DOI 10.36146/2026_01_10

Deep learning-powered system for automated detection and quantification of Vickers indentations

F. Bertolini, M. Mariani, N. Lecis

Hardness testing is a key procedure in materials science for evaluating mechanical properties and process quality. Traditional Vickers hardness measurement relies on manual identification of indentation diagonals, a process that is slow, subjective, and prone to variability. This work introduces a deep learning-based pipeline for fully automated Vickers hardness measurement, combining instance segmentation via Mask R-CNN with sub-pixel geometric fitting for diagonal extraction. A dataset of 403 micrographs of samples under loads from 10 gf to 2000 gf was assembled and annotated for training and validation. Hyperparameter optimisation was performed using a Taguchi design of experiments, and the final model achieved near-perfect segmentation accuracy (overall AP ≈ 90.5%) on the test set. Measurement accuracy was assessed against manual ground truth, yielding mean relative errors of 1.6-1.9% for the two diagonals, with most cases within 2-3%. These results demonstrate that the proposed system provides robust detection, high metrological precision, and reproducible performance across diverse imaging conditions, paving the way for reliable, high-throughput hardness testing in industrial and research settings.

KEYWORDS: MACHINE LEARNING; ARTIFICIAL INTELLIGENCE; VICKERS HARDNESS; METALLOGRAPHY; MASK R-CNN; TAGUCHI DOE; DEEP LEARNING;

INTRODUCTION

Hardness testing remains a fundamental procedure in materials science and engineering for assessing material quality, mechanical properties and process efficacy [1]. This method employs a pyramidal diamond indenter with a 136° angle, and the resulting hardness number is derived from the applied test force (F) and the measured average diagonal length (d) of the approximately square-shaped indentation [2]. Hardness testing is valuable because the measured metric tends to correlate with key mechanical properties such as tensile strength, ductility and wear resistance, thus providing information on the effects of thermomechanical processing [1, 2]. Accurate measurement of the indentation diagonal in the Vickers test is essential: owing to the quadratic dependence on the diagonal length, any error in its measure is amplified in the final value [2].

Although routine, manual measurement of Vickers indentations has its drawbacks. The process is tedious and repetitive, and for each indentation it may take

Francesco Bertolini, Marco Mariani, Nora Lecis
Politecnico di Milano, Italy

a couple of minutes, making it impractical for high-throughput settings. Moreover, when an operator measures the diagonal length, the result depends on their viewpoint, skill and fatigue, which introduces subjectivity and variability [1, 3, 4]. In addition, the specimen and image quality add complications: the indentation edges are not always sharply defined and may appear more like shadows than crisp lines. Real-world conditions further undermine accuracy: variable lighting, reflective or rough surfaces, etching marks, and defects such as grooves, cracks or pile-up/sink-in around the indent all affect the measurement [1, 3, 4].

Classical automated methods based on traditional computer-vision techniques, such as image thresholding, edge detection or Hough-transform-based approaches, can be useful but face limitations in many practical indentation-measurement settings [5, 6, 7, 8, 9, 10]. For example, thresholding often works when the indentation contrasts cleanly with the background, but it becomes unreliable if illumination is uneven or the contrast is low. Similarly, Hough-transform or straight-edge detectors may yield acceptable results when indentation boundaries are crisp and well aligned, but their accuracy decreases when the actual boundaries are curved (due to pile-up/sink-in or surface preparation) or when the indentation is rotated relative to the image axes. Overall, while these classical techniques can perform adequately under controlled conditions, their robustness across the wide variety of materials, surface finishes and imaging conditions found in industrial practice remains limited. Some improved methods, including active-contour models, focus-assessment routines and custom image-processing pipelines, extend capabilities further, but they still often require fine-tuning of parameters (illumination, focus, threshold) and may be sensitive to surface artefacts or process variations [11, 12, 13, 14].

To overcome the robustness limitations of classical automation, Convolutional Neural Networks (CNNs) have been increasingly investigated for Vickers and Brinell indentation analysis, offering improved automatic feature extraction capabilities [15, 16, 17, 18, 19, 20, 21]. Approaches vary from directly predicting the Vickers hardness value via regression to image-processing pipelines utilising CNNs or Fully Convolutional Networks (FCNs) for indentation localisation and segmentation. Object detection models such as Faster R-CNN- or YOLO-based variants have been employed to predict the indentation as a bounding box, typically as an intermediate localisation step within broader pipelines [16,17]. While effective for initial localisation, the bounding-box approach fundamentally limits precision, especially for slightly rotated or irregularly shaped indentations, as it does not capture pixel-level contour information that is essential for accurate metrology [6,16]. Other segmentation-based methods, including FCNs and active-contour-assisted schemes, aim to predict pixel-level masks and achieve good performance in controlled conditions [6, 8, 16]. However, in practical metallographic micrographs, the imprint boundary can be degraded by heterogeneous microstructures, polishing scratches, debris, and pile-up/sink-in effects, so that the limiting factor becomes accurate boundary delineation rather than coarse localisation. In these cases, semantic masks or corner-only regression may under-represent locally distorted or concave edges, and small boundary errors can propagate nonlinearly into diagonal estimation and hardness due to the quadratic dependence on diagonal length. Deep learning approaches in general have shown clear potential for automating hardness indentation evaluation, but many reported studies focus on either relatively small datasets or in ideal conditions (e.g. reference hardness blocks under controlled imaging), so their behaviour on more heterogeneous materials remains less systematically explored [11,14,16,17]. Table 1 summarises the main recent deep-learning approaches and compares them with the present work.

The goal of this work is to present a robust, pragmatic, and high-accuracy pipeline for automatic Vickers hardness test detection and analysis. We achieve this by combining the precision of a Mask R-CNN-based instance segmentation model for pixel-accurate indentation boundary detection with a dedicated geometric fitting procedure for diagonal extraction. Mask R-CNN, an instance segmentation framework, provides precise segmentation masks superior to bounding-box approximations or corner-only detection, thereby addressing the crucial problem

of accurate boundary delineation in noisy images [22]. To ensure the robustness and reproducibility of the system's performance across diverse operational settings, we systematically investigate the influence of various training conditions using a Taguchi L16 Design of Experiments (DoE) approach [23, 24]. This paper presents this integrated methodology as a demonstrated, alternative, and practical route to achieve objective and reliable automatic Vickers test detection and hardness values measurement.

Tab.1 - Comparison of recent deep-learning approaches for automated Vickers indentation analysis, reporting architecture, diagonal (or hardness) extraction strategy, dataset setting, and the main accuracy metric as reported in each study.

