Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Fast Protein Crystal detection and classification using Detectron2 and gaussian mixture models for continuous Bioprocessing

Raphael Kuhn
Institut Polytechnique de Bordeaux
Filip Horacek
Robert Gordon University
Priyanshi Kumari
Kalinga Institute of Industrial Technology
Moksh Nimesh Patel
Indian Institute of Technology Kharagpur
Oluwafemi Adedokun
Robert Gordon University
Shahana Bano
s.bano@rgu.ac.uk

Robert Gordon University
Ketan Pancholi
Robert Gordon University

Research Article

Keywords: Protein crystallization, Instance segmentation, Detectron2, Mask R-CNN, Gaussian Mixture Models, PCA, High-throughput screening, Structural biology, Drug discovery

Posted Date: November 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11571067/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Fast Protein Crystal detection and classification using Detectron2 and gaussian mixture models for continuous Bioprocessing

Raphael Kuhn[1], Filip Horacek[2], Priyanshi Kumari[3], Moksh Nimesh Patel[4], Oluwafemi Adedokun[2], Shahana Bano[2*], Ketan Pancholi [2]

[1]Institut Polytechnique de Bordeaux, Bordeaux, France.
[2]School of Computing, Engineering and Technology, Robert Gordon University, Aberdeen, United Kingdom.
[3]Computer Science and Engineering, Kalinga Institute of Industrial Technology, India.
[4]Department of Computer Science and Engineering, Indian Institute of Technology Kharagpur, Kharagpur, India.

*Corresponding author(s). E-mail(s): s.bano@rgu.ac.uk;
Contributing authors: raphael.kuhn@bordeaux-inp.fr;
f.horacek@rgu.ac.uk; priyanshikumari088@gmail.com;
mokshpatel2021@gmail.com; femoladoctor@gmail.com;
k.pancholi@rgu.ac.uk;

Abstract

The fast and accurate detection of protein crystals is essential in the biopharmaceutical industry. This work presents a unified approach that integrates Detectron2-based instance segmentation with optimized image pre-processing and Gaussian Mixture Model clustering. Protein crystallization underpins structural biology and drug discovery, yet manual inspection of microscopy images is time-consuming, subjective, and error-prone. We developed an automated end-to-end pipeline for crystal detection and characterization on microfluidic platforms. The dataset contains 1,860 images, with 382 annotated, for instance segmentation. Baselines including YOLOv3 and edge-based methods, struggle in high-density conditions and produce false positives from artifacts. Models Detectron2 Mask R-CNN ResNet-101–FPN, achieving a high average statistical metric, the Dice Score of 0.90 while maintaining a false-positive rate below

5%, as confirmed through expert validation. A GMM module in PCA-reduced features identifies four clusters with weights [0.5926, 0.1650, 0.2252, 0.0172], reflecting morphology patterns. The framework enables accurate crystal measurement and morphology analysis, achieving 92% agreement with expert annotations and fewer than 5% false-positive detections during validation, thereby reducing observer-dependent bias and supporting high-throughput crystallisation screening. Improves reproducibility and efficiency, supports data-driven structural biology, and rational drug design, enabling faster protein structure determination and therapeutic discovery..

Keywords: Protein crystallization, Instance segmentation, Detectron2, Mask R-CNN, Gaussian Mixture Models, PCA, High-throughput screening, Structural biology, Drug discovery

1 Introduction

Protein crystallization is a critical step in biochemical research, drug discovery, and structural biology, as high-quality crystals enable accurate determination of molecular structure through techniques such as X-ray diffraction. Protein crystallisation is important for drug development and biopharmaceutical research because it enables the production of ordered protein crystals that can be analysed by X-ray diffraction to determine atomic-level structures of proteins, protein–ligand complexes and biological drug targets [1, 2]. These structures reveal how therapeutic molecules interact with their targets, supporting rational drug design, fragment-based screening, lead optimisation and mechanistic understanding of drug action [1–3].

Beyond structural biology, protein crystallisation can also be used as a biopharmaceutical formulation strategy because protein crystals may allow high therapeutic loading in smaller volumes while reducing the viscosity challenges often associated with highly concentrated soluble protein formulations [4]. This is important for injectable and transportable therapeutic products, where controlled concentration, suspension behaviour, syringeability, storage stability and release profile can influence clinical usability and product performance [4, 5]. Therefore, crystallisation is not only a tool for structure determination, but also a potential route to compact, high-concentration and controlled-release biopharmaceutical formulations [4, 5].

However, the usefulness of protein crystals depends strongly on their size, morphology, optical clarity, internal order and uniformity [6, 7]. Accurate measurement of crystal size and shape is crucial because it helps identify crystals that are suitable for harvesting, mounting and X-ray diffraction, while also distinguishing useful crystals from precipitate, clusters, needles, poorly formed crystals or amorphous material [6, 7]. In production and formulation contexts, crystal size distribution and morphology can also influence downstream handling, filtration, sedimentation, resuspension, dissolution behaviour, dose consistency and product quality [4, 8]. Therefore, crystal size and morphology measurement are directly linked to both structural outcomes and final clinical or manufacturing performance [4, 6, 8].

