kappa}$
L	1907	.34	.08
$L^{\setminus \text{cursive}}$	1537	.17	.09
$L^{\setminus \text{typo}}$	1837	.34	.08
$L^{\setminus \text{correction}}$	1809	.28	.09

6 Conclusion

The presented work produced the *StabLe* dataset, which allowed to examine how the legibility of handwriting was perceived by different raters. The data set comprises images of individual handwriting samples, sensor data recorded while writing, information on the produced writing like the presence of spelling errors, and lastly ratings on its legibility according to four different criteria. The rater agreement was analyzed for each criterion and compared with the agreement reported in related work (**R1**). The performance achieved with the *Curation Beauty* dataset and the model architecture used in the corresponding work was measured and interpreted (**R2**). The models were trained with samples of *StabLe* to see whether reducing the variety of written texts helped the models to pick up legibility related features (**R3**). The dataset was used to examine whether the reliability was higher for the legibility criteria that appeared to be more specific (**R4**). Accordingly, it was tested whether the models could predict more specific criteria more accurately (**R5**). Performances were compared for models trained with different label merging strategies to explore how the uncertainty between raters can be addressed when training supervised models (**R6**). In the following, these questions are addressed and answered based on the results of this work as far as possible.

6.1 Rater Agreement in the *StabLe* Dataset

R1 *Is the rater agreement found in the StabLe dataset comparable to the agreement reported in related research?*

Results For each handwriting sample and each of the four legibility criteria, ratings were collected from two to six raters. With these redundant ratings, the raters' agreement was measured (Section 3.2.4). In studies on legibility assessment, the ICC was the most commonly reported metric on inter-rater reliability and served as the best option to compare rater agreement across different studies. In the *StabLe* dataset the highest pairwise ICC of 0.68 was found for raters '3' and '4' on the ratings for *global-legibility*. Averaged among the pairs of raters, the highest $\overline{ICC}$ of 0.54 was found for question Q1 as well. Following the interpretation of ICC values, that means that the two raters with the highest agreement fall into the higher range of moderate agreement. The mean of pairwise agreements falls into the lower range of moderate agreement with poor agreement detected for 33 out of 44 pairs of raters. For the other legibility criteria, the agreement was found to be poor across the board. For the two questions Q3 (*letter-formation_rnh*) and Q4 (*letter-formation_ad*) a $\overline{ICC}$ of 0.36 was measured. For question Q4 (*slant-consistency*) the lowest $\overline{ICC}$ of 0.35 was found.

Interpretation For the handwriting scales reviewed in Section 2.1 the reported inter-rater reliability of either the total scores or per-item ratings laid between 0.39

and 0.99. The agreement found in the *StabLe* dataset is comparable to the agreements reported for the ratings of the mETCH handwriting scale and is significantly lower than the agreements reported for the other scales. To conclude these findings, the measured agreement could not compete with that reported for renowned handwriting scales.

6.2 Reproducing Results on the *Curation Beauty* Dataset

R2 *Can the results of previous work on automated legibility assessment be reproduced and are the results meaningful?*

Results As described in Section 2.2.3, the work by Grabmann (2023) was understood to represent the status quo for predicting the legibility of handwriting from the writing of movement sensor data. Accuracies obtained by five-fold cross-validation laid between 37.77% and 45.54% for different deep learning models. Of the four models, the described CNN architecture was recreated and trained on the *Curation Beauty* dataset to verify the results. Despite small differences in the training setup (balancing), the reported and recreated training runs appeared similar. The highest validation accuracies of 55% and 57% seemed comparable.

Interpretation To conclude, the reported results could not be recreated exactly, but were found to be reasonably similar given the differences in the training runs.

The reported accuracies were obtained on a highly unbalanced dataset. This is believed to diminish the expressiveness of the results. Furthermore, the strong test accuracy oscillation in the previous work indicates that the model failed to generalize from training samples to unseen samples. If weight adjustments from a single epoch lead to higher differences in validation or test metrics than on the training set, this indicates that the model is probably not learning generalizable features. Instead, it is assumed that the model mainly memorizes the training samples, which by chance leads to more accurate predictions on unseen samples in one epoch and to lower accuracy in another.

Looking at an application of the described models to assess how legible students write, it must be concluded that the models, which represent the status quo, are far from being able to approximate ratings given by humans sufficiently.

6.3 Effects of Focusing on a Reduced Reference Text

R3 *Does reducing the variety of texts in the dataset help models to find patterns that are related to legibility?*

Results Two pairs of models were trained, where both models of a pair were the same except for the dataset they were trained on. One model was trained on the *Curation Beauty* and the other one on the *StabLe* dataset. This allowed for a comparison between a dataset with a wide variety of texts and one with only ten reference sentences. In both pairs, the model trained on *StabLe* performed better on all tracked metrics.

Interpretation From these observations, it was concluded that the ratings in the *StabLe* dataset are more closely correlated with the patterns in the sensor data than is the case for the ratings and sensor data of the *Curation Beauty* dataset. One key difference between the two datasets was that *StabLe* consists of samples corresponding to a fixed set of reference texts. It is assumed that the reduced variety in texts also reduced the variety in the sensor signals, so that the differences between samples are smaller and related to the level of legibility more closely. This suggests that reducing the variety of texts helped the models to find patterns related to legibility.

6.4 Reliability of Different Criteria

R4 *Is the rater agreement higher on the criteria that are assumed to be more specific?*

Results With the *StabLe* dataset, the hypothesis was tested that fine-grained criteria are assessed more reliably than broader criteria. The highest and lowest per-sample variances were found for the two criteria of *letter-formation_rnh* and *letter-formation_ad*. These criteria were both believed to be very specific because the raters were instructed to inspect specific letters in detail. The variance measured for the least specific criterion of *global-legibility* was between the variances of the specific ones. With respect to ICC, a fair agreement was found among raters for *global-legibility* on average, and the agreement was poor for the other three criteria.

Interpretation The per-sample variance did not exhibit the assumed relationship, and the ICC agreement values contradict the assumption as well. Consequently, it cannot be said that the criteria that were said to be more specific were rated more reliably.

6.5 Prediction of Different Criteria

R5 *Do models perform better in assessing criteria that are assumed to be more specific?*

Results Similarly to the reliability of the ratings, it was assumed that the models would perform better in predicting the labels of *StabLe*(Q3) and *StabLe*(Q4), because those criteria are determined by the shapes of specific letters. How well-formed these shapes are in a handwriting sample was believed to be indicated by patterns in the sensor data. Following this intuition, the models should be able to predict the corresponding labels. The same argument suggested that detecting the *slant-consistency* would be feasible from the sensor data. The perceived *global-legibility*, on the other hand, seemed to be less closely related to specific writing movements, so that identifying corresponding features in the sensor data would pose a more challenging task.

The experiments did not show success in predicting the labels for questions Q2, Q3, and Q4. For the least specific question Q1, models were able to learn the labels to some degree, but these results had shown to be mainly a product of discriminating between cursive and block letter writing styles.

Interpretation The results do object the assumption that models would perform better on the more specific criteria. Due to the poor performance of all trained models, a clear conclusion about which criteria can best be predicted from sensor data cannot be drawn.