Comparison with prior works				
Study	Architecture	Diagonal extraction	Data (size & setting)	Reported measurement accuracy
Tanaka et al. (2020) [17]	Dual CNN (rough BB + corner refinement)	Two-stage regression (pixels)	Large datasets (e.g., 4140+2400 and 3840+3200 images)	Diagonal MAE ≈ 0.4–2.0 μm
Jalilian & Uhl (2021) [21]	FCN (RefineNet)	Linear curve fitting + ROI vertex refinement	Two industrial datasets: DA=150, DB=216 images (1280×1024)	Diagonal MAE: 7.03 px (DA), 3.24 px (DB)
Li & Yin (2021) [18]	FCN-ED (U-Net-based)	Oriented bounding box (OBB) on predicted mask	Augmented dataset 12,000 (8:1:1 split); boundary annotated via sampled points	Diagonal MAE ≈ 0.5–5.4 μm; max relative error ≈ 0.39–1.67%
Cheng et al. (2022) [19]	Multi-task learning (MTL) CNN	Direct hardness prediction (regression)	105 base indentations; augmented mixed images (train/val/test 5000/500/100); +59 unseen images	Hardness MAE ≈ 19.7 HV
Privezentsev et al. (2019) [20]	Hybrid (object detection + image processing)	Contour selection from detected imprint	108 indentations	Geometrical relative MAE < 4%
This work	**Mask R-CNN (instance segmentation)**	**Sub-pixel fitting via signed distance fields**	**403 images (10–2000 gf); COCO masks**	**Diagonal MAE ≈ 2.3–2.7 px; relative MAE ≈ 1.8–2.0%**

MATERIALS AND METHODS

Dataset Acquisition and Preparation

A dataset of Vickers microhardness indentations was assembled from laboratory measurements conducted with an FM-180 microindenter by FUTURE-TECH CORP on polished metallic and ceramic samples under different loads, ranging from 10 gf up to 2000 gf. Micrographs were acquired using an optical microscope integrated within the hardness tester. Each image contained one or two indentations exhibiting typical variations in contrast, surface finish, and minor optical artefacts commonly encountered in metallographic imaging. A total of 403 images were collected.

All micrographs were manually annotated using 4-point polygonal masks tightly enclosing each indentation, through LabelMe. Annotations were exported in COCO-compatible format to enable direct use within the Mask R-CNN framework. The dataset was then divided into independent training (75%), validation (15%), and test (10%) subsets, ensuring that no visually similar images appeared across different splits (three-way split) [25].

Model Architecture

Indentation segmentation was performed using a Mask R-CNN architecture [22], employing a ResNet-101 backbone and Feature Pyramid Network (FPN). This

configuration extracts multi-scale features to ensure robustness across indentation sizes, enabling the two-stage pipeline to jointly localise and segment instances at the pixel level.

This design, originally proposed for high-precision instance segmentation tasks, is particularly suitable for Vickers impressions, where accurate delineation of the indentation edges is required for geometric measurement. Compared with single-stage object detectors, the two-stage Mask R-CNN paradigm typically provides higher segmentation fidelity, which is essential for the subsequent extraction of diagonals.

The network weights were initialised from a COCO-pretrained model to leverage generic visual features. Only one object class ("indentation") was used.

Model Training and Hyperparameters Optimisation

The objective of training was to obtain high-fidelity segmentation masks while maintaining sufficient recall to detect all impressions present in an image. Hyperparameters influencing convergence and mask quality were explored through a structured Taguchi design of experiments ($L_{16}(4^4)$) [26], enabling systematic variation of four key factors: learning rate, weight decay, number of training epochs, and the RPN non-maximum suppression threshold. This design allowed systematic sampling of the hyperparameter space while limiting the total number of training runs to sixteen. Each configuration (see table 2) was trained independently on the same train/validation split, and segmentation performance was quantified on the validation set using the COCO mask average precision (segmAP).

To ensure that performance differences observed across the DoE were not attributable to stochastic training variability, each of the sixteen hyperparameter configurations was trained three times with different random seeds; occasional unstable runs were discarded and replaced with the mean of the corresponding stable repetitions. The three validation scores obtained for each configuration were then aggregated (mean and variance), providing a more reliable estimate of the true performance associated with each hyperparameter combination.

The hyperparameter configuration yielding the highest validation segmentation accuracy was selected for final training. The final model was retrained on the combined training and validation sets using the optimal configuration identified through the Taguchi analysis and compared against the baseline model (retrained on the same combined training and validation data).

Geometric Measurements and Hardness Computation

After instance segmentation, each detected indentation mask was processed by a geometric fitting routine to recover the two Vickers diagonals using the OpenCV library [27]. For each detected indentation mask, a signed distance field was computed from the binary region using standard distance-transform formulations [28], [29], and the 0-level isocontour was extracted via a marching squares scheme [30]. The resulting sub-pixel contour was partitioned into four arcs using the top, bottom, left, and right extrema, each arc corresponding to one side of the rhomboidal imprint. A straight line was then fitted to each arc using an orthogonal (total least-squares) regression [31]. Intersections between adjacent fitted lines yielded four sub-pixel vertices of the indentation.

The two Vickers diagonals were obtained from opposite vertex pairs. Their arithmetic mean, d, was used in the standard Vickers hardness as in equation 1 [2]:

$$HV = 1.8544 \frac{F}{d^2} \qquad [1]$$

Where F is the applied load in kgf and d is expressed in mm.

Basic quality-control criteria were applied, excluding cases where the predicted shape was excessively distorted, too small for reliable measurement, or located near the image boundary. Surviving instances were retained for hardness computation and downstream analysis.

Tab.2 - Taguchi Design of Experiment employed for the hyperparameter optimisation.

Taguchi DoE ($L_{16}4^4$)				
Run	**base_lr**	**weight_decay**	**max_epochs**	**rpn_nms_thresh**
0	5e-4	5e-6	5	0.3
1	5e-4	1.7e-4	10	0.5
2	5e-4	3.35e-4	15	0.7
3	5e-4	5e-4	20	0.9
4	7e-3	5e-6	10	0.7
5	7e-3	1.7e-4	5	0.9
6	7e-3	3.35e-4	20	0.3
7	7e-3	5e-4	15	0.5
8	1.35e-2	5e-6	15	0.9
9	1.35e-2	1.7e-4	20	0.7
10	1.35e-2	3.35e-4	5	0.5
11	1.35e-2	5e-4	10	0.3
12	2e-2	5e-6	20	0.5
13	2e-2	1.7e-4	15	0.3
14	2e-2	3.35e-4	10	0.9
15	2e-2	5e-4	5	0.7
Baseline	**2.5e-4**	**1e-4**	**12**	**0.7**

Evaluation Protocol

Segmentation performance was evaluated using COCO mask average precision (segmAP). During the hyperparameter study, the mean validation segmAP over three training seeds was used as the response variable for each Taguchi configuration, with standard deviation as an indicator of stability. For the baseline and final models, segmAP and size-specific APs (AP_{50}, AP_{75}, AP_s, AP_m, AP_l) were computed on the independent test set.