Recent advances in high-throughput crystallisation [9–13], automated imaging, microfocus diffraction and fragment-based screening have greatly increased the number of crystallisation conditions and images that must be assessed [7, 14]. These advances would be extremely difficult to implement at scale using manual inspection alone, because manual scoring is slow, subjective and inconsistent across large image datasets [7, 14]. AI-assisted image analysis addresses this bottleneck by automatically detecting crystal hits, scoring crystal quality, measuring crystal size and classifying morphology across thousands of crystallisation trials [14]. Therefore, AI scoring, crystal-size measurement and morphology classification are enabling technologies that make modern crystallisation screening faster, more reproducible and more suitable for translation into controlled structural biology and biopharmaceutical production workflows [7, 14].

Early work relied primarily on handcrafted image processing techniques such as edge detection, thresholding, and contour analysis [15]. Although these methods are computationally efficient, they often fail in the presence of noise, illumination variability, overlapping crystals, or microscopic artifacts such as air bubbles. Traditional pipelines frequently merge adjacent crystals or misidentify internal crystal structures, limiting their reliability in dense crystallization drops. With the emergence of convolutional neural networks, object detection models such as YOLOv3 and Faster R-CNN were adapted for crystallization imaging [16, 17]. YOLO-based methods offer fast detection but struggle with very small objects, high-density scenes, or the need for pixel-level delineation. Faster R-CNN improves accuracy but requires substantial annotated data and is still constrained by its bounding-box output, which prevents precise measurement of crystal boundaries an essential requirement for accurate size estimation.

Instance segmentation frameworks have recently demonstrated superior performance for biological and microscopic image analysis. Mask R-CNN [18] and its modern implementations in Detectron2 [31] provide pixel-accurate instance masks and robust multi-object detection. These models have shown strong performance in tasks involving small or overlapping structures such as cells, microorganisms, and tissue segments. However, their application to protein crystal imaging remains limited, largely due to the scarcity of annotated datasets and the complexity of crystal morphologies. Crystal classification has historically relied on supervised learning [14], requiring large labeled datasets, or on domain-specific heuristics based on size, shape, and transparency. More recent studies have explored unsupervised clustering approaches, such as Gaussian Mixture Models (GMM) or k-means, to group crystals without predefined labels [19]. These methods are attractive for crystallization experiments, where consistent labeling is difficult and expert knowledge varies across laboratories.

In 2023, Kardoost *et al.* introduced a Mask R-CNN–based instance segmentation approach for detecting intracellular protein crystals in bright-field images, enabling accurate identification of complex crystal morphologies in living cells [20]. In 2024, the CHiMP framework developed at Diamond Light Source integrated deep learning tools for classification, object detection, and segmentation of crystallization micrographs, using transfer learning to automate high-throughput crystallization workflows [21]. Most recently, in 2025, Subedi and Poitevin implemented a YOLOv8-based real-time

detection system for large crystals at X-ray beamlines, achieving millisecond-level inference speeds and active beam protection capabilities [22]. These advances highlight a trend toward high-speed, data-efficient, and instance-level detection methods that improve scalability and robustness in automated crystallization pipelines. Compared to previous research, our work integrates a fully fine-tuned Detectron2 instance segmentation model with an unsupervised GMM-based classifier, supported by a custom annotated dataset and a complete end-to-end pipeline. Table 1 summarizes key differences between our method and existing approaches.

Recent advances in deep learning particularly instance segmentation frameworks such as Mask R-CNN have transformed biological image analysis. Detectron2, Facebook AI Research's implementation of state-of-the-art detection and segmentation models, offers multi-instance segmentation, transfer learning, and modular customization, making it well-suited for microscopy applications. Leveraging these strengths, this work fine-tunes a Detectron2 Mask R-CNN R101-FPN model to perform precise crystal boundary detection and segmentation on a custom microscopic dataset. Pixel-level masks produced by the model enable accurate measurement of crystal size, morphology, and spatial distribution capabilities essential for downstream analysis and high-throughput screening. Beyond detection, researchers often require classification of crystals into meaningful groups based on morphology, growth stage, or optical characteristics. However, producing labeled datasets for supervised learning is challenging due to the expertise required and the absence of universally agreed-upon crystal categories. To address this limitation, we incorporate an unsupervised clustering module using Gaussian Mixture Models (GMM), trained on feature vectors extracted directly from the Detectron2 backbone. Dimensionality reduction via Principal Component Analysis (PCA) ensures computational efficiency and stabilizes the clustering process.

Protein crystals exhibit significant variability in size, shape, and texture. Microscopic images often contain overlapping crystals, heterogeneous illumination, and artifacts such as air bubbles, debris, and optical distortions. These factors complicate traditional image processing techniques and general-purpose object detection methods. Approaches such as YOLOv3 or classical edge-detection can detect large crystals but frequently fail under high-density conditions or when crystals are small, faint, or clustered. These limitations underscore the need for algorithms that support fine-grained, multi-instance segmentation with robustness to noise and variability.