6.6 Effects of Label Merging Strategies

R6 *How can the uncertainty inherent in assessing legibility be addressed when training supervised models?*

Results The different ratings of each sample were used to derive ground-truth labels for the training of supervised models. Using the mean of the ratings yielded the best performance regarding MSE and performed second best with respect to rater Agreement. Rounding the mean ratings to integer labels led to worse performance and using the majority vote performed worst.

Training with the ratings of one individual rater was assumed to allow the model to capture the rater-specific perception of what makes handwriting legible. However, the rater-specific model achieved lower agreement with the rater it was trained on than the model trained on the mean ratings.

Lastly, a model was trained with labels randomly chosen from the available ratings of each sample. As expected, the training loss showed that this prediction task was more challenging because labels were not deterministic and, therefore, could not be memorized for individual samples.

The training and validation loss did not diverge in later epochs as it did with differently derived labels. This suggests that the harder training task offers less opportunity for the model to memorize and overfit the training data. In other words, the features learned on the training data seemed to generalize to unseen samples better.

Training on random labels led to the highest agreement with the raters on unseen samples. Thus, introducing the uncertainty of legibility ratings into the training process benefited model performance with respect to agreement with a group of raters.

Interpretation These findings indicate that the average of the raters' opinions presented the most consistent approximation of what legibility is.

The ratings of an individual rater seemed to be internally inconsistent because the model was unable to detect patterns that correlate with the ratings. The ratings from a single rater might tend to vary due to unwanted effects during the rating process. For example, the perception of when a sample is not legible could change depending on how many non-legible samples the rater has seen recently. After grading many samples that were perfectly legible, the judgment could have shifted to be more rigorous when the next lower legible sample comes along. Such effects would decrease the internal consistency of ratings from a single rater and were assumed to be counteracted by using the average of several ratings

Training with randomly drawn ratings is believed to introduce the degree of uncertainty inherent in the ratings into the training process, acting as a way of regularization that benefits the generalizability of learned features.

7 Limitations

7.1 Low Comparability of Reported Rater Agreements

Comparing ICC values reported in different studies does not tell the whole story. The ICC values of the reviewed handwriting scales (Table 3) were obtained for different measurements. For the HLS the given ICC refers to the agreement between raters with respect to a differentiation between three legibility classes derived from the overall score ranging from zero to 25. For SOS-2 the ICC was calculated directly on the total score ranging from zero to 12. These differences limit the comparability of the ICC values and prohibit a final conclusion on whether the agreement was lower in *StabLe* than in the related work (**R1**).

7.2 Confounding Variable in Comparing Datasets

The comparison model performances obtained on two different datasets suggested that focusing on a set reference text benefited the model with respect to learning generalizable features for determining legibility labels. However, another difference between the two datasets was the rating scheme. The ratings of *Curation Beauty* presented a count of different violations, the ratings in *StabLe* state the legibility directly. It seems plausible that these direct ratings were more strongly correlated with patterns in the data than the counts, which could result from different combinations

of violations that each show differently in the sensor data. This second uncontrolled variable reduces the expressiveness of the presented results to tell whether reducing the variety of texts was the reason for the improvement of model performance (**R3**).

7.3 Limited Comparison of Agreement on Criteria

The lowest variance in ratings was found for question Q1, for which the highest number of ratings per sample was collected. Consequently, the expressiveness of comparing the variances and agreement metrics between the questions was decreased by different and rather small sample sizes (two to six ratings per sample). Another possible reason for the low agreement found for more specific criteria are insufficient instructions. Although most raters are assumed to have a notion of what handwriting they perceive as legible, the same does not hold for the other criteria, which are probably less natural. The criterion of *global-legibility* refers raters to their own opinion of what writing is legible, which in turn means that the instruction does not affect the ratings that much. For the other criteria, instructions are needed to sensitize the raters to actively look for the specified characteristic of the handwriting. Following this argumentation, insufficient instructions would affect how individual raters understand the more specific criteria more strongly. Each sample was rated by a potentially different subgroup of the raters. So, the reported variances were probably influenced by the random assertion of the raters. These possible causes for different levels of rater agreement impede a definitive conclusion on how the reliability of ratings and the specificity of rated criteria are related (**R4**). As models were trained based on these ratings, limitations affect conclusions on how well models can predict criteria of different specificity as well (**R5**).

7.4 Uncontrolled Variables Affecting Evaluation

Considering how much of the performance on *StabLe*(Q1) originated from discriminating between cursive and printed typefaces, it seems plausible that the only features that the model was able to pick up were the characteristics of the writing process, but not the product. Whether a sample was written in cursive or print letters can likely be detected from the frequency with which the pen is lifted off the paper or the number of stops and starts with regard to accelerations. Longer pauses also seem likely to be correlated with low legibility. These characteristics are all only indirectly related to legibility in that they could not be obtained by looking at the handwriting product itself. Legibility, as defined for the scope of this work, is understood as a quality of the handwriting product. Following this explanation for the observed model performances, it must be concluded that the models were probably unable to detect the features of legibility itself. The best models are believed to have identified features of the handwriting process that were correlated with low legibility. The findings question results of previous work, where factors like writing style and the frequency of spelling errors and corrections were not examined. This

highlights the importance of controlling for independent variables such as gender, handedness, and writing style to evaluate whether models actually detect legibility.

7.5 Unbalanced Data Affecting Evaluation

A main limiting factor throughout this work was that few samples with low legibility were available. This shortage gave models less opportunity to find patterns in the sensor data that correlate with low legibility. Issues regarding the expressiveness of metrics in light of a highly unbalanced test set were identified in previous work but were not overcome in this work. MSE, MAE, ICC and Kappa weigh errors on samples with low ratings equally as errors on samples with high ratings. *StabLe* comprised more samples with low ratings (legible) than with high ratings (illegible). This was counteracted by balancing the training set, but the test set remained unbalanced. Therefore, being able to produce correct predictions for the many samples with low target labels benefited the named metrics more than being able to produce correct predictions for the few samples with high target labels. Especially with an outlook towards using such models for diagnostic purposes, where identifying weak students (high ratings) is the goal, this is a problem.

7.6 Improper Use of the Validation Set

In the comparative experiments described in Section 3.3.2, the validation set was used for the model selection within training runs. The model with the weights from the epoch with the lowest validation loss was then evaluated on the validation set again to obtain metrics like the ICC. This drastically decreased the meaningfulness of the reported metrics because those were no results on unseen data, but results on data the model was cherry-picked for among the models of all other epochs. In hindsight, using the mean of validation metrics in the last few epochs or even just the last epoch would have allowed for a better approximation of how the model would later perform on the test set than the described practice. Instead, performances were believed to be reasonably close to training performances based on the cherry-picked validation metrics, just to be undercut by the results on the test set. Cherry-picking models for validation had forestalled realizing how low the generalization from the training data to unseen data was.