Diagonal-measurement accuracy was assessed on the test subset with manual reference diagonals as ground truth. For each detected indentation, d_1 and d_2 were compared through absolute and relative errors, tolerance-band statistics, correlation coefficients and Bland-Altman analysis. An additional multiphase microstructure example was used to show the model performance; the pixel-to-micron conversion was manually calculated from the scale bar in the image and given to the code as input.

RESULTS AND DISCUSSION

Baseline model training

The baseline Mask R-CNN model was first trained using the default hyperparameter configuration described in the Methods section. The training progressed smoothly, as shown in figure 1, with all loss components decaying

monotonically and stabilising after the first few hundred iterations. The absence of oscillations or divergence indicates a well-behaved optimisation process even without any hyperparameter tuning.

On the validation set, the baseline model reached a segmentation AP above 86%, with almost perfect AP_{50} and AP_{75} above 97%, confirming that a standard configuration already provides robust indentation detection (tab. 3). Size-specific APs also remained consistently high, suggesting that the network generalised well across the typical range of indentation dimensions encountered in the dataset.

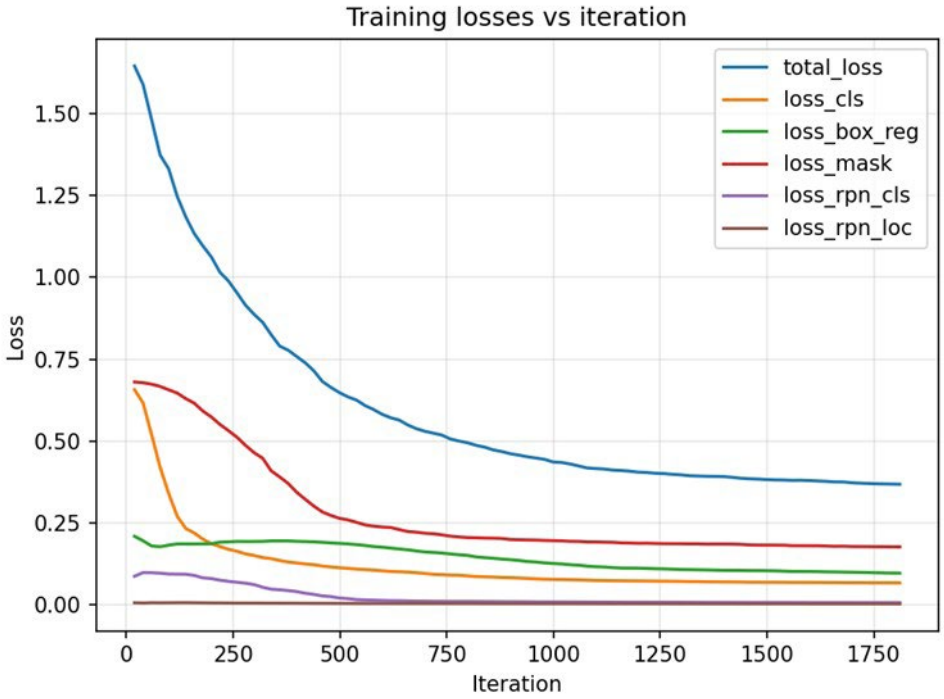

Fig.1 - Training loss profiles vs. iteration number for the baseline model trained on the train subset.

Tab.3 -Segmentation performance of the baseline model on the validation subset.

Baseline on validation subset						
Run	**segmAP [%]**	**AP50 [%]**	**AP75 [%]**	**APs [%]**	**APm [%]**	**APl [%]**
Baseline	86.5±0.1	99.7±0.3	95.4±0.1	83.3±0.4	83.3±0.1	90.9±0.1

Hyperparameters optimisation (Taguchi DoE)

Across the 16 Taguchi configurations, the mean validation AP spanned a wide range, from mid-84% for the weakest settings to more than 93% for the best-performing ones. This substantial spread already indicates a strong dependence of segmentation accuracy on the chosen hyperparameters. For each configuration, the average AP and its standard deviation across seeds are reported in table 4. High-performing runs generally exhibit low variability, whereas lower-performing setups show larger fluctuations or, in a few cases, instability. Divergence occurred in two repetitions of run 14 and in one repetition of run 15, exclusively in configurations combining relatively high learning rates with short training schedules or insufficient optimisation time. These unstable repetitions were excluded and replaced by the mean of the valid runs.

The ANOVA (tab. 5) identifies max_epochs as the dominant factor ($F \approx 9.51$, $p \approx 0.048$), and the main-effects re-

sponse table (fig. 2) confirms this, showing the largest Δ (≈ 6.7 AP points) and a strong monotonic increase in performance from 5 to 20 epochs. The RPN NMS threshold is the second most influential factor, with a Δ ≈ 3.1 AP points. Intermediate NMS levels provide the best balance between suppressing redundant proposals and retaining closely spaced indentations.

In contrast, base learning rate and weight decay show considerably smaller Δ values (≈ 1.9 and 1.3 AP points, respectively) and no statistically significant effects in the ANOVA ($p > 0.5$ for both). The small coefficients of these terms in the linear model further confirm that, within the tested ranges, the model is relatively insensitive to moderate variations of these two parameters. Overall, the response analysis demonstrates that adequate training duration and appropriate NMS filtering are the key drivers of segmentation performance, while learning rate and weight decay exert only secondary, fine-tuning effects. Accordingly, the best-performing configuration identified by the Taguchi design corresponds to a learning rate of 0.007, a weight decay of 3.35×10^{-4}, a training schedule of 20 epochs and an RPN NMS threshold of 0.30.

Tab.4 - Validation segmentation AP (mean ± std) for each Taguchi DoE run over three random seeds.

