Materials & Design 257 (2025) 114424

Contents lists available at ScienceDirect

Materials & Design

journal homepage: www.elsevier.com/locate/matdes

An automated Machine Learning based approach for a reproducible and efficient evaluation of industrial Charpy V-notch specimens

Adrian Herges [a,*], Björn-Ivo Bachmann [a,b], Sebastian Scholl [c], Frank Mücklich [a,b]

[a] Department of Materials Science, Saarland University, 66123 Saarbrücken, Germany
[b] Material Engineering Center Saarland (MECS), 66123 Saarbrücken, Germany
[c] AG der Dillinger Hüttenwerke, 66763 Dillingen/Saar, Germany

ARTICLE INFO

Keywords:
Charpy V-notch
Fracture surfaces
Machine learning
Fracture surface classification
Ductile-Brittle-Transition-Temperature

ABSTRACT

An objective and automated method for the quantification of macroscopic images of tested Charpy V-notch specimens, specifically focusing on their ductility/brittleness characteristics based on a realistic, homogeneous and industrial environment is proposed. Our approach involves a multi-step preprocessing routine that incorporates color thresholding and connected component analysis to first detect the various Charpy V-notch specimen bundles according to their sample material affiliation and testing temperature. Subsequently, a U-Net was trained to further partition the preprocessed images into background, notch, and regions of ductile or brittle fracture, respectively through semantic segmentation. Thereby, a quantification of brittle and ductile fractions of each individual sample focusing on only the fracture surfaces can be conducted. The results obtained are then evaluated using Intersection over Union (IoU) metrics, a tailored domain-specific matrix and SEM images incorporating more objective annotations based on the high resolution and the higher depth of focus to assess the model performance. The findings presented in this study highlight the significant potential of machine learning and computer vision in the realm of a reproducible and objective, automated macroscopic fracture analysis on an industrial scale, providing valuable benefits for materials engineering and quality control in manufacturing processes.

1. Introduction

Fractography is a well-established approach for assessing the mechanical characteristics of materials in order to determine microstructure-property relations and for quality control. Apart from determining the brittle-ductile transition temperature (DBTT) based on measured impact energies [1,2], experts examine the macroscopic images of fracture surfaces from Charpy V-notch specimens to quantify the proportions of ductile and brittle fracture areas. This quantification is not merely descriptive: higher proportions of brittle fracture are typically associated with the lower shelf of the impact energy curve, where materials exhibit low energy absorption and poor toughness. In contrast, ductile fracture dominates in the upper shelf region, where specimens absorb significantly more energy during testing. Evaluating the temperature range in which this transition between fracture modes and properties occurs, the so-called DBTT is essential for evaluating the suitability of steels. However, the visual assessment of fracture surfaces is typically carried out manually and remains prone to subjectivity and limited reproducibility. The conventional practice involves macroscopic inspection of fracture samples, where experts assess the surface features and assign values for shear area (ductility) or crystallinity (brittleness). This approach, while widely accepted and applied, is heavily reliant on human evaluation. In this study, we replicated the traditional visual evaluation method using digital camera images to reflect standard industrial workflows. Previous efforts have attempted to automate this assessment through image processing techniques. For instance, a study by Park et al. proposed a method for quantifying ductile and brittle fracture regions by converting RGB images into binary masks using predefined thresholds, followed by pixel-wise area analysis [3]. While this approach improves reproducibility compared to manual evaluation, it remains limited in adaptability and generalization. In contrast, our method employs a machine learning-based semantic segmentation to enable flexible, data-driven classification after training a suitable model using expert-annotated data that can adapt to variations in imaging and

* Corresponding author.
E-mail addresses: adrian.herges@uni-saarland.de (A. Herges), bjoernivo.bachmann@uni-saarland.de (B.-I. Bachmann), sebastian.scholl@dillinger.biz (S. Scholl), muecke@matsci.uni-sb.de (F. Mücklich).

https://doi.org/10.1016/j.matdes.2025.114424
Received 25 April 2025; Received in revised form 7 July 2025; Accepted 17 July 2025
Available online 18 July 2025

specimen conditions. The objective is to reduce subjectivity and improve reproducibility in the quantification of fracture surface features, thereby establishing process-microstructure-property relations more reliably. In our study, specimens are grouped in sample boxes to guarantee uniform experimental conditions, with only specific cases subjected to investigation using scanning electron microscopy (SEM) due to cost and time efficiency. Consequently, ensuring precise and reproducible evaluations of these macroscopic fracture samples is of utmost importance, while more detailed investigation is usually not feasible. The visual application of machine learning and deep learning is now well established in everyday fields from autonomous driving [4] or to cancer cell detection in biomedicine [5]. Now, these approaches are increasingly being applied to materials science since artificial intelligence offers significant advantages in this domain. Notable examples include failure detection in materials [6], the segmentation and thereby quantification of grain boundaries [7] or complex phases [8] and other microstructural constituents and features. Furthermore, machine learning has also been applied in the field of fractography: Schmies et al. [9] utilized deep learning and topography data together with optical and electron-microscopy images. Their study showed the possibility to automatically segment microscale images of fractures with up to 18 different fracture classes utilizing expert annotations as a ground truth. In this context, a labeling scheme for electron microscope images was established, called "Fractographics v0.60" [10,11]. In a previous study, Rosenberger et al. used deep learning approaches for semantic segmentation of macroscopic fracture surfaces to assess the crack size [12]. Their work used images from various laboratories, acquired under diverse conditions, pursuing a robust and well generalizing solution. Afterwards the data was manually annotated and used for the training of a semantic segmentation approach. While that work primarily addressed various types of sample specimens and also determined ductility and brittleness fractions, our methodology aims to advance this concept to an industrial scale. An essential difference to this work is that in our approach several samples are extracted autonomously and analyzed within one image, as it is common practice in industry not to take individual images of each sample due to time scarcity. Building on this topic, Rosenberger et al. more recently investigated Charpy V-notch specimens using unsupervised machine learning to project static force–displacement curves onto fracture surfaces [13]. While this approach provides a novel perspective by linking mechanical tests with fracture morphology and represents another example of the potential of machine learning in fractography, it was conducted within a highly controlled experimental framework. In contrast, the method presented in this work is tailored to industrial conditions, where multiple specimens are imaged within sample boxes under limited resolution and different influences, such as lightning and sample arrangement, among others. Our focus lies in achieving robust and scalable fracture surface evaluations using simple imaging adapted to practical, high-throughput use cases. In