8 Future work

8.1 Increasing the Comparability of Rater Agreement

For better comparison, future work should use ratings to discriminate between groups of students similarly to the groups derived in related work, so that inter-rater reliability can be compared for similar rating tasks. Furthermore, it is plausible that

conducting the annotation process online brought limitations, especially compared to setups where the raters were instructed together and in person. Instructions being supplied differently and varying in their extent should be examined with respect to the effect on measured rater agreement, as well as the effect of the rating setup (in person, on paper, online). In addition, a more controlled setting for the annotation process, where a fixed group of raters all provide ratings for the same samples, could provide data for a more meaningful examination of the suspected relationship between the reliability of ratings and the specificity of a given legibility criterion.

8.2 Modeling a Renowned Handwriting Scale

In this work, legibility criteria were adopted from related work and instructions on how to rate them were created. Instead of creating a new handwriting scale for the annotation, future work could focus on ratings from an existing and proven handwriting scale with higher inter-rater reliability. This would probably increase the reliability of collected ratings and would allow to use measurements of agreement to test results reported for the given scale. Furthermore, modeling a handwriting scale that is already being used in practice increases the applicability of the trained models in case they reach sufficient performance.

8.3 Recording of Suitable Handwriting Samples

For this work, the supervisors instructed the students to write the sentences in one go and to repeat the recording when they needed to correct their writing or when they paused writing mid-sentence. The high frequency of corrections found in the manual validation of the dataset showed that this might not be sufficient. In a setting where each student is overlook by a supervisor, who interrupts when necessary, the recorded sensor data could be assured to capture only the movement related to writing the sentence. This in turn could benefit models trained on this data. However, training only on this kind of high-quality data might miss the purpose of models trained for legibility assessment. Another approach could investigate how sensor data can be preprocessed to reduce the variety of sensor signals originating from writing pauses or corrections. Furthermore, future data acquisition should attempt to collect more illegible handwriting samples to avoid problems that originated from strongly unbalanced data. For example, it could be examined whether instructing students to write quickly results in a greater share of hard to read samples.

8.4 Detecting and Controlling the Confounding Variables

As pointed out earlier, confounding variables, like the writing style, strongly affect model performance and decrease the expressiveness of evaluations. It is believed that characteristics like the writing style or the presence of corrections can be detected by machine learning models. This would allow to add this meta information to all

samples of a dataset, so that the variables can be accounted for in training and evaluating models.

8.5 Refining the Deep Learning Approach

All models trained in this work were simple CNNs of moderate size. Only one-dimensional kernels were used so that sensor channels were processed separately. It is suspected that the use of other deep learning architectures could drastically improve the performance of the model. Different approaches to modeling, such as allowing features to be learned for channels jointly, using RNNs or Transformers to capture the sequential nature of the sensor data, or pretraining on unlabeled handwriting sensor data, are believed to present opportunities for more accurate predictions of legibility labels. In case future models perform better, the rater-specific approach could be revisited to train several models that imitate different individual raters. These models could then form an ensemble and provide a distribution of ratings per sample instead of a single value prediction.

A Appendix

A.1 Descriptions of Reviewed Handwriting Scales

SEMS and (SOS-2) The German Systematische Erfassung motorischer Schreibstörungen (SEMS) is a measurement tool to identify children with handwriting difficulties adopted from the Dutch SOS-2. Waelvelde et al. (2012) reported high construct validity as well as high intra- and interrater reliability for the original Dutch SOS-2. The construct validity was evaluated using the scale to discriminate between children with and without motor difficulties. Suspects perform a near-copy writing task. The scales evaluate both the legibility of the produced writing as well as writing speed. Franken and Harris (2021) found that the SEMS score can be used to accurately identify children with handwriting problems in the second grade, but its sensitivity as a diagnostic test decreased when used in fourth grade. In the corresponding questionnaire, legibility is assessed through questions about seven criteria. For each legibility criterion, the rater states if it is fulfilled mostly (0), sometimes (1) or rarely (2).

1. Are the letters correctly formed?
2. Are letters written without (...) interruptions?
3. (...) are the joins fluid and correct?
4. Is the child's writing the correct size (...)?
5. Are all of the letters approximately the same size?
6. Does the child leave enough space between words?
7. Does the child write on the line (...)?

HLS Barnett et al. (2018) developed the Handwriting Legibility Scale (HLS) as a quick and easy-to-use tool to assess the legibility of handwriting. Its construct validity was evaluated using the scale to discriminate between children with Developmental Coordination Disorder (DCD) and normal developing children. A holistic legibility score is computed from five criteria that are rated on a five-point Lickert scale following the corresponding questionnaire. The suspect performs a free writing task. After instructions from the authors, the rater makes a judgement on each of the five criteria ranging from A to E. The questionnaire contains verbal explanations of scores between one and five for each criterion. The global legibility score is calculated as the sum of these five scores that capture the impression of the rater:

A An overall impression of global legibility based on your first reading of the text.

B An overall impression of the amount of effort required for you to read the script the first time.

C An overall impression of the layout of the writing on the page.

D An overall impression of letter formation.

E An overall impression of the attempts made to rectify letters within words.

ETCH The Evaluation Tool of Children's Handwriting (ETCH) tests for many aspects of handwriting performance. The suspects participate in several writing tasks. The examiner observes the writing process to assess aspects of the writing process, for example, how the pen is gripped. Afterwards, the writing product is examined. Duff and Goyen (2010) describe it as a criterion-referenced assessment that focuses on the readability of letters, words, and numbers at a glance and out of context. The corresponding examiner's manual by Amundson (2004) provides detailed instructions on the preparation, execution, and interpretation of the proposed assessment. This tool was designed for use by occupational therapists. Duff and Goyen (2010) displayed its construct validity by using the scores to discriminate between children with and without handwriting dysfunctions.

HPSQ Rosenblum (2008) proposed the Handwriting Proficiency Screening Questionnaire (HPSQ) as a standardized practical tool to identify handwriting difficulties among school-age children. To examine its construct validity, participants were divided into two groups based on their HPSQ score. The authors found significant differences between handwriting scores measured in the two groups using the Hebrew Handwriting Evaluation (HHE) which is taken as an indicator that the underlying latent construct of legibility is assessed. It is intended to be used by teachers to assess the handwriting of their students. The questionnaire contains ten questions that the examiner rates from one ("never") to five ("always"). A principal component factor analysis of these ten criteria revealed two main factors. The first comprises questions one, two, four, and ten. The authors summarized those as being related to legibility.

1 Is the handwriting difficult to read?

2 Do you have difficulty reading the handwriting?

4 Does the child often erase while writing?

10 Are you satisfied with the handwriting?

MHT Following the description by Rosenblum et al. (2003), the Minnesota Handwriting Test (MHT) was developed to assist occupational therapists in identifying school children with writing difficulties. Suspects copy a standardized set of words for a fixed period. Subsequently, the examiner checks which statements from a set list of fourteen observations apply and rates the produced writing on six criteria of handwriting quality. For each such criterion, the rater checks which of the three statements best describes the proficiency of the suspect in this area. The statements reflect how the suspect is performing in comparison to his or her peers. The examiner checks whether the suspect performs "Like Peers", "Somewhat Below Peers" or "Well Below Peers". Five of these criteria are related to legibility.