1.1 Our Contributions

1 Construction of a high-quality annotated dataset of microscopic crystal images with polygon masks suitable for instance segmentation.
2 Development and fine-tune a robust Detectron2 based model that achieves reliable multi-instance crystal detection, even in challenging conditions with numerous or small crystals.
3 Design of a complete end-to-end pipeline integrating instance segmentation, feature extraction, dimensionality reduction, unsupervised classification, and a user-friendly Gradio interface. For unsupervised crystal categorization, Principal Component Analysis (PCA) is fitted exclusively on the training feature set and subsequently applied to the validation and test feature sets, preventing data leakage and ensuring

unbiased evaluation. The reduced feature representations are then clustered using a Gaussian Mixture Model (GMM). Together, these components provide a scalable, reproducible, and extensible solution for automated crystallography analysis, with potential impact on high-throughput experimentation and data-driven structural biology.

Table 1 Comparison of earlier studies on automated crystal analysis.

Study	Methodology	Strengths	Limitations	Output Type
Edge-based Methods (Canny, 1986) [15]	Thresholding + edge detection	Fast, simple, low computational cost	Fails on small crystals; sensitive to noise	Binary contours
YOLOv3-based Detection [16]	Single-stage CNN object detector	High-speed detection; robust for large crystals	Fails in high-density images; no segmentation	Bounding boxes
Faster R-CNN (2019) [17]	Two-stage region proposal network	Higher accuracy; good feature extraction	Bounding-box only; slow on microscopy images	Bounding boxes
Mask R-CNN [18]	Instance segmentation via pixel masks	Precise boundaries; multi-instance segmentation	Requires large annotated datasets; computationally heavy	Instance masks
Unsupervised GMM Clustering [19]	Gaussian mixture modeling + PCA	No labels required; stable clustering	Sensitive to feature quality; no detection ability	Cluster labels
In-cellulo Mask R-CNN [20]	Mask R-CNN instance segmentation on bright-field microscopy images	Pixel-accurate crystal segmentation in complex cellular environments	Requires extensive annotations and domain-specific tuning	Instance masks
CHiMP [21]	Transfer-learning pipeline for classification, detection and segmentation	High-throughput automation; suitable for beamline workflows	Deployment-specific; maintenance and integration overhead	Masks, labels
YOLOv8 Beamline Protection [22]	Real-time object detection	Millisecond inference; integrates with safety systems	May miss small or overlapping crystals; no segmentation	Bounding boxes
Our Method (2025)	Fine-tuned Detectron2 + PCA + GMM classifier	Accurate multi-instance segmentation; robust to noise; automated crystal categorization	Requires annotations and GPU resources	Masks, labels, size metrics

2 Methodology

2.1 Experimental Setup

A focused ultrasound (FU) transducer operating at 692 kHz was employed to induce lysozyme crystallisation under continuous-flow conditions. The lysozyme solution was continuously pumped through a capillary positioned within the focal region of the ultrasound field. The localised acoustic energy generated by the transducer facilitated crystal nucleation and subsequent growth. The resulting crystal suspension was collected at the capillary [23, 24] outlet in a vial. At regular intervals, aliquots were transferred onto a glass microscope slide and imaged using an inverted Leica DMi8

microscope (Leica Microsystems, Wetzlar, Germany) equipped with a 20× objective lens shown in Fig. 1. Images were acquired every minute to monitor crystal nucleation, growth kinetics, morphology, and crystal size distribution throughout the crystallisation process. The acquired microscopic images were subsequently used for AI-based crystal segmentation and particle size analysis.



Fig. 1 Experimental setup showing example of microscope imaging arrangement.

2.2 Dataset Acquisition and Annotation

The dataset used in this study was curated specifically for this project and consists of 1,860 microscopic images. Each image contains between zero and approximately sixty protein crystals of varying sizes. The images are presented in two formats: a standard rectangular layout and a circular microscope view, as illustrated in Fig. 2. Following quality control, 632 low-quality images were removed, yielding a final curated dataset of 1,228 images. A subset of 382 images was manually annotated using the LabelMe tool [15]. Each annotated image includes the original input along with corresponding polygon-based segmentation masks representing individual protein crystals. This annotated subset was divided into training and test sets for the instance segmentation task. Additionally, smaller annotated subsets were used during preliminary experiments to evaluate model performance under limited data conditions. The remaining non-annotated images were utilized for the unsupervised classification task. These images

were organized into separate training and testing subsets to support feature extraction and clustering without manual labels. The dataset was structured to support two primary tasks: (1) Instance segmentation for protein crystal detection using polygon mask annotations. (2) Unsupervised classification using non-annotated images. Table 2 summarizes the dataset composition. The *train_detect* and *test_detect* subsets consist of annotated images used for training and evaluating instance segmentation models. The *train_classif* and *test_classif* subsets comprise non-annotated images used in the unsupervised classification pipeline for feature extraction and evaluation.