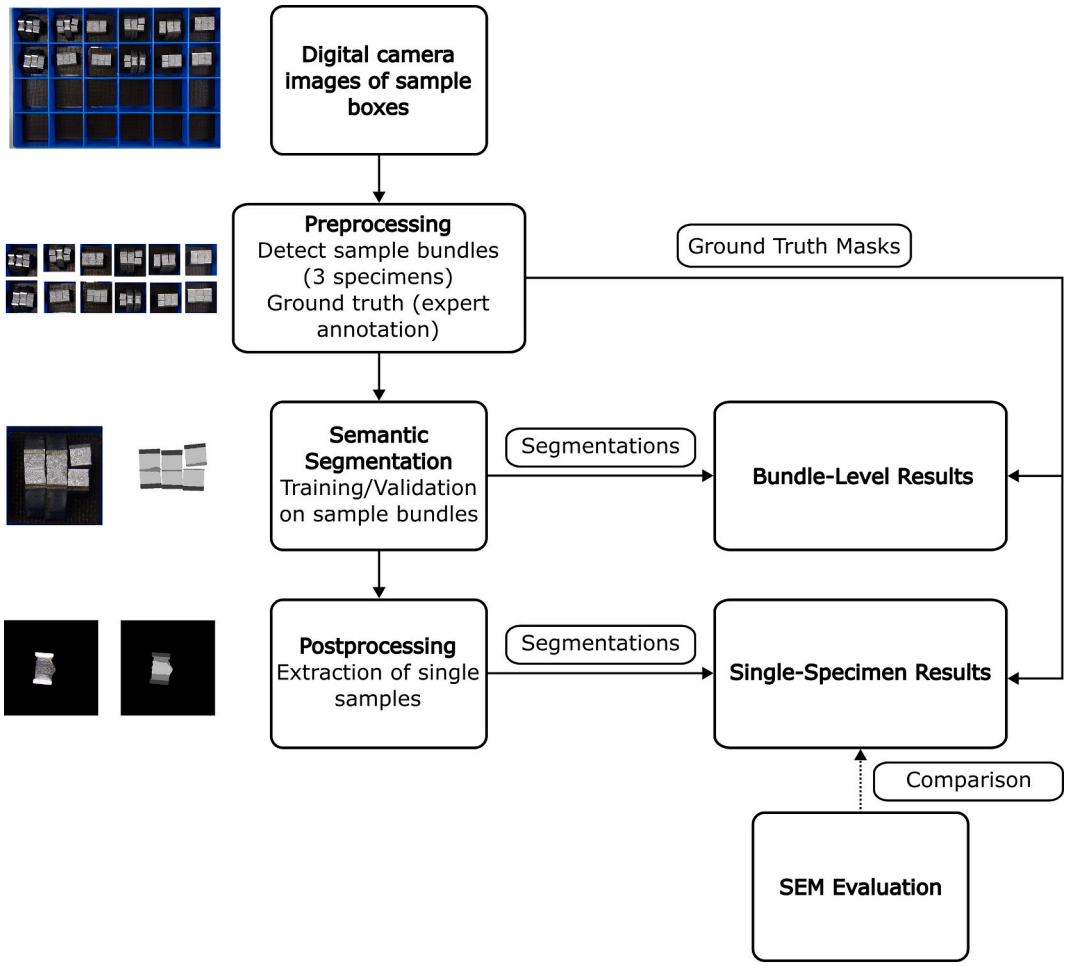


Fig. 1. Flow chart of the segmentation process of Charpy V-notch samples using the developed approach. The routine can be subdivided into three main stages: Preprocessing, Semantic Segmentation and Postprocessing. Arrows indicate the flow of data and the contributions of each step to the results, which can be subdivided into the Bundle-Level Results and the Single-Specimen Results. Bundle-Level results are derived through a pixel-wise comparison between the ground truth masks and the model segmentation. During Postprocessing, individual samples were extracted, enabling a direct comparison of each individual sample segmentation to the corresponding ground truth masks, referred to as Single-Specimen Results. Optional Scanning Electron Microscopy (SEM) images may further enhance evaluations at the specimen level through direct comparison.

previous work [14], groundwork for the automated assessment of fracture sample surface classification has been laid. While these initial steps demonstrated the feasibility of the approach, the present study builds upon and significantly enhances these results. Specifically, in this study, semantic segmentation was employed to refine the classification process and introduce a novel routine for the automated extraction and pixel-wise evaluation of samples, thereby improving both the precision and the scope of the obtained information.

2. Material & methods

The following section outlines the complete methodological pipeline used in this study, which combines standardized image acquisition with a multi-stage image processing and analysis workflow. The aim of this methodology is to enable the automated and reliable characterization of fracture surface features in Charpy V-notch specimens under various testing conditions. The entire processing pipeline was developed in Python and is illustrated schematically in Fig. 1.

The workflow consists of three main stages:

(1) **Preprocessing:** In this step, digital camera images of sample boxes containing specimens were separated into sample bundles, which are defined as a grouped set of three specimens of one testing condition in the sample box (e.g., one testing temperature and position in the weld). This was facilitated by using color thresholding and connected component analysis (CCA), which was implemented using OpenCV [15]. The resulting pictures were manually annotated to generate expert-labeled ground truth data. The preprocessed data then served as the foundation for the subsequent segmentation step.
(2) **Semantic Segmentation:** Using the annotated data, a U-Net-based deep learning approach was trained to classify ductile fracture, brittle fracture, specimen notch and the background. The model was implemented using the segmentation models library [16], using the Keras environment in accordance with the work of Bachmann et al. [7]. The dataset was split into training and validation sets. The output segmentation masks were then compared to the ground truth, enabling an initial evaluation of model accuracy.
(3) **Postprocessing:** To enable accurate evaluation at the single-specimen level, the generated segmentation masks of (2) were used to cut out the fracture surfaces in the sample bundle images. Then postprocessing techniques from the scikit-image library [17] were applied, including watershed and morphological operations. In this way individual samples together with their segmentation masks were attained.

The processing pipeline yields two types of results:

(1) **Bundle-Level Results:** Segmentation outputs for the grouped samples from the **Semantic Segmentation** were compared pixel-wise to the corresponding ground truth masks from **Preprocessing**. This enabled the creation of a pixel-wise evaluation matrix, quantifying model performance and identifying potential misclassifications across different testing conditions.
(2) **Single-Specimen Results:** Following **Postprocessing**, each specimen could be evaluated individually. The extracted fracture surfaces were again compared to the ground truth from the **Preprocessing** on a pixel level, providing error metrics for each single sample and enabling a more detailed interpretation of the fracture behavior. Optionally, Scanning Electron Microscopy (SEM) images could be used to compare the model's evaluation with high-resolution evaluations.

2.1. Material

The investigated specimens were manufactured from a low-alloy, thermomechanically rolled and electron-beam welded S355ML steel. The Charpy V-notch samples were subjected to testing temperatures ranging from −80 °C to 0 °C, with intervals of 20 °C between successive values, although not all temperature settings were tested for each sample configuration. Samples were furthermore taken from different positions, specifically, at the top, quarter, middle, and lower thickness locations, to assess the material's properties across the whole plate thickness. Consequently, the samples exhibited fracture behaviors across a spectrum, from the lower-shelf region with low absorbed impact energies (brittle), through the ductile–brittle transition (mixed), to the upper-shelf region with high absorbed impact energies (ductile). Moreover, neither the samples with ductile nor those with brittle fractures were broken apart after the test to ensure that the fracture surface remained pristine. For each individual testing condition three Charpy V-notch specimens were used to get a statistically more meaningful assessment.

2.2. General image acquisition

Three samples for each testing condition were placed together in one compartment of a box in an upright position. Several samples of different positions and different test temperatures were combined in one sample box according to the underlying industrial procedure. Fig. 2 illustrates an image of a collection of Charpy V-notch samples used as an input in scope of the present approach.

Due to the image acquisition process the quality of the images for each Charpy V-notch sample is lowered compared to acquiring images of every single specimen using a higher resolving image acquisition technique [14]. Due to the lower overall image quality, the standardization of boundary conditions and the recording conditions as far as possible, while paying attention to realistic circumstances of a potential serial application, plays an important role. A conventional digital camera Canon EOS 500D was used to acquire the images, with a resolution of 15.1 megapixels and a focal length of 60 mm. As part of image acquisition, external lighting was used to guarantee the most reproducible and homogeneous illumination possible. The nature of the illumination, and above all the angle of incidence and the light's intensity have a major influence on the appearance and thus for the potential characterization of the fracture surfaces. Given the complexity and variability of the optical features used to distinguish fracture types, lighting conditions, especially regarding time of day and season, should be kept as consistent as possible.