1. Legibility
2. Form
3. Alignment
4. Size
5. Spacing

A.2 List of Examined Labeling Tools

A list of all the labeling tools that were reviewed.

Table 30: Listing of all the reviewed labeling tools. ?? indicates that the documentation did not give sufficient information to judge whether the requirement is met.

Name	Supports classification	Centralized web-app	User management	Simple user-interface
labelImg	no	no	no	no
cvat	yes	yes	yes	no
labelme	yes	no	no	no
VoTT	yes	yes	no	no
imglab	no	no	no	no
YOLO_mark	no	no	no	no
PixelAnnotationTool	no	no	no	no
OpenLabeling	no	no	no	no
imagetagger	no	yes	??	no
ImageAnnotation	no	yes	??	no
deeplabel	no	no	no	no
MedTagger	no	yes	yes	no
LabelBox	??	??	??	??
turktool	no	yes	no	no
Pixie	no	no	no	no
OpenLabeler	??	no	no	no
Anno-Mage	no	no	no	no
CATMAID	??	yes	??	no
make-sense	yes	no	no	no
LOST	??	yes	yes	no
Annotorious	yes	yes	no	no
Sloth	??	no	no	no
Pixano	yes	yes	no	no
Alp's labeling too (ALT)	no	no	no	no
Classifai	no	no	no	no
COCO Annotator	no	yes	yes	no
commacoloring	no	no	no	no
DataGym.ai	yes	yes	??	??
diffgram	yes	yes	??	??
dsgou/annotator	no	no	no	no
Etiketai	no	no	no	no
(FIAT)	no	no	no	no
Grid-Annotation-Tool-2	no	no	no	no

Table 31: Continuation of the listing of all the reviewed labeling tools.

Name	Supports classification	Centralized web-app	User management	Simple user-interface
labelImg	no	no	no	no
ilastik	no	no	??	??
imannotate	no	yes	yes	??
labelml	??	??	??	??
labeld	??	yes	??	??
label-studio(community)	yes	yes	no	no
react-image-annotation	yes	no	no	no
scalabel	??	yes	??	??
tator	no	yes	yes	no
universal-data-tool	yes	yes	no	no
Slicer	??	no	no	??
anylabeling	no	no	no	no
autodistill	no	??	no	??
bbox-visualizer	no	no	no	no
BoundingBoxEditor	no	no	no	no
knossos	no	no	no	no
labelCloud	no	no	no	no
labelflow	??	yes	??	??
myvision	??	no	no	??
OHIF/Viewers	??	no	no	no

A.3 Instruction Texts for the Four Criteria

A.3.1 Q1 global-legibility

Bitte lies den unten angezeigten Satz. Bewerte anschließend, wie leserlich du die Schrift findest. Achte darauf, ob du den Satz mehrfach lesen musst, oder ob du ihn in einem Schwung lesen kannst. Musst du den Satz oder einzelne Wörter mehrfach lesen, dann ist der Satz weniger leserlich. Vergleiche den Satz mit den angezeigten Beispielen, um besser einschätzen zu können, welche Bewertung zutreffend ist.

1 - Die Bewertung 1 bedeutet, dass der Satz sehr leserlich ist. Du konntest ihn schnell und problemlos lesen. Du bist nicht in Stocken gekommen.

5 - Die Bewertung 5 bedeutet, dass der Satz sehr schwer leserlich ist. Selbst nach mehrfachem Lesen ist mindestens ein Wort nicht eindeutig zu entziffern.

A.3.2 Q2 consistent-slant

Bitte lies den unten angezeigten Satz. Nachdem du den Satz einmal gelesen hast, schaue dir einen Buchstaben nach dem anderen an. Achte auf die Neigung der vertikalen Striche. Sind alle Buchstaben in die gleiche Richtung geneigt (nach links

oder rechts)? Ist die Neigung bei allen Buchstaben gleich, ähnlich, oder unterschiedlich? Vergleiche den Satz mit den angezeigten Beispielen, um besser einschätzen zu können, welche Bewertung zutreffend ist. Die Bewertung richtet sich nach dem größten Neigungsunterschied zweier Buchstaben und nicht nach der durchschnittlichen Abweichung.

1- Die vertikalen Striche aller Buchstaben sind einheitlich geneigt. Mit dem blossen Auge lässt sich keine Abweichung im Winkel erkennen.

5- Verschiedene Buchstaben sind unterschiedlich ausgerichtet. Zwischen mindestens zwei Buchstaben mit vertikalem Strich ist eine starke Abweichung des Winkels zu erkennen.

A.3.3 Q3 letter-formation_rnh

Bitte lies den unten angezeigten Satz. Achte dabei auf die Buchstaben 'r', 'n' und 'h'. Diese drei Buchstaben entstehen durch einen ähnlichen Schwung des Stiftes. Sie unterscheiden sich nur darin, wie weit oben der vertikale Strich beginnt und wie weit unten der Bogen endet. Sind diese Buchstaben einzeln eindeutig zu erkennen? Oder sieht eines der 'n' eher aus wie ein 'r' oder 'h' bzw. andersherum.

1 - Alle Vorkommen der Buchstaben 'r', 'n' und 'h' sind wohlgeformt und eindeutig zu erkennen. Der vertikale Strich des 'n' ist deutlich kürzer als der eines 'h', deshalb sind die beiden Buchstaben leicht zu unterscheiden. Der Bogen des 'n' reicht weiter hinunter zur Grundlinie als der eines 'r', deshalb sind die beiden Buchstaben leicht zu unterscheiden. Es bedarf nicht den Kontext im Wort, um zu wissen, dass es sich um den jeweiligen Buchstaben handelt.

5 - Es gibt mindestens ein 'n', das eher wie ein 'r' oder 'h' aussieht, bzw. andersherum. Ohne den Kontext im Wort könnte man diesen Buchstaben auch für einen anderen halten.

A.3.4 Q4 letter-formation_ad

Bitte lies den unten angezeigten Satz. Achte dabei auf die Buchstaben 'a' und 'd'. Diese zwei Buchstaben entstehen durch einen ähnlichen Schwung des Stiftes. Sie unterscheiden sich nur darin, wie weit nach oben der vertikale Strich gezogen wurde. Sind diese Buchstaben einzeln eindeutig zu erkennen? Oder sieht eines der 'a' eher aus wie ein 'd', bzw. andersherum.

1 - Alle Vorkommen der Buchstaben 'a' und 'd' sind wohlgeformt und eindeutig zu erkennen. Der vertikale Strich des 'a' ist deutlich kürzer als der eines 'd', deshalb sind die beiden Buchstaben leicht zu unterscheiden. Es bedarf nicht den Kontext im Wort, um zu wissen, dass es sich um den jeweiligen Buchstaben handelt.

5 - Es gibt mindestens ein 'a', das eher wie ein 'd' aussieht, bzw. andersherum. Ohne den Kontext im Wort könnte man diesen Buchstaben auch für einen anderen halten.