Taguchi DoE ($L_{16}4^4$)			
Run	**SegmAP [%]**	**Run**	**SegmAP [%]**
0	85.6±0.3	8	90.3±0.4
1	88.1±1.1	9	90.5±2.2
2	90.4±0.5	10	86.4±2.2
3	91.4±0.3	11	92.6±1.3
4	90.5±1.4	12	92.8±1.3
5	85.5±1.9	13	92.9±1.1
6	93.6±0.6	14	84.9±0
7	92.4±0.9	15	84.0±4.4

Tab.5 - Analysis of variance (ANOVA) for the Taguchi design.

Taguchi DoE ($L_{16}4^4$)			
Source	**Adj Mean Square**	**F-value**	**p-value**
base_lr	3.138	0.81	0.568
weight_decay	1.347	0.35	0.796
max_epochs	36.973	9.51	0.048
rpn_nms_thresh	7.361	1.89	0.307
Residual Error	3.887		

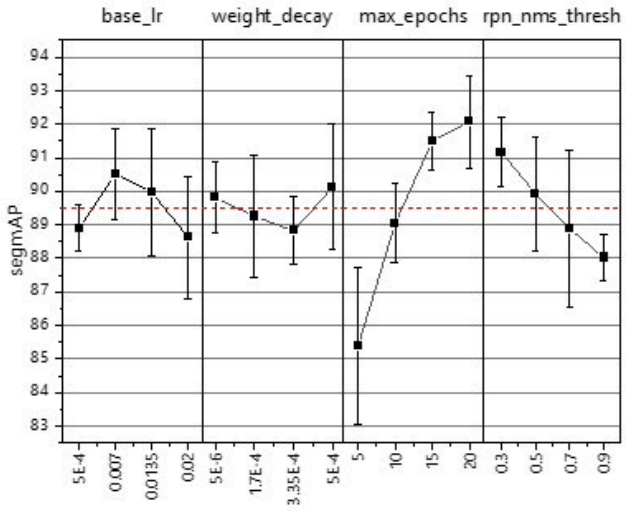

Fig.2 - Average Precision dependence over the analysed model factors optimised by Taguchi DoE.

Best model training and comparison with baseline

The best-performing hyperparameter configuration identified by the Taguchi design was then used to retrain the final Mask R-CNN model on the combined training and validation sets. A direct comparison of the training dynamics in figure 3 highlights the substantial impact of this optimisation on the learning behaviour. In the baseline run (letft panel), convergence is relatively gradual: the total loss decreases slowly and requires on the order of 1000 iterations to reach a stable plateau at about 0.38. By contrast, the optimised model (right panel) exhibits much more efficient learning, with the total loss dropping steeply within the first ≈250 iterations and stabilising at a markedly lower value, close to 0.20.

In addition to this overall reduction, the mask loss remains consistently lower for the optimised configuration throughout training. This indicates that the tuned hyperparameters enable the network to resolve indentation boundaries with higher fidelity and confidence, effectively reducing pixel-level segmentation errors that would otherwise propagate into the diagonal measurements and, ultimately, into the computed hardness values.

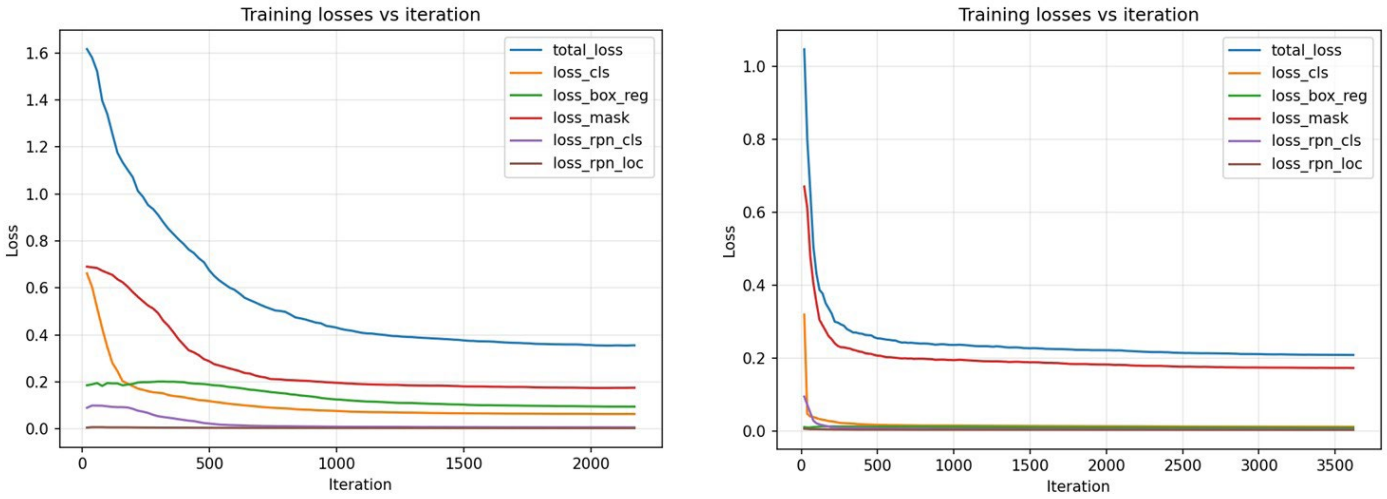

Fig.3 - Comparison of training loss profiles vs. iteration number for the baseline (left) and the optimised model (right), trained on combined training and validation data.

Even in its baseline configuration, the Mask R-CNN demonstrates strong segmentation capabilities: it consistently detects indentation regions and produces masks that align reasonably well with the true imprint geometry. This behaviour is evident in the top panel of figure 4, where the baseline model correctly identifies and segments most indentations, yielding visually coherent masks for medium and large imprints.

However, qualitative inspection also highlights clear limitations of the baseline model, particularly on more challenging samples. In low-contrast regions or in the presence of strong background texture, polishing scratches or debris, the predicted boundaries tend to be slightly irregular and, in some cases, the model produces spurious detections. While these issues do not drastically undermine average performance, they are critical from an operational standpoint: in an industrial context, the system must be highly reliable, and any false positive or irregular mask is unacceptable, as it directly corrupts the subsequent diagonal estimation and hardness computation.

The optimised model, obtained through the Taguchi hyperparameter exploration and final retraining, mitigates these weaknesses. The bottom panel of figure 4 shows the corresponding predictions from the best model for the same three test images. In the first example, the optimised model produces a cleaner and more tightly aligned mask around the indentation edges, markedly reducing the small boundary irregularities still visible in the baseline output. In the second, more challenging example, where the baseline model produced two false positives, the optimised model correctly identifies a single indentation with no spurious detections. In the third example, representing a large and well-defined imprint, both models perform well, but the optimised model exhibits sharper contour definition and a more consistent alignment between the mask and the underlying imprint geometry.