SEM images were also captured to validate the results on selected sample bundles, utilizing a Helios G4 PFIB CXe scanning electron microscope. The images were taken at an acceleration voltage of 20 kV using a secondary electron detector with a horizontal field width of 2590 μm and 4096x3536 pixels per image. To automate the complete visual observation of fracture surface details, a script based on the microscopes Python API was developed to automatically acquire SEM images in a grid-like pattern. These images were subsequently stitched together using Image Composite Editor 2.0.3.0 (Microsoft) to obtain a full image of the surface features. Furthermore, annotations were done as part of the ground truth assignment using GIMP 2.10.38 [18] and adapted symbols from the FractoGraphics database.

2.3. Image preprocessing and sample-wise patch extraction

To address the challenges presented by multiple samples in one picture first, a color threshold was utilized (Fig. 1 pre-processing). After normalization, the blue color channel enabled the separation between the different sample box compartments and thus distinguishing between

Fig. 2. Industrial recording to document the fracture behavior of the different test conditions. One compartment within the box represents a specific test configuration – each tested using 3 samples.

the different specimen states. Beyond this a "connected component analysis" (CCA) from the Opencv library was employed to differentiate the various sample states and extract images equal to the number of states examined for this material. This allowed us to subdivide the different states, meaning a full box of 24 squares would be divided into 24 patches. Empty compartments can be identified by the model of the semantic segmentation approach accordingly, as no sample-related classes such as ductile, brittle or notch are recognized. The resulting

Fig. 3. Subdivision of boxed samples using a color threshold and edge detection algorithm (CCA).

separation of specimen conditions after preprocessing is shown in Fig. 3.

Comparing Figs. 2 and 3, all relevant sample states were successfully extracted. The extracted and preprocessed samples were then used as input for training and evaluating the semantic segmentation model, as outlined in the next section.

3. Semantic segmentation model

3.1. Ground truth annotation and dataset creation

Based on the automatically extracted sample bundles within the cropped compartments, a training data set for the subsequent machine learning approach could be created. For this purpose, the images were evaluated using GIMP 2.10.38 and manually annotated to obtain masks with the different fracture surface areas of interest. One example is shown in Fig. 4.

3.2. Data augmentation and model training

The used data augmentation, a common practice to increase the data set and model performance [19], consisted of discrete 90° rotations, mirroring, distorting, focus alterations as well as color and contrast changes. The overall goal was to increase the size of the data set and the robustness of the model for variations in image acquisition and sample appearance [7]. The final training dataset consisted of 700 augmented training images, based on 100 annotated raw images as shown in Fig. 4 with dimensions of 512×512px, respectively. The remaining 25 annotated images were withheld for validation. Although this number of training images may appear low compared to segmentation datasets in other domains, recent publications have emphasized that microstructure images in materials science are typically considered data-rich. Consequently, deep learning studies in this field often operate under data-scarce conditions without a loss in effectiveness or acceptance [7,8,20], in particular also in regard to fracture surfaces of Charpy V-notch specimens [13]. The segmentation approach was defined to be a multiclass segmentation problem considering the four classes: ductile, brittle, notch and background. As model architecture, a U-Net [21], based on an ImageNet [22] pre-trained DenseNet-201 [23] backbone was used, using non-frozen encoder weights. For training, image data was normalized by dividing each color channel by 255, and the masks were one-hot encoded. As loss function a linear combination of a categorical crossentropy loss, a Jaccard loss as well as a class specific focal loss for the two classes ductile and brittle was used. Training took place for 10 epochs at a learning rate of 0.0001 using ADAM optimizer, followed by a fine-tuning for 5 epochs at a learning rate of 0.00001. As the final prediction result for each individual pixel, the class with the highest softmax probability output was chosen.

Model evaluation

One possible approach to interpret the accuracy of a semantic segmentation model is the so-called Intersection over Union (IoU). The IoU correlates the correctly classified pixels over the whole area of pixels. It is a commonly used and significant metric for segmentation tasks as shown in Eq. (1):

$$\mathrm{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}} \tag{1}$$

This standard metric, while useful for overall performance evaluation, proved insufficient for our specific purpose, as it does not provide detailed insight into the fracture-based misclassifications. Specifically, IoU alone makes it challenging to distinguish whether discrepancies are consistent across all samples or caused by individual outliers, particularly when considering sample bundles containing three distinct Charpy V-notch specimens. Therefore, class-wise IoU metrics were calculated for all samples to evaluate the overall robustness of the model, and a pixel-wise domain specific matrix approach was developed additionally, optimized for this specific use case focusing on the correctness of the identification of the fracture surfaces – not considering the background and sample notch. This approach was aimed at assessing any model biases toward particular classes.

Evaluation matrix

A domain-specific matrix, called the evaluation matrix, was generated by comparing the manually created ground truth mask with the model's prediction on a pixel-wise basis. The corresponding pixels of the ground truth state (y-axis) were compared with the prediction state (x-axis) across all four aforementioned classes and normalized by the total number of pixels in the image. Thus, a matrix with only diagonal entries would indicate perfect agreement between the segmentation results and the ground truth. Additionally, the sum of all entries in the matrix is 1.

3.3. Postprocessing and quantification

The arrangement of bundles of samples in close proximity, see Fig. 4, poses challenges since the specimens are arranged randomly with irregular empty spaces in between each other, therefore in some cases leading to segmentations where adjacent samples have connected pixels to the surrounding samples and therefore falsely being counted as one. There, the post-processing allows separating individual samples by altering the segmentation result. For this purpose, postprocessing masks were evaluated by their pixel values, to discern samples from background and afterwards binary operations like watershed, which is based on topographic distance in the scikit-Image library [17], and eroding were employed to separate samples as shown later in Fig. 9. Only by separating the individual samples can a corresponding specimen specific statistic be collected, based on which the pixelwise quantification of the

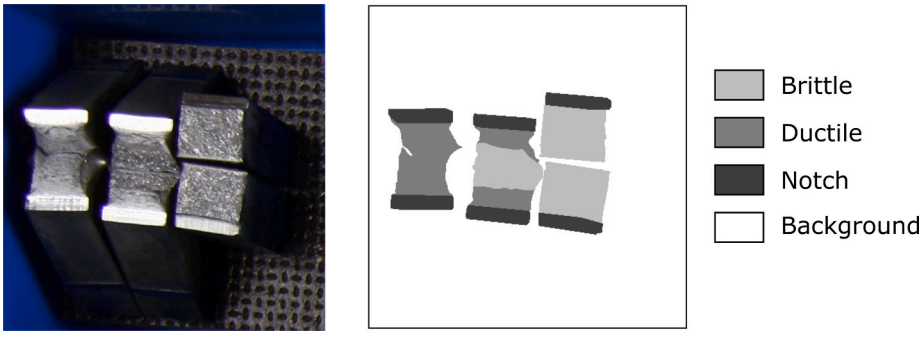


Fig. 4. Collection of fracture samples of one specific testing condition accompanied by the expert-annotated mask. Light grey denotes brittle regions, while dark grey indicates ductile fracture surfaces. The specimen notch is delineated in a black tone, against a white background. In this case a fully ductile, brittle and a mixed evaluated fracture sample are shown. Sample examinations were performed analogously to Fig. 4 on 125 images, each showing 3 tested Charpy V-notch samples, leading to a total amount of 375 unique examples.

different sample conditions can be carried out. The postprocessed segmentation outputs provided the foundation for the evaluation and analysis presented in Section 4, enabling a detailed assessment of model performance on both bundle and single-specimen levels.

4. Results and discussion

In the following, the results and discussion focus on the validation images. These were excluded from the training set and therefore offer an unbiased assessment of the model's performance. To analyze classification accuracy in detail, pixel-wise evaluations and an evaluation matrix were applied to determine the proportion of each class. Fig. 5 shows an example input image, along with the manually annotated ground truth and the prediction generated by the trained U-Net.