A.4 User Interface of the Annotation App

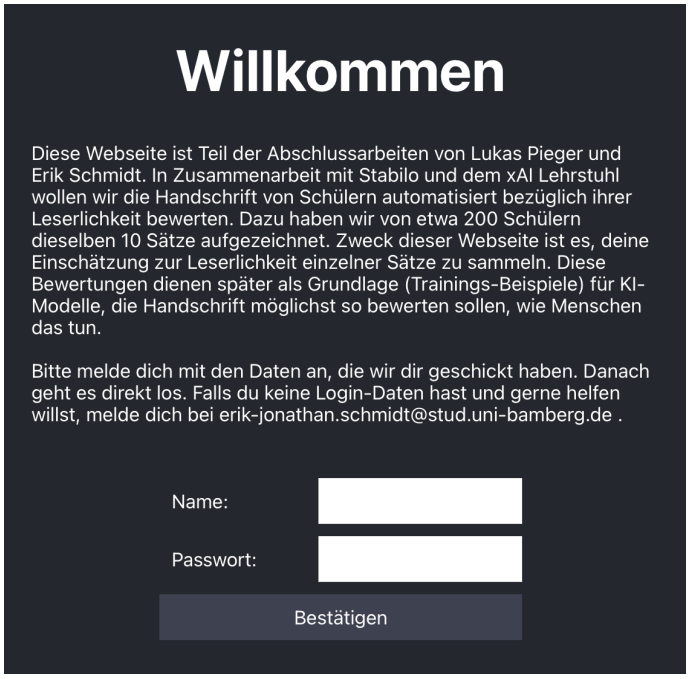


Figure 28: The login page of the Labeling App.

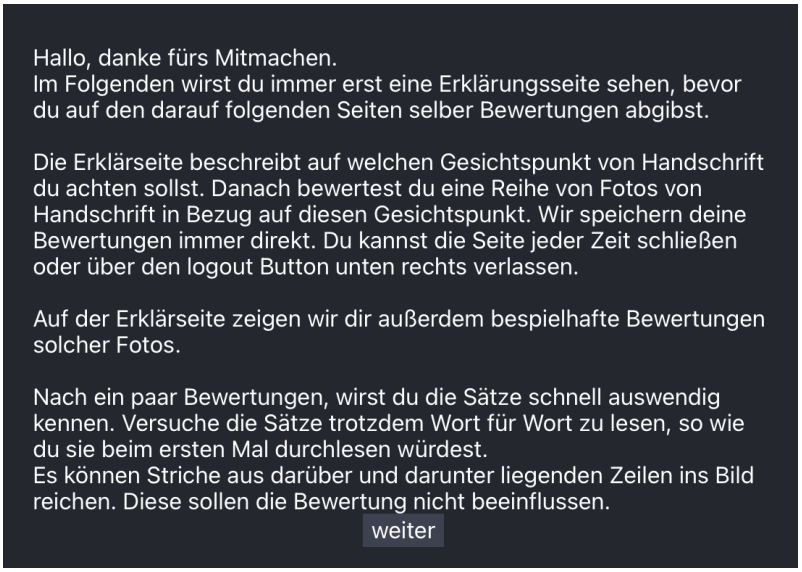


Figure 29: The welcome page of the Labeling App with brief introduction to the project.

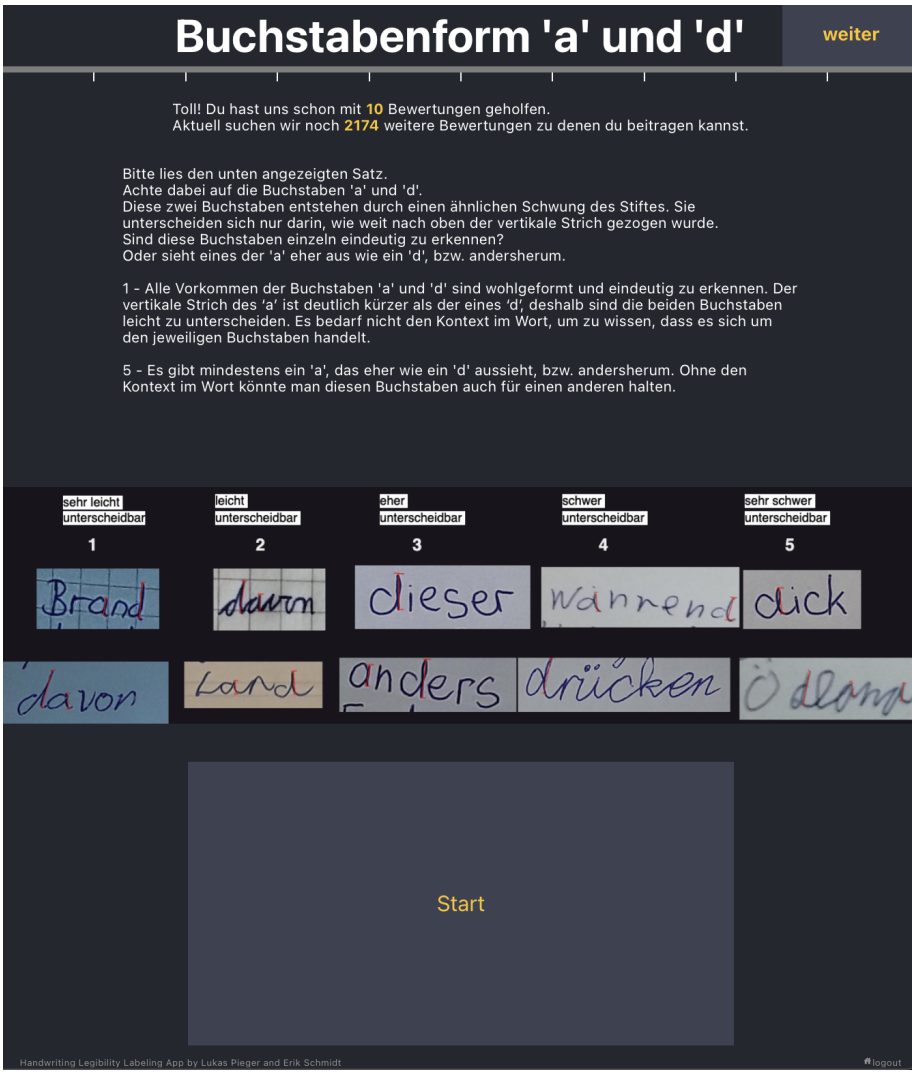


Figure 30: The batch introduction page of the Labeling App. The criterion which was to rate on the next sides was explained in a text and with an example image.