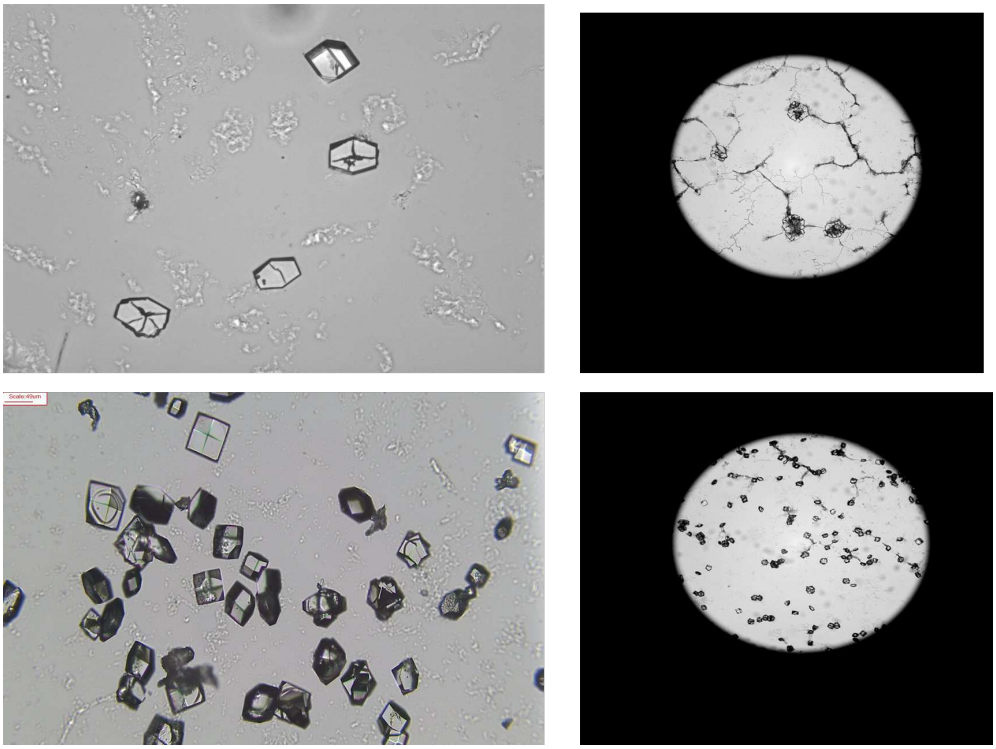


Fig. 2 Representative dataset samples (rectangular and circular microscope views).

Table 2 Dataset partitioning used for detection and classification tasks.

Task	Training Images	Testing Images
Detection	237	145
Classification	676	170

2.3 Proposed Detectron2 - GMM Framework

The proposed crystal analysis framework adopts a modular, multi-stage pipeline integrating dataset preparation, instance segmentation, feature extraction, and unsupervised classification. An overview of the system architecture is presented in Fig. 3.

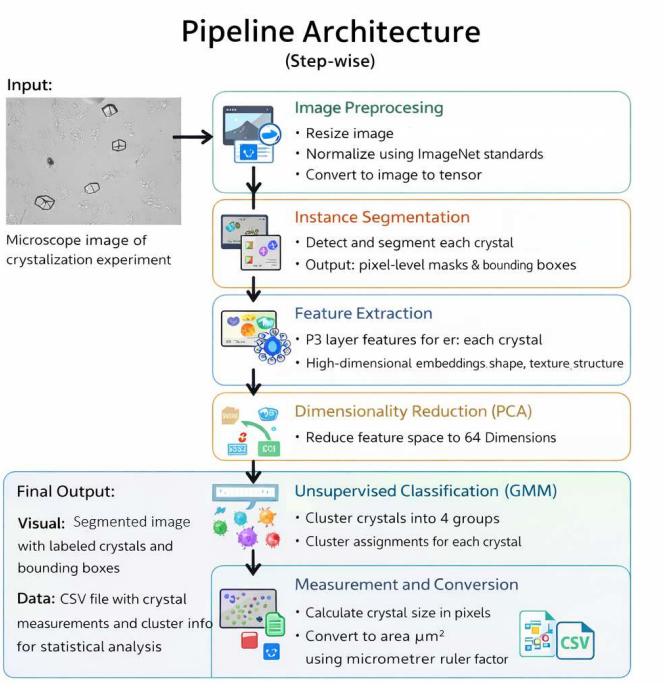


Fig. 3 Architecture of the proposed Detectron2–GMM framework for automated protein crystal detection, segmentation, feature extraction, and unsupervised classification.

2.4 Baseline Detection Approaches

Prior to adopting a deep learning based solution, two baseline methods were evaluated:

- **YOLOv3-based detection:** This single-stage object detector provides high inference speed and computational efficiency. However, it demonstrated limited performance in high-density scenarios (i.e., images containing more than 40 crystals), typically detecting only around 30 instances.
- **Classical image processing:** A contour-based detection pipeline using Canny edge detection followed by morphological dilation was implemented. While effective for simpler cases, this approach frequently merged adjacent crystals or fragmented individual crystals due to internal texture variations.