The visual comparison provided by Fig. 5 is valuable for identifying both similarities and differences in the classification of relevant areas, which provides essential insights to further refine the computer vision approach. In this case, both ground truth and model prediction are in good alignment, especially for validation images 23 and 25, while small differences around the brittle areas are observed in validation image 15. Given the four classes involved, four distinct class specific IoU values were calculated over all 125 images (see 3.2).

Table 1 displays that IoU values of approximately 90 % were reached for all classes except the validation of brittle fractures. Values of 50 % or higher are typically regarded as positive detections for bounding boxes in the recognition of cars [24], with so-called high-quality detections starting at 70 % [25]. Although these threshold metrics stem from object detection contexts, they provide a useful orientation for evaluating segmentation accuracy in industrial applications. The present work exceeds these values and is more in line with other semantic segmentation results in Material Science. For example, Rosenberger et al. reported IoU values in their results of fracture surface evaluations of around 60 % for ductile fracture surface features and above 90 % for brittle [12]. To further assess model biases evaluation matrices were created for all evaluated pictures. These could lend an insight into class biases, most importantly to biases between ductile and brittle areas, which are of the highest importance in this study.

Table 1
IoU values for the evaluated classes for the training and validation data.

	IoU Ductile	IoU Brittle	IoU Background	IoU Notch	Mean IoU
Training	94.3 %	90.2 %	93.9 %	98.9 %	93.8 %
Validation	90.1 %	82.6 %	93.9 %	98.6 %	90.2 %

Fig. 6 reveals that around 20 % of the total test image corresponds to the fracture surface, which is typical of the preprocessed sample bundle images. Furthermore, for both less than 1 % of the image belongs to the fracture surface not recognized by the model (red) and the wrongly recognized fracture surface (green), confirming the generally good performance of the model. Therefore, the image mostly consists of the background and sample notch, with around 80 %. Differing evaluations between experts and the model can easily be visualized as deviations from the matrix's diagonal. To evaluate the performance of the model, all these areas of the evaluation matrix were investigated and are

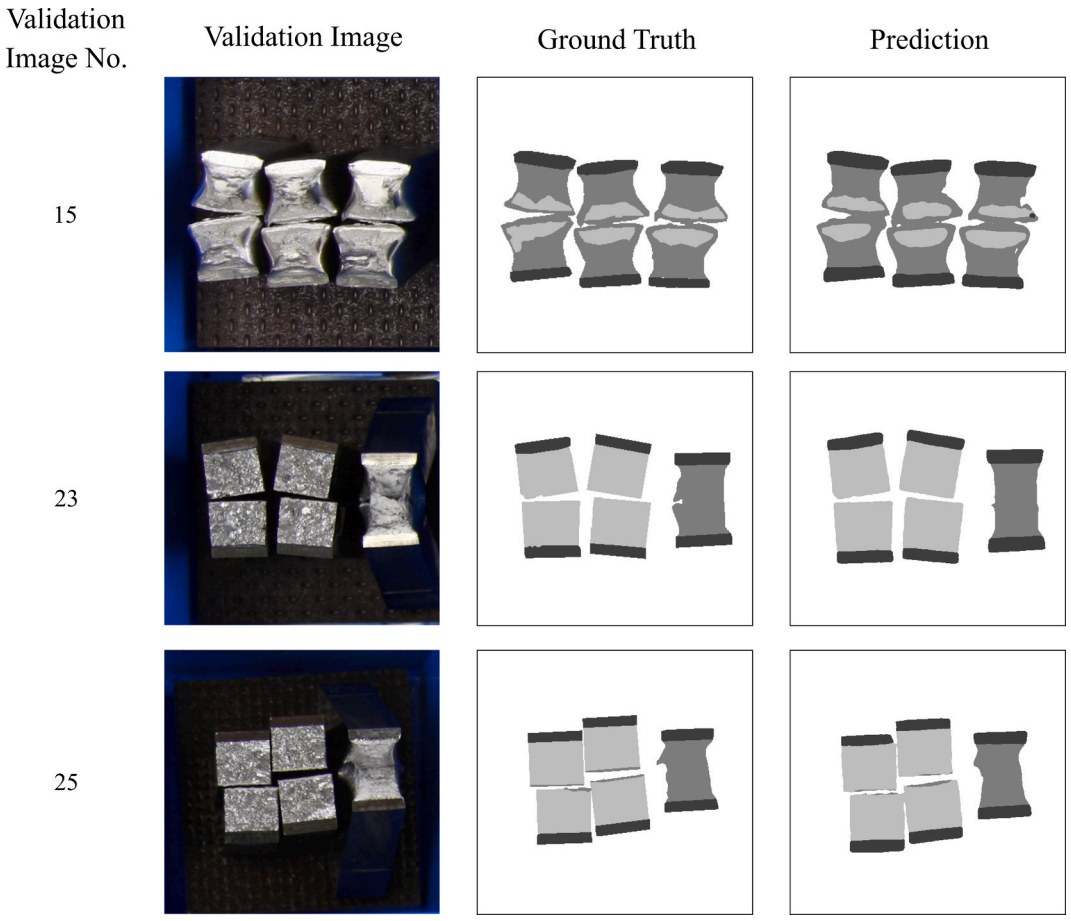


Fig. 5. Semantic segmentation prediction on unseen test images. On the left, the preprocessed test image is shown. The expert-assessed ground truth is presented in the center, and the model's prediction is displayed on the right, with brittle areas in light gray, ductile areas in dark gray, the notch in a black tone, and the background in white. The label applies to both ground truth as well as to the prediction image. The validation image numbers also directly correspond with the labels in Fig. 8.

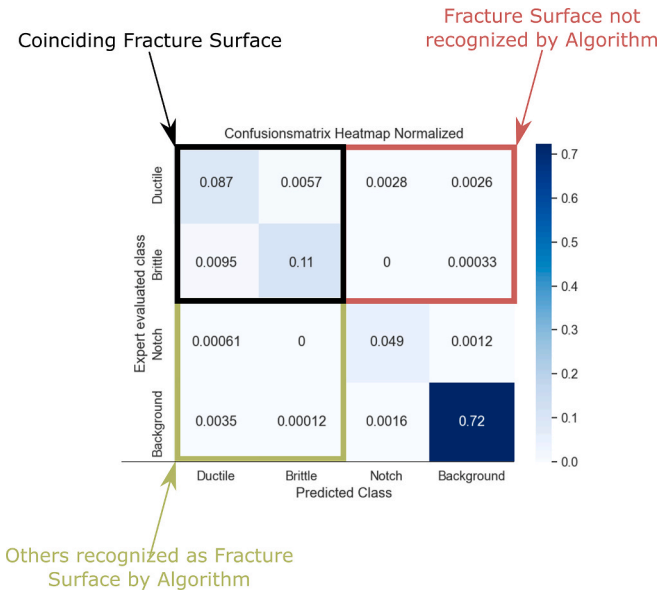


Fig. 6. Evaluation Matrix of an evaluated image comparing model prediction and expert evaluation of the 4 classes: Ductile, Brittle, Notch and Background. Further domain specific subdivisions were executed, showing the agreed upon fracture surface (black), the fracture surface not recognized by the segmentation model (red), the wrongly recognized fracture surface (green) and the non-fracture surface (no color). These diagrams were further evaluated in Fig. 7.