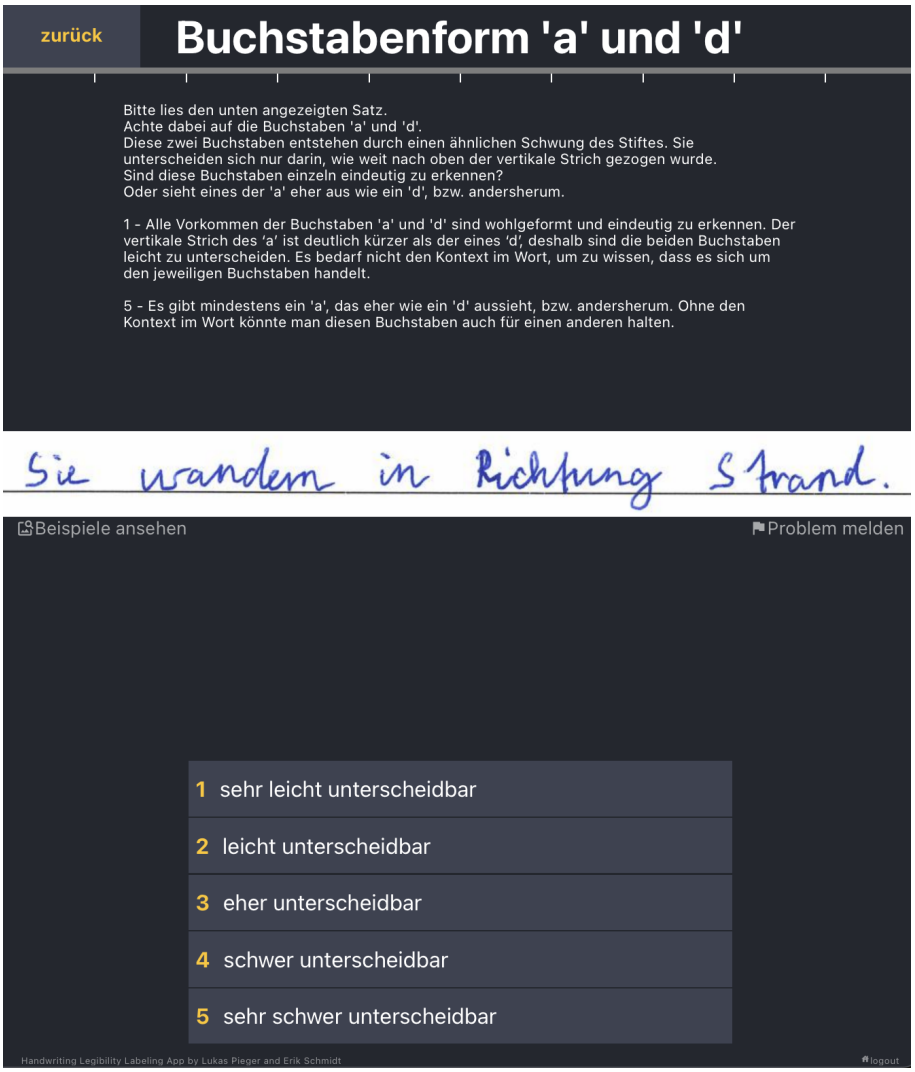


Figure 31: The scoring page of the Labeling App, where raters inspect the image of a sentence and select a suitable score. Raters can review the example image from the batch introduction page by clicking "view example" underneath the image. "report issue" did open a popup where raters could leave a message.

A.5 Data Splits

Table 32: Composition of training, validation and test splits. Distribution of the ratings given in response to questions Q1, Q2, Q3 and Q4.

Split and Question	# samples	# students	# rating 0	# rating 1	# rating 2	# rating 3	# rating 4
Q1 Train	1398	140	2758	2133	1535	719	191
Q1 Validation	410	41	836	672	417	174	31
Q1 Test	209	21	270	282	272	202	45
Q2 Train	1398	140	1598	1596	763	206	20
Q2 Validation	410	41	444	469	222	70	8
Q2 Test	209	21	177	225	157	52	0
Q3 Train	1398	140	1732	1211	619	313	134
Q3 Validation	410	41	523	352	168	99	24
Q3 Test	209	21	225	178	90	62	32
Q4 Train	1398	140	1085	462	203	79	24
Q4 Validation	410	41	317	136	49	19	10
Q4 Test	209	21	145	80	35	14	0

A.6 Outlier Samples

Figure 32 shows handwriting samples with best and worst scores in *StabLe*(Q1) (a) (b), *StabLe*(Q2) (c) (d), *StabLe*(Q3) (e) (f), and *StabLe*(Q4) (g) (h).

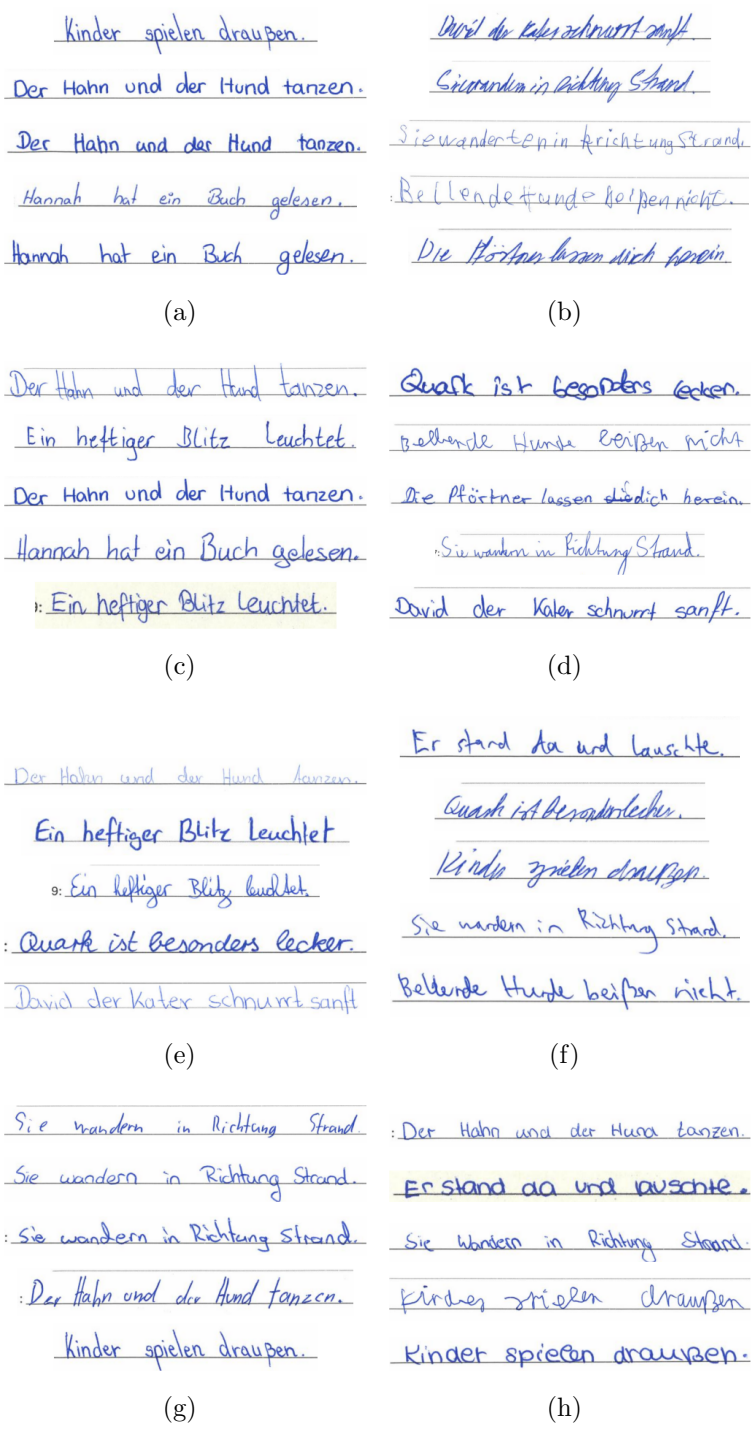


Figure 32: The five best and worst scores handwriting samples per question.