Quantitatively, the optimised configuration delivers a clear and consistent improvement over the baseline on the test subset, as summarised in table 6. In the COCO framework, the overall AP is the primary summary metric: it averages detection performance over a range of IoU thresholds (typically from 0.50 to 0.95), so it rewards models that are not only able to detect objects but also to delineate them accurately across different levels of overlap. A higher AP therefore indicates a globally more reliable segmentation behaviour, both in terms of finding indentations and in terms of matching their true shape.

The individual components AP_{50} and AP_{75} provide additional insight. The first measures performance at a relatively loose overlap threshold (IoU ≥ 0.5), reflecting the ability of the model to locate indentations in approximately the right position. AP_{75}, computed at a stricter threshold (IoU ≥ 0.75), is more sensitive to precise contour alignment and boundary quality. In our case, both models already reach perfect AP_{50}, indicating that almost all indentations are detected without gross localisation errors. The advantage of the optimised model emerges at higher IoU and in the global AP: AP_{75} reaches essentially perfect levels, and the mean AP increases, showing that the optimised network segments indentation contours more accurately rather than merely "finding" them.

Size-specific metrics AP_s, AP_m and AP_l further characterise the behaviour across different indentation sizes, grouping small, medium and large imprints, respectively. Improvements in AP_s and AP_m are particularly relevant here, because smaller and medium-sized indentations are more susceptible to noise, contrast variations and polishing artefacts. The optimised model achieves higher AP_s and AP_m, indicating more robust performance on these more difficult cases, while AP_l also increases, confirming that large, well-defined imprints are segmented with very high fidelity.

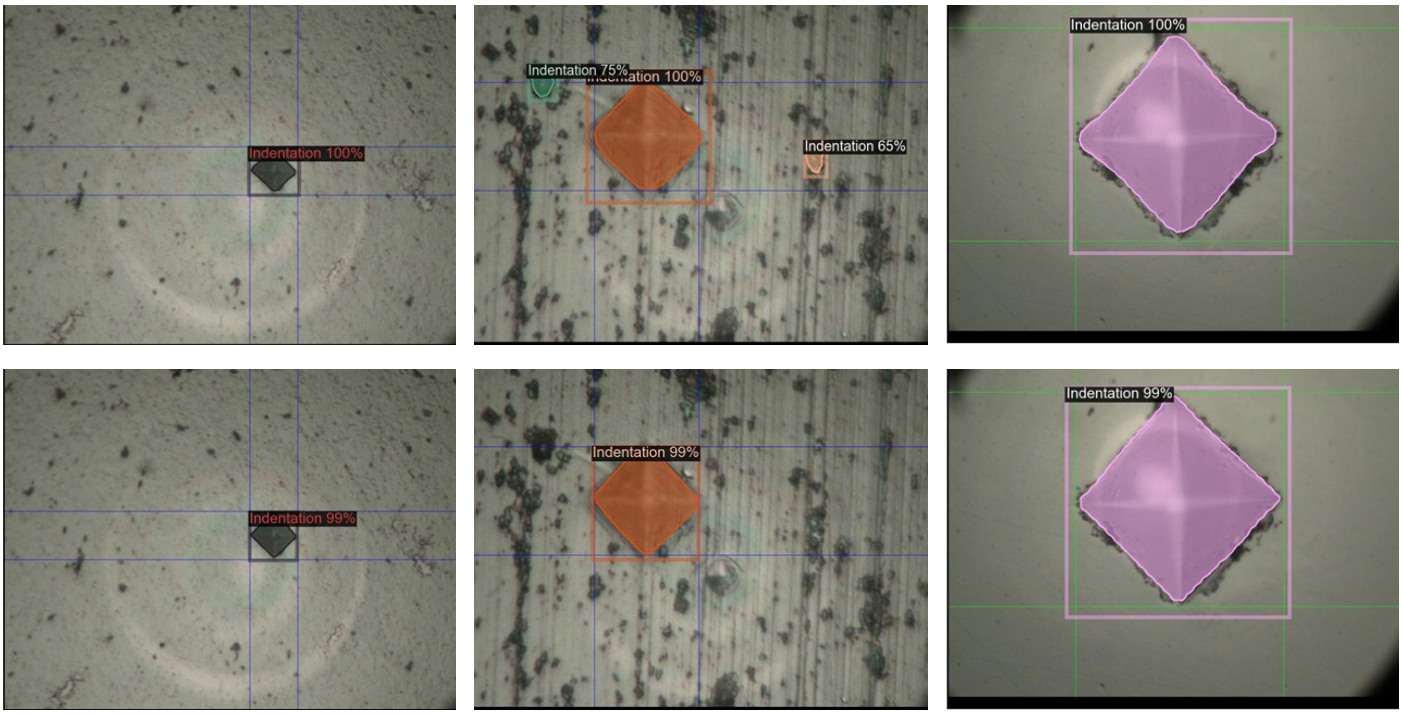

Fig.4 - Qualitative comparison of segmentation results: top row - baseline model; bottom row - optimised model.

Tab.6 - Comparison of segmentation performance between the baseline and optimised models on the test subset

Baseline and Optimal on test subset						
Run	**segmAP [%]**	**AP50 [%]**	**AP75 [%]**	**APs [%]**	**APm [%]**	**APl [%]**
Baseline	87.7±0.4	100	95.6± 0.2	83.3±0.5	83.4±0.2	91.0±0.1
Optimal	90.5±0.5	100	100	90.0	87.7±0.4	95.5±0.5

Diagonal measurements performance

The agreement between automatic and manual diagonal measurements is consistently high and remains stable across the full range of indentation sizes. As shown in the scatter plots (fig. 5), the predicted diagonal lengths follow the ground-truth values almost perfectly, with no evident scale-dependent deviations from the identity line. This linear behaviour is confirmed quantitatively by the determination coefficients ($R^2 > 0.998$ for both d_1 and d_2; tab. 7), demonstrating that the model generalises effectively across different magnifications and indentation sizes rather than regressing toward an average value learned from the training set.

The error distribution relative to practical tolerances is reported in table 8. Approximately 70% of all indentations fall within a strict ±2% relative error, while relaxing the tolerance to ±5%, a range often cited as the inter-operator variability band in manual Vickers testing [4, 32], raises the acceptance rate to about 94% for d_1 and ≈88% for d_2. Beyond ±10%, the method essentially saturates, covering more than 99% of cases. The low median relative error (≈1.2–1.3%), compared with the slightly higher mean relative error (≈1.8–2.0%), indicates a mildly right-skewed distribution: typical predictions are highly accurate, and the mean is affected primarily by a small number of challenging images rather than by systematic model drift.