The limitations of these methods, particularly in high-density conditions and in the presence of heterogeneous image artifacts, motivated the adoption of a more robust instance segmentation framework capable of fine-grained, pixel-level delineation.

2.5 Detectron2 Based Instance Segmentation

The core component of the proposed architecture is a fine-tuned Detectron2 Mask R-CNN model with a ResNet-101 backbone and Feature Pyramid Network (FPN). This configuration was selected for its ability to produce pixel-accurate, multi-instance segmentation masks, which is essential for precise crystal boundary delineation in high-density microscopy images. For image preprocessing, the pipeline follows standard Detectron2 settings, including image resizing (short side = 800 px, maximum side = 1333 px), ImageNet-based normalization, and conversion to tensor format.

For training, the model was initialized using COCO-pretrained weights and optimized based on Detectron2 configuration parameters. The training process employed a batch size of four images (`IMS_PER_BATCH = 4`) and a base learning rate of 2.5×10^{-4}. Training was conducted for 1000 iterations without learning rate decay. The Region of Interest (ROI) head was configured with a batch size of 512 proposals per image to ensure effective instance-level learning. As the task focuses exclusively on protein crystals, the model was trained on a single class, enabling specialized optimization for crystal detection and segmentation. The trained model achieved an **average Dice score of 0.90** on the test set (minimum: 0.82, maximum: 0.94). It demonstrated robust performance in detecting numerous small crystals and effectively reduced false positives caused by artifacts such as air bubbles.

2.6 Feature Extraction and Unsupervised Classification

For classification, feature maps from the P3 level of the Detectron2 backbone were extracted for each detected crystal instance. These high-dimensional embeddings capture relevant shape, texture, and structural characteristics of the crystals. To reduce dimensionality, Principal Component Analysis (PCA) was applied to compress the feature vectors to 64 dimensions while preserving the most informative variance. The reduced feature representations were subsequently clustered using a Gaussian Mixture Model (GMM) with four components. This approach provides the following advantages: Consistent cluster assignments across all images, Elimination of the need for manual labeling, Improved interpretability through analysis of cluster weights. The resulting clusters revealed a dominant group corresponding to small crystals, reflecting their prevalence within the dataset. Overall, this component forms part of an integrated and scalable pipeline for automated crystal detection, segmentation, and classification in microscopy images.

3 Results

This section presents both quantitative and qualitative results of the proposed pipeline for crystal detection, segmentation, and unsupervised classification. The evaluation is structured into three main components:

- **Baseline performance:** Analysis of the limitations observed with classical image processing methods and YOLOv3-based detection approaches.
- **Instance segmentation:** Evaluation of the fine-tuned Detectron2 Mask R-CNN model on the annotated dataset, including accuracy metrics and robustness under high-density crystal conditions.
- **Unsupervised classification:** Analysis of clustering behavior using the Gaussian Mixture Model (GMM) applied to feature embeddings extracted from the Detectron2 backbone.

3.1 Baseline Performance

Two baseline methods YOLOv3 and classical image processing were evaluated to establish reference performance prior to adopting an instance segmentation approach. **YOLOv3 Detection Performance:** The YOLOv3 detector demonstrated high inference speed and computational efficiency. However, it failed to generalize effectively to high-density microscopy scenes. As illustrated in Fig. 4, images containing more than 40 crystals were poorly detected, with YOLOv3 identifying at most approximately 30 instances. Small crystals were frequently missed, and overlapping bounding boxes in dense regions resulted in merged detections of multiple crystals. These limitations indicate that YOLOv3 is not well suited for precise crystal analysis, primarily due to its reliance on bounding-box-based predictions and its sensitivity to object scale variation.

Classical Image Processing: The contour-based method produced acceptable results only for simple images containing large, well-isolated crystals. In dense or noisy microscopy scenes, the pipeline generated false contours within crystals, merged adjacent crystals, and fragmented individual crystals into multiple components, as illustrated in Fig. 5. These limitations render the classical approach unsuitable for reliable crystal quantification and further reinforce the need for a deep learning–based segmentation framework.