A.7 Models

Table 33: Listing of the models that were trained for the different experiments. Models differ in the architecture of the convolutional layers and the prediction head, in the data they were trained on, the label merging strategy, and the prediction task they were trained for.

Model	Experiment	CNN	Head	Dataset	Label	Task
A	Repr, CnnArc, CompDs	broad	1l	*Curation Beauty*	-	3 class
B	CnnArc, CompDs	cone	1l	*Curation Beauty*	-	3 class
C	CnnArc, CompDs	broad	1l	*StabLe*(Q1)	mean-rounded	3 class
D	CnnArc, CompDs	cone	1l	*StabLe*(Q1)	mean-rounded	3 class
E	CnnArc	broad	1l	*StabLe*(Q1)	mean-rounded	5 class
F	CnnArc, HeadArc	cone	1l	*StabLe*(Q1)	mean-rounded	5 class
G	HeadArc	cone	3l	*StabLe*(Q1)	mean-rounded	5 class
H	HeadArc, LabMerg	cone	1l	*StabLe*(Q1)	mean-rounded	reg
I	HeadArc	cone	3l	*StabLe*(Q1)	mean-rounded	reg
J	LabMerg	cone	1l	*StabLe*(Q1)	majority	reg
K	LabMerg, HypPar	cone	1l	*StabLe*(Q1)	random	reg
L	LabMerg, HypPar, Eval	cone	1l	*StabLe*(Q1)	mean	reg
L*	Eval	cone	1l	*StabLe*(Q1)*	mean	reg
$L^{\setminus cu}$	Eval	cone	1l	*StabLe*(Q1\cu)	mean	reg
$L^{\setminus ty}$	Eval	cone	1l	*StabLe*(Q1\ty)	mean	reg
$L^{\setminus co}$	Eval	cone	1l	*StabLe*(Q1\co)	mean	reg
M	LabMerg, Eval	cone	1l	*StabLe*(Q1)	rater-specific	reg
N	Eval	cone	1l	*StabLe*(Q2)	mean	reg
O	Eval	cone	1l	*StabLe*(Q3)	mean	reg
P	Eval	cone	1l	*StabLe*(Q4)	mean	reg

B Collaboration in this Work

For this work I collaborated with colleagues at STABILO, members of the Schreibmotorik Institut (SMI) and Lukas Pieger to realize different parts of the described project. Employees of STABILO helped with the data recordings and supported the creation of the *StabLe* dataset by giving valuable feedback. Furthermore, they supported the training of the models by setting up the GPU server and giving advice when needed. Members of the SMI supported by providing relevant literature on the assessment of legibility, in choosing the legibility criteria, in designing the reference sentences, and by labeling the handwriting samples. Lukas Pieger conducted a master thesis project that was carried out in parallel to this work. Where this work focused on sensor data, he was concerned with image data. We collaborated to create the reference sentences and to create the sheet for the recordings. We both supervised the two recording sessions and developed the annotation app together, where he fully implemented the sentence extraction. The descriptive analysis of the collected ratings was also done collaboratively.

C Use of Generative AI

Generative AI, such as ChatGPT has been used as search engine and to formulate passages based on bulletpoints by the author, to paraphrase text written by the author and to help summarize text by the author for reference in other sections. All generated text was carefully proofread, manually adjusted, and paraphrased.

Bibliography

S.J. Amundson. *Evaluation Tool of Children's Handwriting: ETCH Examiner's Manual*. O.T. KIDS, 2004. URL `https://books.google.de/books?id=y4x0oAEACAAJ`.

Eunice H. Au, Annie McCluskey, and Natasha A. Lannin. Inter-rater reliability of three adult handwriting legibility instruments. *Australian Occupational Therapy Journal*, 59(5):347–354, 2012. doi: https://doi.org/10.1111/j.1440-1630.2012.01035.x. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1440-1630.2012.01035.x`.

Anna L. Barnett, Mellissa Prunty, and Sara Rosenblum. Development of the handwriting legibility scale (hls): A preliminary examination of reliability and validity. *Research in Developmental Disabilities*, 72:240–247, 2018. ISSN 0891-4222. doi: https://doi.org/10.1016/j.ridd.2017.11.013. URL `https://www.sciencedirect.com/science/article/pii/S0891422217303074`.

Y. Bengio and X. Glorot. Understanding the difficulty of training deep feed forward neural networks. *International Conference on Artificial Intelligence and Statistics*, pages 249–256, 01 2010.

Virginia W. Berninger and Maggie O'Malley May. Evidence-based diagnosis and treatment for specific learning disabilities involving impairments in written and/or oral language. *Journal of Learning Disabilities*, 44(2):167–183, 2011. doi: 10.1177/0022219410391189. URL `https://doi.org/10.1177/0022219410391189`. PMID: 21383108.

Virginia W. Berninger and Todd L. Richards. Building a reading brain neurologically. In *Brain literacy for educators and psychologists*, 2002. URL `https://api.semanticscholar.org/CorpusID:141848561`.

Mugdim Bublin, Franz Werner, Andrea Kerschbaumer, Gernot Korak, Sebastian Geyer, Lena Rettinger, Erna Schönthaler, and Matthias Schmid-Kietreiber. Handwriting evaluation using deep learning with sensogrip. *Sensors*, 23(11), 2023. ISSN 1424-8220. doi: 10.3390/s23115215. URL `https://www.mdpi.com/1424-8220/23/11/5215`.

Laura Dinehart. Handwriting in early childhood education: Current research and future implications. *Journal of Early Childhood Literacy*, 15, 03 2014. doi: 10.1177/1468798414522825.

Laura Dinehart and Louis Manfra. Associations between low-income children's fine motor skills in preschool and academic performance in second grade. *Early Education and Development*, 24:138–161, 02 2013. doi: 10.1080/10409289.2011.636729.

Sharon Duff and Traci-Anne Goyen. Reliability and validity of the evaluation tool of children's handwriting-cursive (etch-c) using the general scoring criteria. *The*

American journal of occupational therapy : official publication of the American Occupational Therapy Association, 64:37–46, 01 2010. doi: 10.5014/ajot.64.1.37.

Katya P Feder and Annette Majnemer. Handwriting development, competency, and intervention. *Developmental Medicine & Child Neurology*, 49(4):312–317, 2007. doi: https://doi.org/10.1111/j.1469-8749.2007.00312.x. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-8749.2007.00312.x`.

Anita Franken and Susan Harris. Teachers' perceptions of handwriting legibility versus the german systematic screening for motoric-handwriting difficulties (sems). *OTJR: Occupation, Participation and Health*, 41:153944922110338, 07 2021. doi: 10.1177/15394492211033828.

Lea Grabmann. Automatic evaluation of the readability of handwriting using a sensor enhanced digital pen. Master's thesis, Friedrich-Alexander-Universität, 2023.