Tab.7 - Summary of diagonal-measurement accuracy for the best-performing model. Metrics are averaged across repetitions and reported with standard deviations. MAE, median absolute error (MEDAE) and P95 refer to relative errors; R^2 quantifies the agreement between predicted and reference diagonals; Bland–Altman statistics are reported as mean bias and limits of agreement (LOA), expressed in both pixels and relative terms.

Diagonal Measurement Accuracy (Absolute & Relative Errors, Correlation, Bland–Altman)				
Diagonal	**MAE [%]**	**MEDAE [%]**	**P95 [%]**	**R^2**
d_1	1.8±0.2	1.3±0.2	5.1±0.1	0.999 ± 0.001
d_2	2.0±0.1	1.2±0.1	6.6±0.3	0.998 ± 0.001
Diagonal	**BA bias [px]**	**BA bias [%]**	**LOA [px]**	**LOA [%]**
d_1	+1.3 ± 0.5	+0.9 ± 0.3	[-4.23 – 6.67]	[-2.78 – 4.39]
d_2	−0.1 ± 0.5	−0.1 ± 0.3	[-8.64 – 8.16]	[-5.84 – 5.51]

Tab.8 - Fraction of predicted diagonals falling within ±2%, ±5% and ±10% of the manual reference, averaged across.

Threshold within Relative Error			
Diagonal	**±2% [%]**	**±5% [%]**	**±10% [%]**
d_1	69.5±0.1	94.0±0.1	98.7±0.2
d_2	70.8±0.1	87.8 ±0.2	99.1±0.1

The Bland-Altman analysis (fig. 5) clarifies the nature of these deviations. The mean bias is negligible (about +1 pixel for d_1 and approximately 0 pixels for d_2), showing that the segmentation step does not consistently enlarge or shrink the indentation outlines. The limits of agreement remain roughly constant across the full range of diagonal lengths (on the order of ± 6-8 px), meaning that the magnitude of the errors does not systematically increase for larger or smaller indentations. This behaviour indicates that most discrepancies arise from local, pixel-level uncertainty at the indentation edges rather than from any scaling distortion or drift in the geometric fitting. Because this pixel-level uncertainty is essentially constant, its relative impact is naturally higher for very small indentations, which explains the few outliers in the error distribution.

A modest asymmetry is observed between the two diagonals: d_1 consistently shows slightly narrower limits of agreement than d_2. Nevertheless, both diagonals achieve extremely high linear agreement, and the correlation for d_2 remains above $R^2 = 0.998$, fully consistent with the behaviour observed for d_1.

Finally, given that Vickers hardness is inversely proportional to the square of the diagonal, the observed error magnitudes ($MAE(d_1) = 2.3±0.3$ px; $MAE(d_2) = 2.7±0.3$ px) translate into modest and well-bounded variations in the computed HV. In practical terms, the automatic system delivers a measurement repeatability that matches, and in many cases may exceed, the consistency of manual microscopic readings.

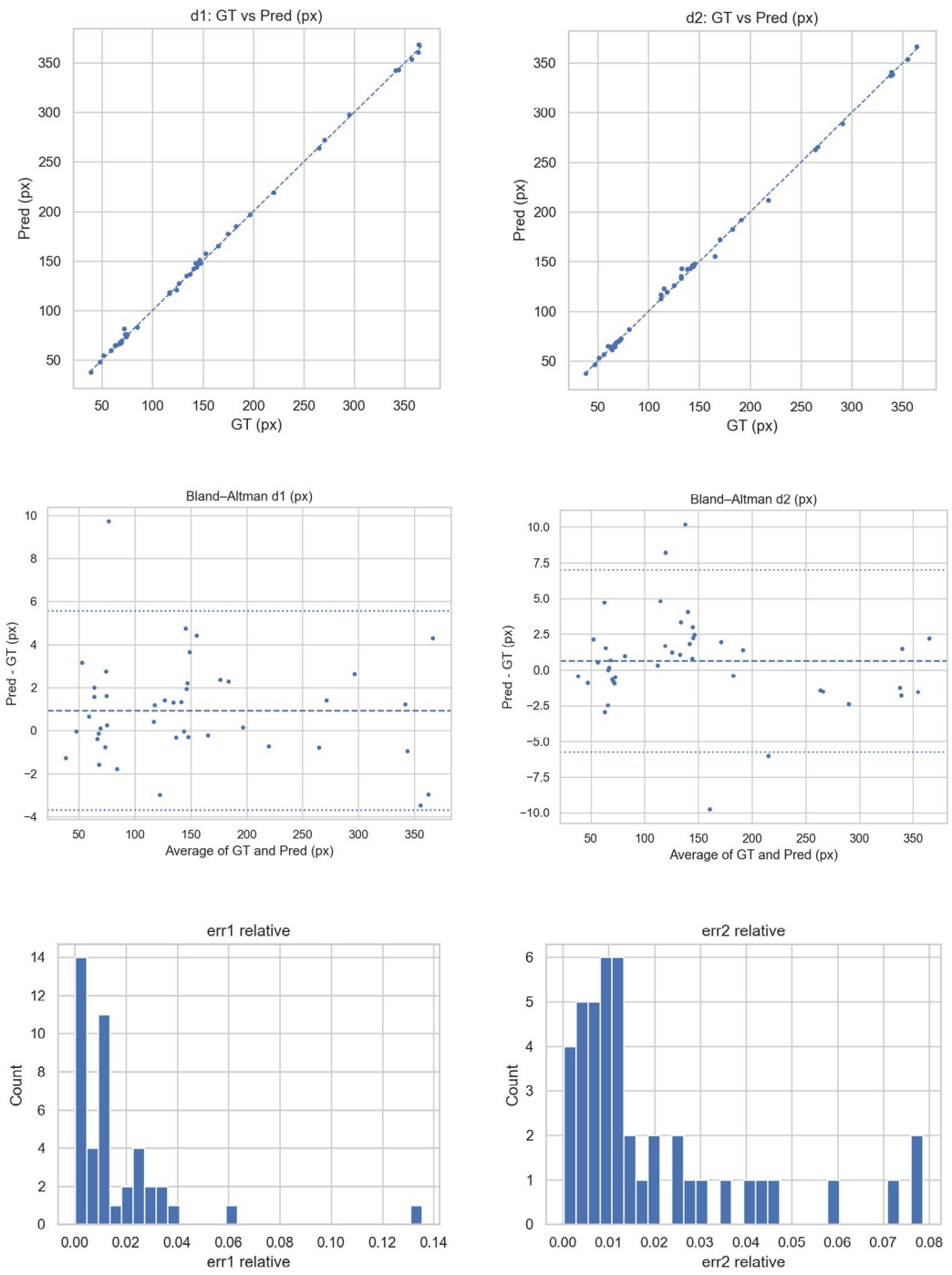

Fig.5 - Metrological validation of the diagonals for a single repetition (seed 65). (Top) Scatter plot comparing the automatic predictions against the manual ground truth measurements. (Middle) Bland–Altman plot displaying the measurement differences against the average of the two methods, indicating the systematic bias (dashed line) and limits of agreement (dotted lines). (Bottom) Histogram showing the frequency distribution of the relative errors.