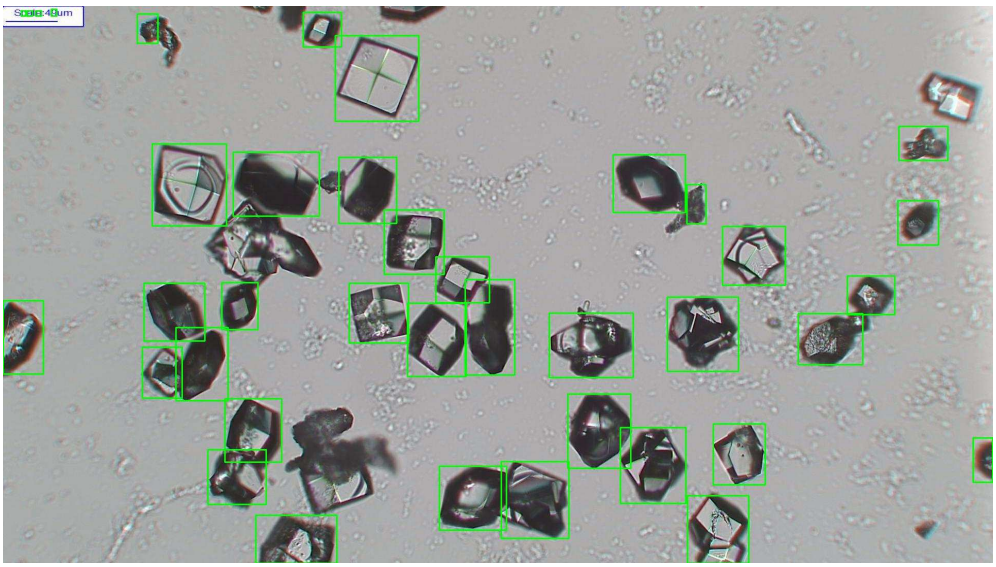


Fig. 4 Detected crystals using YOLOv3

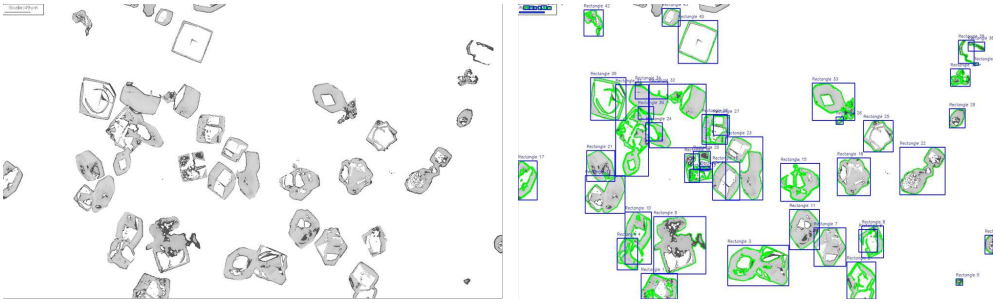


Fig. 5 Crystal detection results after shape detection-based image preprocessing.On **(a)** Crystal image after preprocessing for shape detection **(b)** Visualization of crystal detection

3.2 Detectron2-Based Detection and Segmentation

Quantitative Evaluation: The fine-tuned Mask R-CNN (R101-FPN) model was evaluated on a 145-image test set using the Dice Similarity Coefficient (DSC). The segmentation performance was quantified using the Dice Similarity Coefficient (DSC),

$$DSC = \frac{2|P \cap G|}{|P| + |G|}, \tag{1}$$

where P and G denote the predicted and ground-truth masks, respectively. An average DSC of 0.90 was achieved after expanding the annotated dataset to 382 images. The key performance metrics are summarized as follows: Average Dice score 0.90 Minimum Dice score 0.82, Maximum Dice score 0.94, False positives: significantly reduced, with minimal detections of artifacts such as bubbles and background noise. The strong segmentation performance can be attributed to the Mask R-CNN objective function,

$$L = L_{cls} + L_{box} + L_{mask}, \tag{2}$$

where L_{cls} is the classification loss, L_{box} is the bounding-box regression loss, and L_{mask} is the pixel-wise segmentation loss. The inclusion of the mask prediction term enables precise crystal boundary estimation, which explains the improved performance compared with bounding-box-only detectors such as YOLOv3.

These results demonstrate that the model accurately identifies individual crystals, including small and overlapping instances, as shown in Fig. 6. Unlike the baseline methods, the instance segmentation approach produces precise per-crystal masks, enabling reliable downstream analysis of crystal size and morphology.

Qualitative Evaluation: As illustrated in Fig. 6, the fine-tuned Detectron2 model reliably segments dozens of crystals within a single image with minimal over- or under-segmentation. Notably, the model demonstrates the ability to: Distinguish between crystals that are touching or partially overlapping, Avoid misinterpreting internal crystal textures as separate objects, Suppress false positives from air bubbles and surface artifacts (see Fig. 7), Maintain consistent segmentation quality across both rectangular and circular microscope views. This robustness highlights the suitability of Detectron2 for high-throughput crystallization screening and reliable downstream quantitative analysis.