Yahia Hamdi, Hanen Akouaydi, Houcine Boubaker, and Adel M. Alimi. Handwriting quality analysis using online-offline models, 2020. URL `https://arxiv.org/abs/2010.06693`.

C.W. Harris and American Educational Research Association. *Encyclopedia of Educational Research: A Project of the American Educational Research Association*. Macmillan, 1960. URL `https://books.google.de/books?id=D_RXAAAAMAAJ`.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

Karin H. James and Laura Engelhardt. The effects of handwriting experience on functional brain development in pre-literate children. *Trends in Neuroscience and Education*, 1(1):32–42, 2012. ISSN 2211-9493. doi: https://doi.org/10.1016/j.tine.2012.08.001. URL `https://www.sciencedirect.com/science/article/pii/S2211949312000038`.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Li Mae Y Koo T. K. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *PMC*, 2016.

Edward R. Lewis and Hilda P. Lewis. An analysis of errors in the formation of manuscript letters by first-grade children. *American Educational Research Journal*, 2(1):25–35, 1965. doi: 10.3102/00028312002001025. URL `https://doi.org/10.3102/00028312002001025`.

Jinyuan Liu, Wan Tang, Guanqin Chen, Yin Lu, Changyong Feng, and Xin Tu. Correlation and agreement: overview and clarification of competing concepts and measures. *Shanghai archives of psychiatry*, 28:115–120, 04 2016. doi: 10.11919/j.issn.1002-0829.216045.

Christian Marquardt, Marianela Diaz Meyer, Manuela Schneider, and René Hilgemann. Learning handwriting at school – a teachers' survey on actual problems and future options. *Trends in Neuroscience and Education*, 5(3):82–89, 2016. ISSN 2211-9493. doi: https://doi.org/10.1016/j.tine.2016.07.001. URL `https://www.sciencedirect.com/science/article/pii/S2211949316300126`. Writing in the digital age.

Margaret Martlewm. Handwriting and spelling: Dyslexic children's abilities compared with children of the same chronological age and younger children of the same spelling level. *British Journal of Educational Psychology*, 62(3):375–390, 1992. doi: https://doi.org/10.1111/j.2044-8279.1992.tb01030.x. URL `https://bpspsychub.onlinelibrary.wiley.com/doi/abs/10.1111/j.2044-8279.1992.tb01030.x`.

D McCarney, L Peters, S Jackson, M Thomas, and J Kirby. Does poor handwriting conceal literacy potential in primary school children? *International Journal of Disability, Development and Education*, 60(2):105–118, May 2013. doi: 10.1080/1034912X.2013.786561. URL `http://researchspace.bathspa.ac.uk/8422/`.

Kathleen McHale and Sharon A. Cermak. Fine motor activities in elementary school: preliminary findings and provisional implications for children with fine motor problems. *The American journal of occupational therapy : official publication of the American Occupational Therapy Association*, 46 10:898–903, 1992. URL `https://api.semanticscholar.org/CorpusID:33368309`.

Jiri Mekyska, Zoltan Galaz, Katarina Safarova, Vojtech Zvoncak, Lukas Cunek, Tomas Urbanek, Jana Marie Havigerova, Jirina Bednarova, Ján Mucha, Michal Gavenciak, Zdenek Smekal, and Marcos Faundez-Zanuy. *Assessment of Developmental Dysgraphia Utilising a Display Tablet*, page 21–35. Springer Nature Switzerland, 2023. ISBN 9783031454615. doi: 10.1007/978-3-031-45461-5_2. URL `http://dx.doi.org/10.1007/978-3-031-45461-5_2`.

Taivo Pungas. awesome-data-annotation. `https://github.com/taivop/awesome-data-annotation`, 2022.

Oreste Renato Rondinella. *An Evaluation of Subjectivity of Elementary-School Teachers in Gradig Handwriting*. PhD thesis, Fordham University, 1962.

Sara Rosenblum. Development, reliability, and validity of the handwriting proficiency screening questionnaire (hpsq). *The American journal of occupational therapy : official publication of the American Occupational Therapy Association*, 62:298–307, 05 2008. doi: 10.5014/ajot.62.3.298.

Sara Rosenblum, Patrice Weiss, and Shula Parush. Product and process evaluation of handwriting difficulties. *Educational Psychology Review*, 15:41–81, 01 2003. doi: 10.1023/A:1021371425220.

Angelika Rüb. *Leserlichkeit der Handschrift von Schreibanfängern : eine empirische Studie zur Erfassung und Bedeutung der Leserlichkeit*. PhD thesis, Universität Bamberg, Bamberg, 2018. Jahr der Erstpublikation: 2017.

Liora Epsztein Sara Rosenblum and Naomi Josman. Handwriting performance of children with attention deficit hyperactive disorders: A pilot study. *Physical & Occupational Therapy In Pediatrics*, 28(3):219–234, 2008. doi: 10.1080/01942630802224934. URL https://doi.org/10.1080/01942630802224934. PMID: 19064457.

Patrick Shrout and Joseph Fleiss. Intraclass correlations: Uses in assessing rater reliability. *Psychological bulletin*, 86:420–8, 03 1979. doi: 10.1037/0033-2909.86.2.420.

Julius Sim and Chris C Wright. The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy*, 85(3):257–268, 03 2005. ISSN 0031-9023. doi: 10.1093/ptj/85.3.257. URL https://doi.org/10.1093/ptj/85.3.257.

Thorarinn Stefansson and Ragnheidur Karlsdottir. Formative evaluation of handwriting quality. *Perceptual and motor skills*, 97:1231–64, 01 2004. doi: 10.2466/PMS.97.8.1231-1264.

Karen R. Harris Steve Graham. *Handbook of Educational Psychology and Students with Special Needs*, chapter Writing and Students with Learning Disabilities. Routledge, Taylor & Francis Group, 2020.

Piotr Szymczak. Grading for translation quality or legibility? a challenge to objective assessment of translation quality in handwritten samples. *Journal of Translator Education and Translation Studies*, 2016. URL https://api.semanticscholar.org/CorpusID:54976277.

Max Tkachenko. awesome-data-labeling. https://github.com/HumanSignal/awesome-data-labeling?tab=readme-ov-file, 2022.

Alex Strick van Linschoten. awesome-open-data-annotation. https://github.com/zenml-io/awesome-open-data-annotation, 2024.

Hilde Van Waelvelde, Tinneke Hellinckx, Wim Peersman, and Bouwien C. M. Smits-Engelsman. Sos: A screening instrument to identify children with handwriting impairments. *Physical & Occupational Therapy In Pediatrics*, 32:306 – 319, 2012. URL https://api.semanticscholar.org/CorpusID:7120214.

Declaration of Authorship

Ich erkläre hiermit gemäß §9 Abs. 12 APO, dass ich die vorstehende Abschlussarbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Des Weiteren erkläre ich, dass die digitale Fassung der gedruckten Ausfertigung der Abschlussarbeit ausnahmslos in Inhalt und Wortlaut entspricht und zur Kenntnis genommen wurde, dass diese digitale Fassung einer durch Software unterstützten, anonymisierten Prüfung auf Plagiate unterzogen werden kann.

Place, Date

Signature