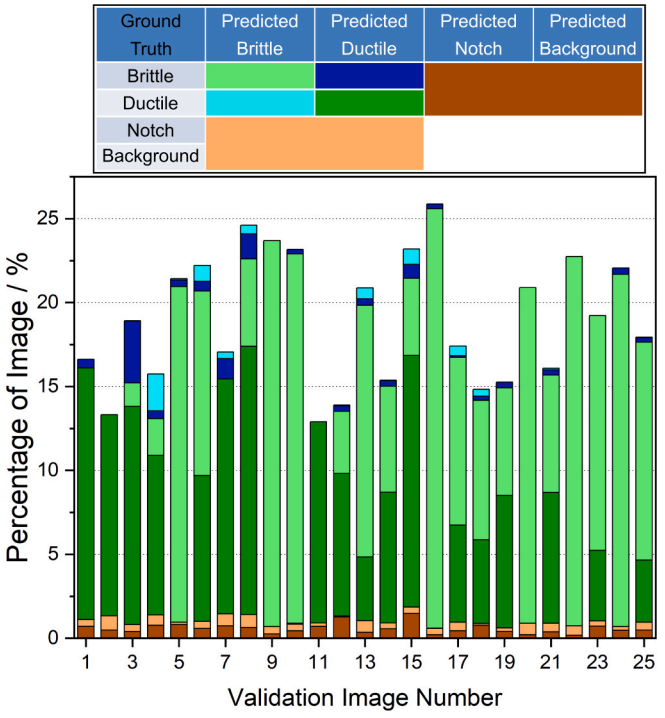


Fig. 7. Comparison of image area using the evaluation matrices of Fig. 6 of the validation dataset. Correct classifications of the fracture surfaces (ductile and brittle) are shown in corresponding shades of green, which resemble the majority of the evaluations, indicating a generally high segmentation quality. Misclassifications within the fracture surface classes, between ductile and brittle, are represented in shades of blue. Fracture surface areas where the model and the expert disagreed on whether the region was part of the fracture surface are marked in shades of brown. The missing fraction to achieve 100% of the entire image corresponds to the non-fracture-surface classes notch and background.

summarized in Fig. 7 for all validation images.

Fig. 7 illustrates that generally there was a slight tendency for the model to miss portions of the fracture surface identified by experts (dark brown) and to overpredict the extent of brittle areas (dark blue) compared to expert assessments. However, as already shown in Table 1, high accuracy for the segmentation results (green) was achieved for all validation samples. Especially, complete ductile and brittle fractures, like Validation Images 11 and 20 were evaluated in accordance with the ground truth, but also the mixed fracture sample bundles, like Validation Image 14 and 18 showed an almost exact match. The two samples with the highest observed evaluation errors are discussed in detail in the following section.

As shown in Fig. 8, the validation images with the highest error reveal two distinct types of discrepancies. In validation image 3, the model predicts a higher percentage of brittle fracture compared to the expert assessment. On closer inspection, only one of the three specimens in the bundle have been evaluated very differently. This specimen displays an unusual appearance, with a reflective, brittle area at the center, surrounded by a wrinkled topography. The other two specimens in this bundle exhibit separations that suggest ductile behavior in the main fracture plane, leading to the hypothesis that the last specimen might also be partially ductile [26]. This fracture morphology was unique within the dataset, posing a challenge for the model's evaluation. One possible explanation for this result could be that the model arbitrarily prefers the brittle class in its segmentation, when encountering difficult image features, therefore also leading to worse IoU values for this class (see Table 1) and more differing evaluations (dark blue in Fig. 7). Furthermore, the present topography increased the reflectivity of the investigated area, which in this case made it visually similar to other brittle fractures. This issue could be addressed by adapting the loss function, therefore reducing the model's inherent bias towards the brittle class and increasing the overall dataset to include more of these types of fractures. In contrast, validation image 4 features three predominantly ductile mixed fractures. The model's prediction indicates good model performance, though the image itself appears slightly blurred, which may contribute to the observed error. Also, small discrepancies can be identified when observing the overall shape of the samples. These can be explained by the image capturing process, making the sample boundary unclear, and therefore leading to small errors in the background evaluation. In summary, utilizing a simple setup for a challenging task, good segmentation results were achieved comparable to previous studies [12] as the discussed examples are only the worst segmentation results obtained. However, it is also obvious that additional data, especially for unusual fracture surfaces appearances, as well as higher resolution and focus depth could enhance the obtained results (see SEM Image Analysis). Furthermore, an improved camera setup and lighting could yield improved model performance. Combining Figs. 7, 8 and Table 1 the following conclusions for sample bundles could be drawn:

- The sample notch IoU values were high for both training and validation images. This is reasonable because the notch is always very similar in shape in the pictures.
- The background values are also high for both training and validation, although small differences compared to expert evaluation could be observed, partially owed to the image capture process and therefore unclear sample boundaries.
- Both ductile and brittle fractures have lower IoU values in the validation data. Fig. 7 illustrates this fact results mostly from few samples in the validation data set and a small overall bias towards brittle. However, the values generally demonstrate the model's strong performance across all categories.

As part of the postprocessing, individual specimens within the sample bundles using the predicted masks were extracted. These masks were overlaid onto the original image, enabling background removal using

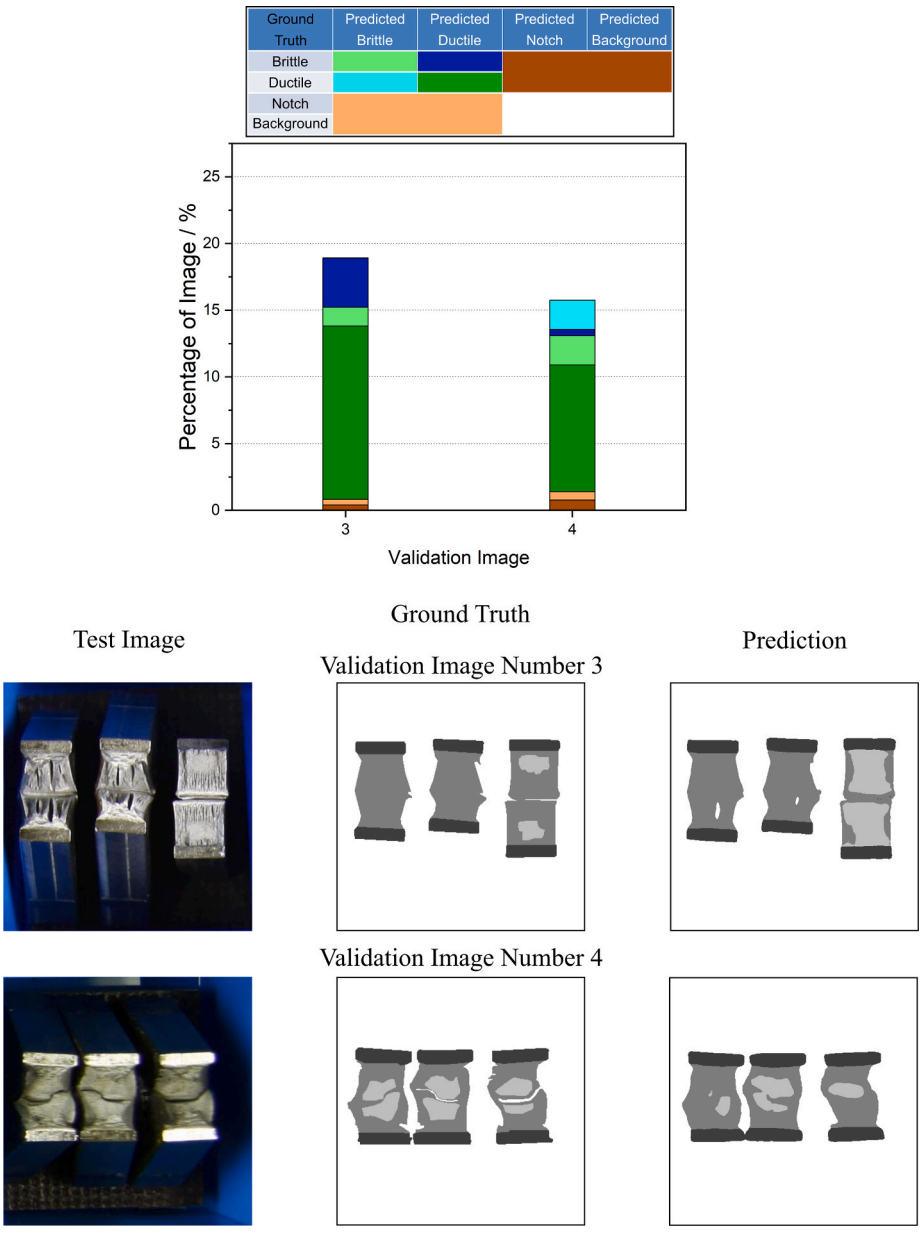


Fig. 8. Sample bundles with the highest error in the validation data (see Fig. 7 Validation Images 3 & 4). The first image shows the specimens with 2 ductile fractures and one mixed fracture. In the second image three mostly ductile mixed fractures can be seen.