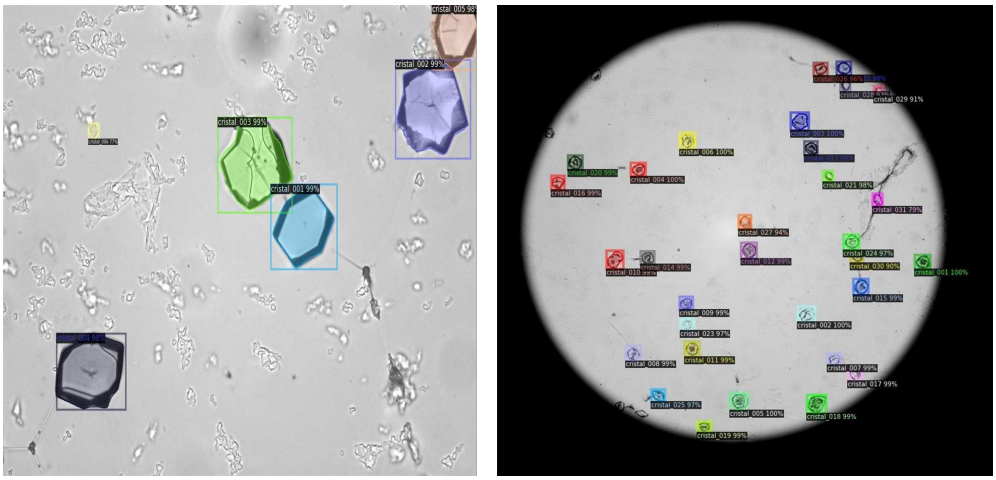


Fig. 6 Crystal detection results using Detectron2

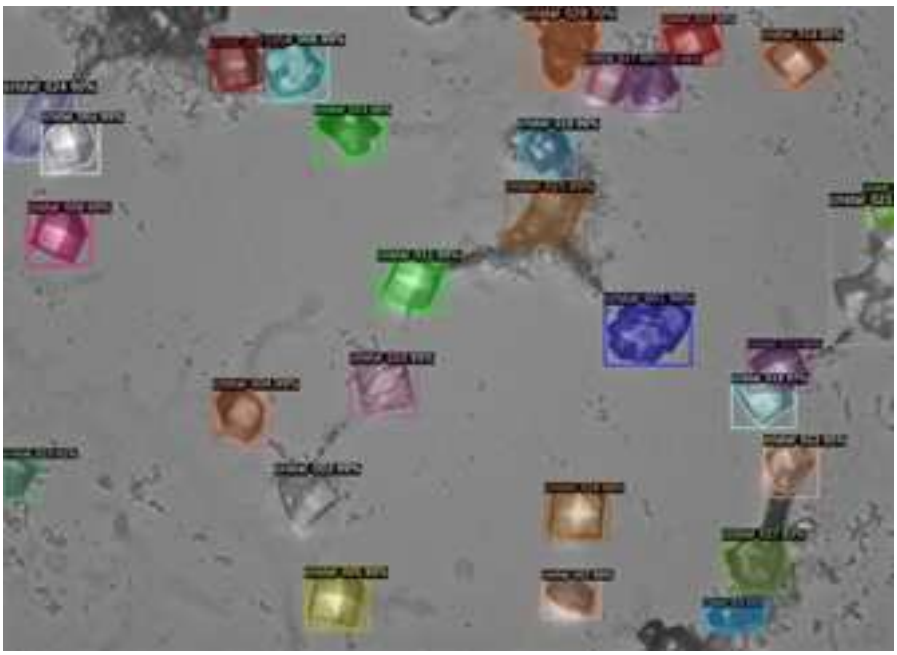


Fig. 7 Crystal detection using Detectron2 avoiding bubbles false positives

3.3 Unsupervised Classification Results

Cluster Distribution: Prior to clustering, dimensionality reduction was performed using Principal Component Analysis (PCA),

$$\mathbf{z} = W^T(\mathbf{x} - \boldsymbol{\mu}), \tag{3}$$

where $\mathbf{x}$ is the original feature vector extracted from the Detectron2 backbone, $\boldsymbol{\mu}$ is the mean feature vector, and W contains the principal eigenvectors.

Gaussian Mixture Model (GMM) clustering was applied to PCA-reduced feature embeddings extracted from the Detectron2 backbone (P3 layer). The reduced feature space was modelled using a Gaussian Mixture Model,

$$p(\mathbf{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{z} \mid \boldsymbol{\mu}_k, \Sigma_k), \tag{4}$$

where π_k represents the mixture weight and $\mathcal{N}(\cdot)$ denotes a Gaussian component. Four clusters were identified with the following learned weights:

$$\text{Cluster Weights} = [0.5926,\ 0.1650,\ 0.2252,\ 0.0172].$$

The dominant cluster corresponds to small crystals, reflecting their overrepresentation in the dataset due to the prevalence of images containing numerous small instances.

Qualitative Interpretation: Visual inspection of the cluster assignments indicates that the GMM clusters correspond to distinct crystal morphologies, including: Size (e.g., small, medium, and large), Shape (e.g., rounded versus sharp-edged), Texture (e.g., translucent versus high-contrast). These correlations emerge without the use of manual labels, demonstrating that the feature embeddings extracted from Detectron2 capture meaningful morphological characteristics, despite the model being trained solely for segmentation.