Implementation example

To demonstrate the model's capability in a realistic metallurgical workflow, the pipeline was tested on an optical micrograph of a MIG soldering pool in S355 steel. The image was acquired from a polished cross-section without etching, and it exhibits marked surface heterogeneity, including polishing scratches and strong phase/reflectivity contrast. In this demonstrative case only,

the pixel-to-micron conversion was obtained manually from the image scale bar and provided as input to compute hardness values in HV units. As illustrated in figure 6, the system successfully detected all four indentations, effectively distinguishing the imprints from background artefacts that typically confound standard thresholding or edge-detection algorithms. The quantitative comparison between the hardness values obtained by automated (HV_AI) and manual (HV_GT) indentation detection is detailed in table 9. The system maintained high metrological accuracy even in this complex landscape, with relative errors ranging from -0.99% (ID 01) to -3.98% (ID 03). The observed Δ% values (≈1–4%) are consistent with the diagonal error distributions reported in the tables 7-8. These results indicate that the segmentation network is sufficiently robust to handle the optical noise and texture variations inherent to routine metallographic inspections of industrial alloys.

Tab.9 - Vickers hardness values calculated by the AI model (HV_AI) compared to the ground truth (HV_GT) manually measured for the indentations in figure 6 and the respective percentage variation.

HV comparison			
ID	**HV_AI**	**HV_GT**	**Δ%**
01	207.5	209.6	-0.99
02	189.7	193.6	-2.03
03	209.9	218.6	-3.98
04	226.1	228.7	-1.16

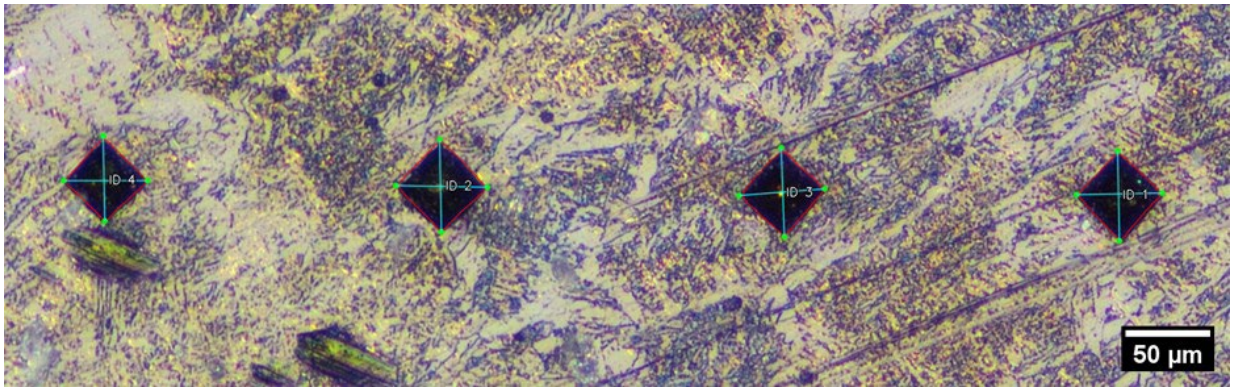

Fig.6 - Example of ML-backed indentations detection and Vickers hardness analysis performed on a multiphasic microstructure with multiple indentations and surface defects.

CONCLUSIONS

This work presented a robust, deep learning-based framework for fully automated Vickers hardness indentation detection and values calculation, effectively addressing the limitations of subjectivity and low throughput inherent to manual testing. By integrating a Mask R-CNN architecture with a rigorous Taguchi-based hyperparameter optimisation, the system achieved precise instance segmentation even under challenging imaging conditions. The experimental analysis showed that training duration and non-maximum suppression thresholds are the critical factors driving segmentation performance, whereas the model proved relatively insensitive to minor variations in learning rate.

Metrological validation against manual ground truth confirmed the system's high accuracy and reliability. The automated diagonal measurements exhibit relative errors consistently confined within narrow industrial tolerances (±5% for most cases). Importantly, the pipeline demonstrates scale invariance and robustness against surface defects, effectively bridging the gap between academic computer vision and practical laboratory requirements.

DATA AND CODE AVAILABILITY

The code developed and the dataset used for this study are not publicly available at this time, as they constitute the basis for several ongoing and planned research works. The dataset used in this study was collected internally and is therefore not publicly released. Reasonable requests for methodological clarification may be addressed by the corresponding authors.