Fig. 9. Charpy V-notch specimen using prediction masks to extract the 3 samples from the sample bundle of Fig. 4.

only brittle, ductile and notch areas of the generated mask. Additionally, binary operations such as Erode and Watershed were applied as post-processing to further separate closely positioned specimens, enabling a more detailed examination on a single-specimen level. An example is shown in Fig. 9.

SEM image analysis

In this part, fractographic analysis of Scanning Electron Microscopy (SEM) images of the investigated Charpy V-notch samples was conducted to further validate the quality of the semantic segmentation results on an extracted sample. Fig. 10 shows a comparison between the original image of a sample bundle and labeled SEM images using FractoGraphics v0.60 [10,11]:

The images reveal distinct fracture modes, including a fully ductile fracture (B.), a mixed fracture (C.), and a predominantly brittle fracture,

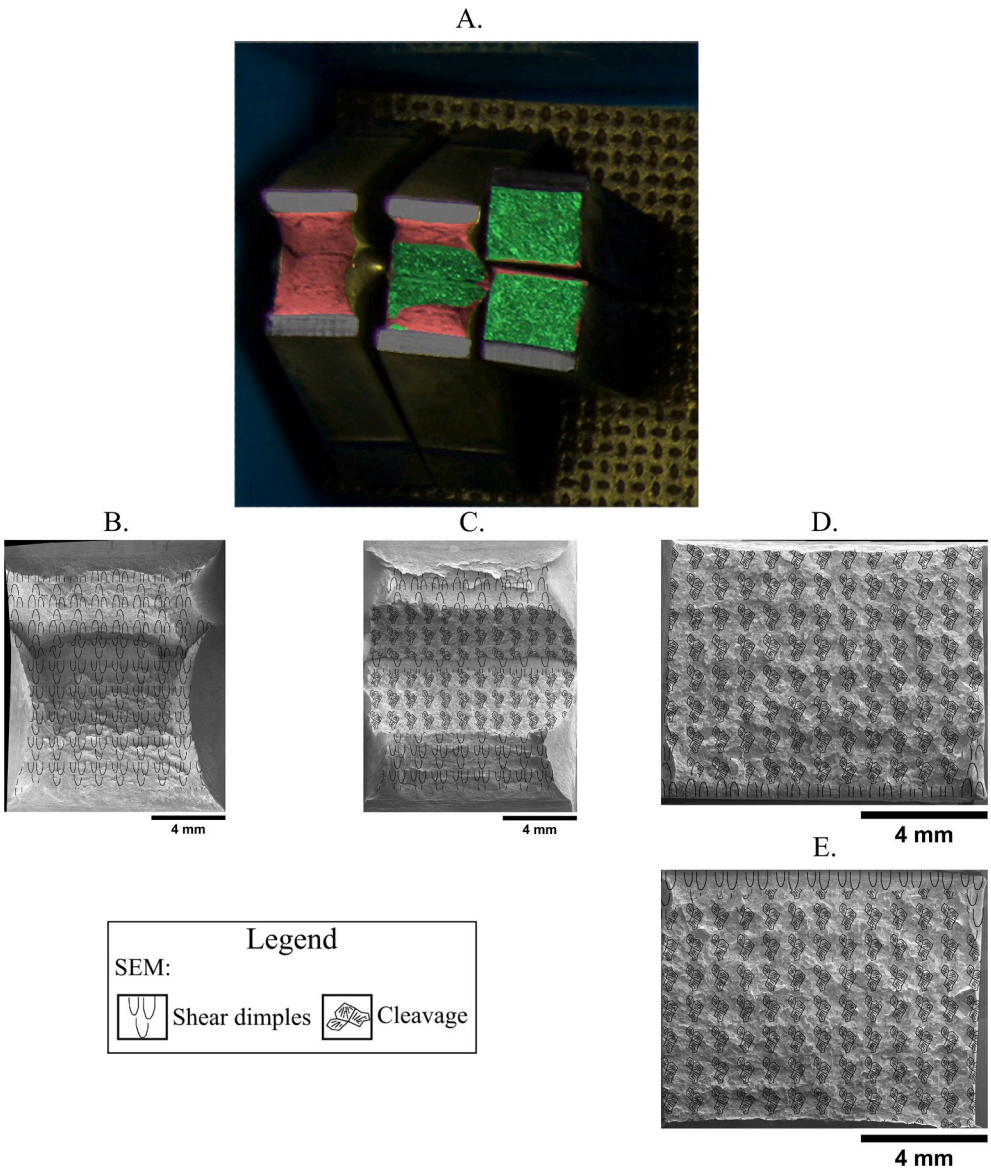


Fig. 10. Original Image (A.) and subsequent SEM analysis using adapted symbols [10,11]. B. shows the uneven spread of dimple orientations in the ductile sample due to its deformation. The mixed fracture (C.) in the middle exhibits a transition from ductile to brittle and back to ductile in the central region of the fracture surface. While the two remaining fracture surfaces (D. & E.) on the right exhibit mostly brittle fracture, except the edge of the samples, with some areas of shear dimples.

leading to two separate fracture surfaces (D. & E.). By incorporating the symbols provided by FractoGraphics v0.60 [10,11], it is possible to analyze the area distributions of the fracture types and the orientations of the dimples in the SEM images, enhancing the visualization of these features. However, a direct comparison of the images highlights distortions caused by the image acquisition process. This discrepancy arises from the angular difference between the digital camera and SEM images, as well as variations in depth of field during image capture. Consequently, image registration was required to align and compare the results accurately. Table 2 illustrates the results based on the registered SEM images.

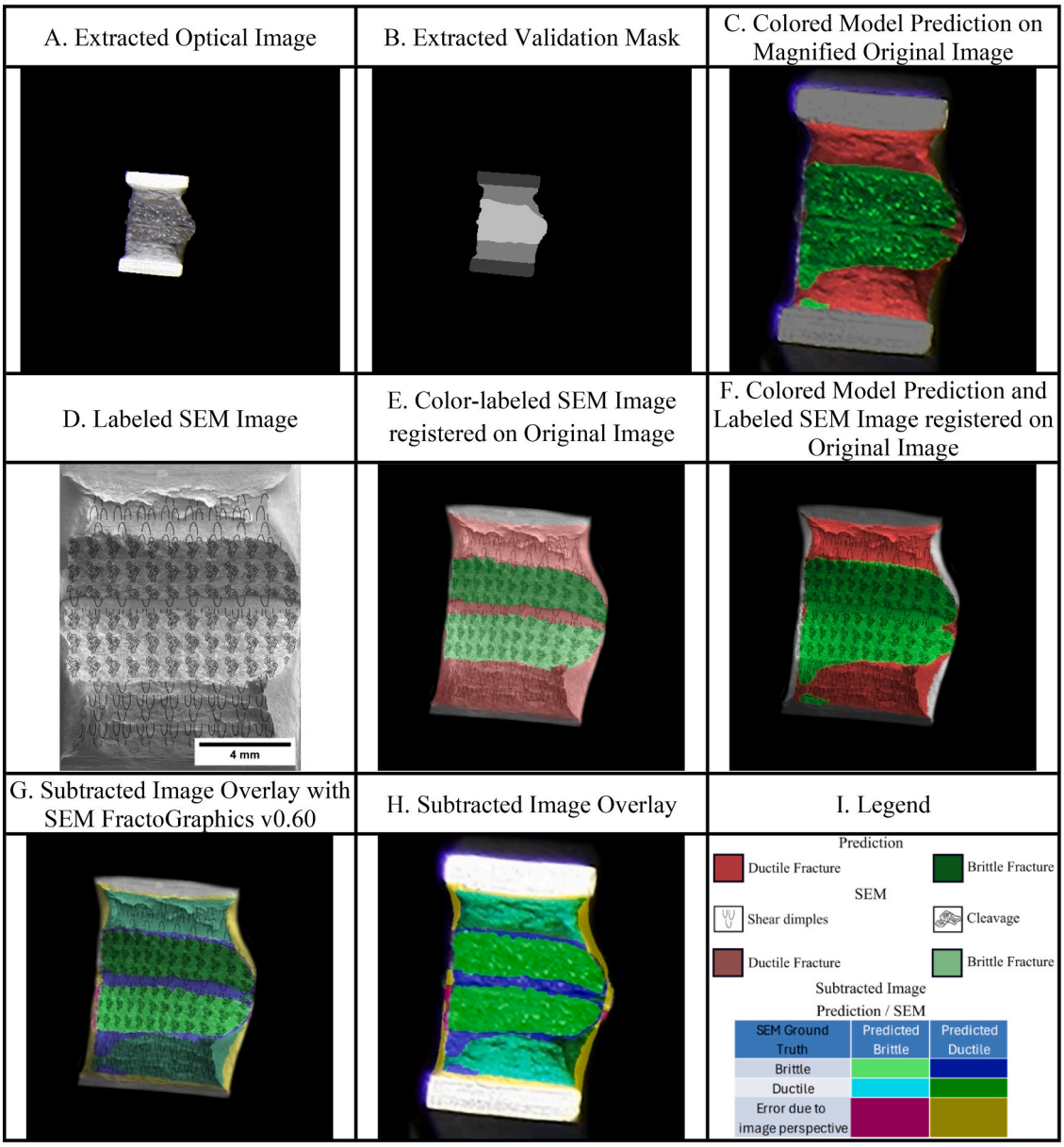
Using the processing pipeline outlined in Fig. 2, the middle sample shown in Fig. 10 was extracted, along with its validation mask (Table 2A. & B.). A direct visual comparison of the validation mask B. with the magnified evaluation results C. demonstrates strong agreement in the assessment of the fracture surface.

Panel D. shows the original, labeled SEM image. Due to inherent distortions when comparing digital camera imaging and SEM imaging, an image registration process was necessary, as illustrated in Panel E., where color labels were applied for improved visualization. By overlaying the SEM labels from D. with the prediction in C., a combined image F. was generated.

From the overlays E. and F., a direct color subtraction was performed to identify areas of disagreement, as shown in Panel G., which is based on the SEM image. In the final step, this subtracted image was also projected onto the digital camera image in H.

The subtracted image in H. highlights areas of agreement (shades of green) and disagreement (shades of purple) between the two evaluations. Extensive green regions indicate a high degree of agreement, confirming the quality of the assessment. Discrepancies were primarily observed near the sample's center and along the brittle fracture edges (dark blue). These differences can be partially attributed to distortions caused by varying imaging techniques. For instance, the ductile area in the center of the sample is clearly identifiable in the top-view SEM image D., while this region appears more obfuscated in the original image A. Additionally, viewing angle differences caused minor distortions near

Table 2
Step-by-step process for evaluation of segmentation results of the model and comparison to the labeled stitched SEM images. Furthermore, the SEM images have been labeled using the symbols from [10,11] to visually label the observed fracture surface. Detailed labels of the images are provided in panel I.



the brittle fracture edges.

Some errors also arose from the image capture process. Areas marked in yellow and red, predominantly near the specimen boundaries, were partially invisible in the original image, leading to gaps in the model's predictions. Finally, a small discrepancy in the bottom-left region of the subtracted image fracture surface reveals overpredicted brittle areas (dark blue) compared to the SEM image. Despite these limitations, the subtracted image effectively visualizes evaluation errors while demonstrating robust performance, even in challenging mixed-fracture regions.

5. Conclusion

In this paper, we present a multi-step machine learning approach to evaluate industrial images of Charpy V-notch specimens. Using color thresholding and connected component analysis, image regions containing three specimens each, referred to as specimen bundles, were successfully extracted from sample boxes. These images then served as a basis to create ground truth assessments, which were subsequently used to train a segmentation model that assigns the four designated classes (background, notch, ductile, and brittle) to each image pixel. Utilizing this method, the model achieved high performance with IoU scores exceeding 82 % for all classes in validation and over 90 % for ductile regions in the validation dataset. These results therefore demonstrate that a machine learning approach can characterize fracture surfaces with high accuracy, offering a reproducible, accurate and operator-independent alternative to traditional visual assessments by experts. Furthermore, customized matrices, optimized for this specific use-case, were utilized to identify possible model biases. This evaluation scheme allowed for the identification of potentially problematic samples

and highlighted the necessity of domain-specific approaches when using AI-based segmentation models. This was particularly beneficial, as standard IoU values alone showed limited applicability when tracking errors across the different specimens and classes for each specimen bundle. In a subsequent step, individual Charpy V-notch specimens were extracted from the previously mentioned images using watershed and morphological operations, enabling a direct comparison with an SEM-based evaluation and highlighting segmentation differences in a pixel-wise manner. This was supported by using symbols from the FractoGraphics v0.60 database [10,11], applied to SEM images which provides a comprehensive understanding of fracture morphology while offering an intuitive and easily visible evaluation method. The primary limitation of this approach lies in the low-resolution imaging of the digital camera, which influences both ground truth assessments and the resulting model performance, especially when compared to the high-resolution SEM-imaging. Improvements such as a higher-resolution camera, enhanced depth of field and a macro lens could mitigate this limitation. In addition, the pixel-wise evaluation of area fractions also presents limitations compared to industry evaluations, with larger deformations in the samples resulting in higher pixel counts for ductile regions compared to brittle, undeformed regions, potentially introducing bias. Furthermore, incomplete fracture specimens obscure parts of the sample, whereas breaking them apart introduces new fracture surfaces after testing. A potential solution could involve a projection-based re-transformation of the data back to the original state of the Charpy V-notch sample geometry to address these issues. Despite these limitations, the presented method achieved robust segmentation results and shows strong potential for objective, scalable fracture analysis. While minor variations in the predicted area fractions may influence fracture classification in certain edge cases, it is important to note that conventional, expert-based assessments are also subject to interpretative variability. In contrast, our approach not only automates this evaluation but also provides a transparent and reproducible pixel-wise basis for comparison and quantification. The integration of SEM-based validation in this study serves to highlight discrepancies and guide interpretation. We therefore see this method as a tool to support, rather than replace, expert evaluation, particularly in high-throughput contexts where manual assessment is impractical, and as a complement to detailed SEM investigations in critical individual cases. Future work may also include the integration of SEM-based ground truth datasets to improve the labeling accuracy, as well as the extension to other fracture sample types, such as CTOD, BDWTT or CT samples, which would broaden its applicability in mechanical testing scenarios.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT to improve the language and readability of the manuscript. Additionally, ChatGPT was used to support the debugging and troubleshooting of code. After using this tool, the authors carefully reviewed and edited all content as needed and take full responsibility for the content of the publication.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Sebastian Scholl is currently employed at “AG der Dillinger Hüttenwerke”. While his expertise contributed to parts of the research, his employment had no undue influence and does not represent a conflict of interest. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Acknowledgments

A special thank you goes to Dr.-Ing. Christoph Pauly for the development of the python script for the automatic image capture process at the SEM. Furthermore, our thanks go to Timm Evers and Dr.-Ing. Simon Olschok for making the electron beam welds. The authors would also like to thank the German Research Foundation and the Saarland government for funding the PFIB/REM (INST 256/510-1 FUGG), on which the SEM images were obtained.