REFERENCES

[1] I. C. Leigh, "Micro-indentation hardness test: The practical realization of a standard," Ph.D. thesis, University of Surrey, 1983.
[2] "Standard Test Methods for Vickers Hardness and Knoop Hardness of Metallic Materials," ASTM International E92 –23, pp. 1–28, 2023, doi: 10.1520/E0092-23.2.
[3] L. Brice, F. Davis, A. Crawshaw, "Uncertainty in hardness measurement," NPL Report (CMAM 87), April 2003. [Online]. Available: https://eprintspublications.npl.co.uk/2615/
[4] G. Barbato, S. Desogus, "Problems in the measurement of Vickers and Brinell indentations," Measurement, vol. 4, no. 4, pp. 137–147, Oct. 1986, doi: 10.1016/0263-2241(86)90006-0.
[5] A. Maier, A. Uhl, "Robust automatic indentation localisation and size approximation for Vickers microindentation hardness indentations," ISPA 2011 - 7th Int. Symp. Image Signal Process. Anal., pp. 295–300, 2011.
[6] T. Sugimoto, T. Kawaguchi, "Development of an automatic Vickers hardness testing system using image processing technology," IEEE Trans. Ind. Electron., vol. 44, no. 5, pp. 696–702, 1997, doi: 10.1109/41.633474.
[7] P. P. R. Filho, et al., "Brinell and Vickers Hardness Measurement Using Image Processing and Analysis Techniques," J. Test. Eval., vol. 38, no. 1, pp. 88–94, Jan. 2010, doi: 10.1520/JTE102220.
[8] A. P. Fedotkin, et al., "Automatic Processing of Microhardness Images Using Computer Vision Methods," Instruments Exp. Tech., vol. 64, no. 3, pp. 357–362, May 2021, doi: 10.1134/S0020441221030180.
[9] J. M. R. Puente, J. M. R. Garnica, C. A. G. Isáis, "Measuring Hardness System Based on Image Processing," 2024 21st International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE), IEEE, Oct. 2024, pp. 1–5. doi: 10.1109/CCE62852.2024.10771036.
[10] Y. Ji, A. Xu, "A New Method for Automatically Measurement of Vickers Hardness Using Thick Line Hough Transform and Least Square Method," 2009 2nd International Congress on Image and Signal Processing, IEEE, Oct. 2009, pp. 1–4. doi: 10.1109/CISP.2009.5305653.
[11] M. Gadermayr, A. Maier, A. Uhl, "Active contours methods with respect to Vickers indentations," Mach. Vis. Appl., vol. 24, no. 6, pp. 1183–1196, Aug. 2013, doi: 10.1007/s00138-012-0478-5.
[12] A. Maier, "Efficient focus assessment for a computer vision-based Vickers hardness measurement system," J. Electron. Imaging, vol. 21, no. 2, p. 021114, May 2012, doi: 10.1117/1.JEI.21.2.021114.
[13] F. D. Lima Moreira et al., "A novel Vickers hardness measurement technique based on Adaptive Balloon Active Contour Method," Expert Syst. Appl., vol. 45, pp. 294–306, Mar. 2016, doi: 10.1016/j.eswa.2015.09.025.
[14] J. D. Polanco, et al., "Automatic Method for Vickers Hardness Estimation by Image Processing," J. Imaging, vol. 9, no. 1, p. 8, Dec. 2022, doi: 10.3390/jimaging9010008.
[15] Y. Tanaka, Y. Seino, K. Hattori, "Vickers hardness measurement by using convolutional neural network," J. Phys. Conf. Ser., vol. 1065, no. 6, p. 062001, Aug. 2018, doi: 10.1088/1742-6596/1065/6/062001.
[16] Y. Tanaka, Y. Seino, K. Hattori, "Measuring Brinell hardness indentation by using a convolutional neural network," Meas. Sci. Technol., vol. 30, no. 6, p. 065012, Jun. 2019, doi: 10.1088/1361-6501/ab150f.
[17] Y. Tanaka, Y. Seino, K. Hattori, "Automated Vickers hardness measurement using convolutional neural networks," Int. J. Adv. Manuf. Technol., vol. 109, no. 5–6, pp. 1345–1355, Jul. 2020, doi: 10.1007/s00170-020-05746-4.
[18] Z. Li, F. Yin, "Automated measurement of Vickers hardness using image segmentation with neural networks," Measurement, vol. 186,, p. 110200, Dec. 2021, doi: 10.1016/j.measurement.2021.110200.
[19] W. S. Cheng, et al.i, "Vickers Hardness Value Test via Multi-Task Learning Convolutional Neural Networks and Image Augmentation," Appl. Sci., vol. 12, no. 21, p. 10820, Oct. 2022, doi: 10.3390/app122110820.

[20] D. G. Privezentsev, A. L. Zhiznyakov, Y. Y. Kulkov, "Automation of Measuring Microhardness of Materials using Metal-Graphic Images," 2019 International Russian Automation Conference (RusAutoCon), IEEE, Sep. 2019, pp. 1–5. doi: 10.1109/RUSAUTOCON.2019.8867750.
[21] E. Jalilian, A. Uhl, "Deep Learning Based Automated Vickers Hardness Measurement," Computer Analysis of Images and Patterns. CAIP 2021. Lecture Notes in Computer Science(), vol 13053. Springer, Cham, 2021, pp. 3–13. doi: 10.1007/978-3-030-89131-2_1.
[22] K. He, G. Gkioxari, P. Dollár, R. Girshick, "Mask R-CNN," 2018. [Online]. Available: https://arxiv.org/abs/1703.06870
[23] W. Sukthomya, J. Tannock, "The optimisation of neural network parameters using Taguchi's design of experiments approach: an application in manufacturing process modelling," Neural Comput. Appl., vol. 14, no. 4, pp. 337–344, Dec. 2005, doi: 10.1007/s00521-005-0470-3.
[24] M. S. Packianather, P. R. Drake, H. Rowlands, "Optimizing the parameters of multilayered feedforward neural networks through Taguchi design of experiments," Qual. Reliab. Eng. Int., vol. 16, no. 6, pp. 461–473, Nov. 2000, doi: 10.1002/1099-1638(200011/12)16:6<461::AID-QRE341>3.0.CO;2-G.
[25] T. Hastie, R. Tibshirani, J. Friedman, "The Elements of Statistical Learning," Springer Series in Statistics. New York, NY: Springer New York, 2009. doi: 10.1007/978-0-387-84858-7.
[26] G. S. Peace, Taguchi Methods: A Hands-On Approach. Addison-Wesley, 1992.
[27] G. Bradski, "The OpenCV Library," Dr. Dobb's J. Softw. Tools, 2000.
[28] G. Borgefors, "Distance transformations in digital images," Comput. Vision, Graph. Image Process., vol. 34, no. 3, pp. 344–371, Jun. 1986, doi: 10.1016/S0734-189X(86)80047-0.
[29] G. J. Grevera, "The 'dead reckoning' signed distance transform," Comput. Vis. Image Underst., vol. 95, no. 3, pp. 317–333, Sep. 2004, doi: 10.1016/j.cviu.2004.05.002.
[30] W. E. Lorensen, H. E. Cline, "Marching cubes: A high resolution 3D surface construction algorithm," ACM SIGGRAPH Comput. Graph., vol. 21, no. 4, pp. 163–169, Aug. 1987, doi: 10.1145/37402.37422.
[31] S. Van Huffel, J. Vandewalle, "The Total Least Squares Problem," Society for Industrial and Applied Mathematics, 1991. doi: 10.1137/1.9781611971002.
[32] G. F. Vander Voort, G. M. Lucas, "Microindentation Hardness Testing," Mechanical Testing and Evaluation, ASM International, 2000, pp. 221–231. doi: 10.31399/asm.hb.v08.a0003272.