Funding

The authors would like to thank the Federal Ministry of Economics and Climate Protection for funding the BMWK project “Development of heavy plates for the high-performance welding of monopiles for the construction of wind offshore energy systems - HL-Blech” (grant number 03EE2028A) based on which this research was conducted.

Data availability

The data presented in this study are not readily available because they are part of an ongoing study. Requests to access the dataset should be directed to Adrian Herges (adrian.herges@uni-saarland.de).

References

[1] Y.J. Chao, J.D. Ward, R.G. Sands, Charpy impact energy, fracture toughness and ductile–brittle transition temperature of dual-phase 590 Steel, Mater. Des. 28 (2007) 551–557, https://doi.org/10.1016/j.matdes.2005.08.009.
[2] C. Shang, D. Zhu, H.-H. Wu, P. Bai, F. Hou, J. Li, S. Wang, G. Wu, J. Gao, X. Zhou, T. Lookman, X. Mao, A quantitative relation for the ductile-brittle transition temperature in pipeline steel, Scr. Mater. 244 (2024) 116023, https://doi.org/10.1016/j.scriptamat.2024.116023.
[3] T.C. Park, B.S. Kim, J.H. Son, Y.K. Yeo, A new fracture analysis technique for Charpy impact test using image processing, Korean J. Metals Mater. 59 (2021) 61–66, https://doi.org/10.3365/KJMM.2021.59.1.61.
[4] A. Faisal, M. Kamruzzaman, T. Yigitcanlar, G. Currie, Understanding autonomous vehicles, J. Transp. Land Use 12 (2019) 45–72, https://doi.org/10.2307/26911258.
[5] G. Karayegen, M.F. Aksahin, Brain tumor prediction on MR images with semantic segmentation by using deep learning network and 3D imaging of tumor region, Biomed. Signal Process. Control 66 (2021), https://doi.org/10.1016/j.bspc.2021.102458.
[6] Xiaodan Wu, Jianwen Wang, A. Flitman, P. Thomson, Neural and machine learning to the surface defect investigation in sheet metal forming, in: ICONIP’99. ANZIIS’99 & ANNES’99 & ACNN’99. 6th International Conference on Neural Information Processing. Proceedings (Cat. No.99EX378), IEEE, n.d.: pp. 1088–1093. https://doi.org/10.1109/ICONIP.1999.844687.
[7] B.-I. Bachmann, M. Müller, D. Britz, A.R. Durmaz, M. Ackermann, O. Shchyglo, T. Staudt, F. Mücklich, Efficient reconstruction of prior austenite grains in steel from etched light optical micrographs using deep learning and annotations from correlative microscopy, Front. Mater. 9 (2022), https://doi.org/10.3389/fmats.2022.1033505.
[8] M. Müller, D. Britz, L. Ulrich, T. Staudt, F. Mücklich, Classification of bainitic structures using textural parameters and machine learning techniques, Metals (Basel) 10 (2020) 630, https://doi.org/10.3390/met10050630.
[9] L. Schmies, B. Botsch, Q.-H. Le, A. Yarysh, U. Sonntag, M. Hemmleb, D. Bettge, Classification of fracture characteristics and fracture mechanisms using deep learning and topography data, Pract. Metallogr. 60 (2023) 76–92, https://doi.org/10.1515/pm-2022-1008.
[10] D. Bettge, L. Schmies, The WG fractography online database – stage of development and planning, Pract. Metallogr. 60 (2023) 569–579, https://doi.org/10.1515/pm-2023-0048.
[11] FractoDB, FractoGraphics v0.60, (n.d.). https://www.fraktographie.bam.de/php/fraktographie/show_view1.php?q=PT006-Symbolik. (accessed May 26, 2025).
[12] J. Rosenberger, J. Tlatlik, S. Münstermann, Deep learning based initial crack size measurements utilizing macroscale fracture surface segmentation, Eng. Fract. Mech. 293 (2023), https://doi.org/10.1016/j.engfracmech.2023.109686.
[13] J. Rosenberger, J. Tlatlik, N. Rump, S. Münstermann, Prediction of statistical force–displacement curves of Charpy-V impact tests based on unsupervised fracture surface machine learning, Eng. Fail. Anal. 175 (2025) 109551, https://doi.org/10.1016/j.engfailanal.2025.109551.
[14] A. Herges, L. Ulrich, S. Scholl, M. Müller, D. Britz, F. Mücklich, Machine learning for the classification of macroscale fracture surfaces, Pract. Metallogr. 60 (2023) 352–362, https://doi.org/10.1515/pm-2023-1042.
[15] G. Bradski, A. Kaehler, The Opencv Library, first ed., O’Reilly Media Inc., 2008.
[16] Iakubovskii P., Segmentation models Pytorch. GitHub Repository., (2019). https://github.com/qubvel-org/segmentation_models.pytorch (accessed May 26, 2025).

[17] S. van der Walt, J.L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J.D. Warner, N. Yager, E. Gouillart, T. Yu, scikit-image: image processing in Python, PeerJ 2 (2014) e453.
[18] The GIMP Development Team, GNU Image Manipulation Program (GIMP), Version 2.10.38. Community, Free Software (license GPLv3), (2024). https://gimp.org/ (accessed May 26, 2025).
[19] D.A. van Dyk, X.-L. Meng, The art of data augmentation, J. Comput. Graph. Stat. 10 (2001) 1–50, https://doi.org/10.1198/10618600152418584.
[20] A.R. Durmaz, M. Müller, B. Lei, A. Thomas, D. Britz, E.A. Holm, C. Eberl, F. Mücklich, P. Gumbsch, A deep learning approach for complex microstructure inference, Nat. Commun. 12 (2021) 6272, https://doi.org/10.1038/s41467-021-26565-5.
[21] O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional Networks for Biomedical Image Segmentation, in: 2015: pp. 234–241. https://doi.org/10.1007/978-3-319-24574-4_28.
[22] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, Li Fei-Fei, ImageNet: a large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255, https://doi.org/10.1109/CVPR.2009.5206848.
[23] G. Huang, Z. Liu, L. van der Maaten, K.Q. Weinberger, Densely Connected Convolutional Networks, (2016). http://arxiv.org/abs/1608.06993.
[24] A. Geiger, P. Lenz, R. Urtasun, Are we ready for autonomous driving? The KITTI vision benchmark suite, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 3354–3361, https://doi.org/10.1109/CVPR.2012.6248074.
[25] Z. Cai, N. Vasconcelos, Cascade R-CNN: high quality object detection and instance segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 43 (2021) 1483–1498, https://doi.org/10.1109/TPAMI.2019.2956516.
[26] S. Scholl, A. Schneider, V. Schwinn, Significance of separations occurring in mechanical testing of thermomechanical rolled pipeline steels, J. Pipeline Sci. Eng. 2 (2022) 100070, https://doi.org/10.1016/j.jpse.2022.100070.

Glossary

CCA: Connected Component Analysis
IoU: Intersection over Union
SEM: Scanning Electron Microscope