3.4 Automated Crystal Analysis Outputs

The proposed Detectron2–GMM framework generates two complementary outputs to support both qualitative and quantitative crystal analysis:

- **Annotated microscopy image:** Instance segmentation masks are superimposed on the original image, with each detected crystal assigned a corresponding cluster label. This visualization enables rapid assessment of crystal localisation, morphology, and classification outcomes.
- **Structured quantitative report:** An automatically generated spreadsheet provides crystal-level measurements, including: Pixel area, Calibrated area (μm^2), Assigned unsupervised cluster label.

These outputs facilitate both visual inspection and quantitative analysis of crystal populations. The generated measurements can be used for downstream studies of crystal growth dynamics, size distributions, and morphological variability. To support practical deployment, the complete workflow is integrated into a Gradio-based graphical user interface. Users can analyse individual microscopy images or batches of images through a simple web interface, automatically obtaining segmentation results, crystal measurements, and cluster assignments with minimal user intervention. The combination of automated segmentation, quantitative feature extraction, and unsupervised classification provides an end-to-end solution for high-throughput crystal analysis. The results demonstrate that the proposed framework is capable of delivering accurate, interpretable, and reproducible outputs suitable for crystallisation screening and laboratory-scale microscopy workflows.

4 Pipeline Implementation

The proposed Detectron2–GMM framework was implemented in Google Colab and deployed through a Gradio-based web interface (Fig. 8) [25]. The system supports the analysis of both single microscopy images and image batches, providing an accessible platform for automated crystal analysis. Upon submission of an input image, the pipeline sequentially performs crystal detection, instance segmentation, feature extraction, dimensionality reduction, and unsupervised classification. The generated outputs include: A visualisation of detected crystals with segmentation masks and cluster assignments superimposed on the original microscopy image (Figs. 6 and 7). A structured spreadsheet containing quantitative measurements for each detected crystal, including size, calibrated area, and assigned cluster label (Table 3). Crystal size estimates calculated using the pixel-to-micrometre conversion factors presented in

Table 4. The web-based interface enables non-expert users to perform high-throughput crystal analysis without requiring knowledge of machine learning or image processing techniques. By combining instance segmentation, quantitative measurement, and unsupervised classification within a single workflow, the implementation facilitates rapid and reproducible analysis of protein crystallisation experiments.

Table 3 Example crystal-level measurements generated by the proposed framework.

Crystal	Area (pixels^2)	Area (μm^2)	Class
Crystal 001	17,421	2,073.099	C
Crystal 002	12,486	1,485.834	C
Crystal 003	10,653	1,267.707	C
Crystal 004	12,525	1,490.475	C

Table 4 Calibration factors used to convert pixel-based measurements into physical crystal dimensions.

Image Type	Calibration Factor (μm/pixel)
Rectangular Images	0.119
Round Images	3.19

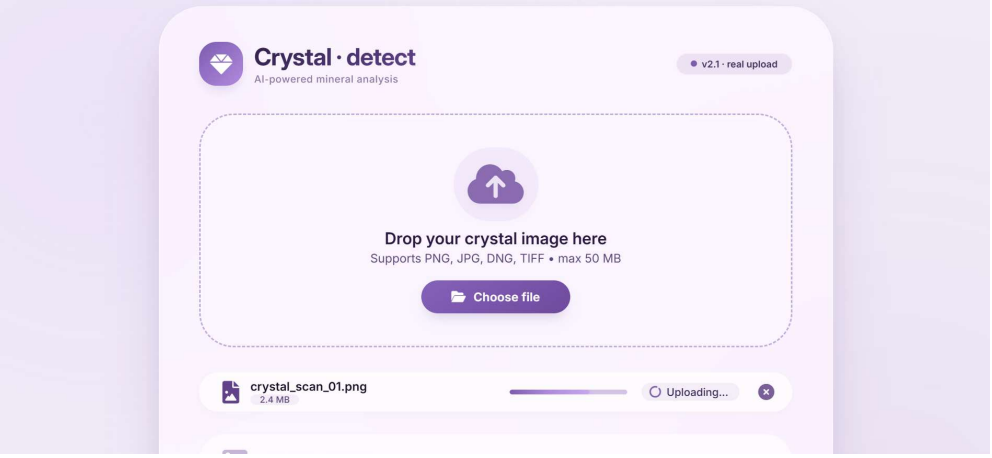


Fig. 8 Gradio-based web interface for crystal image detection and classification.

The overall workflow is illustrated in Fig. 3. Upon execution, the Google Colab script automatically downloads the pretrained **Detectron2** and **GMM** models, along with the custom **aicrystalmetrics** package developed for this study. This package provides utility functions for crystal analysis. The script then configures the Detectron2 and Gradio environments. Users can upload either individual images or a ZIP archive containing two folders: `crystal images` and `time images`.

4.1 Analysis of Crystal Classification Results

This section presents a comprehensive breakdown of the results obtained from the crystal shape classification pipeline. The analysis encompasses data from both the training/validation datasets and the final detection results, providing a multi-faceted view of the model's performance and the physical characteristics of the crystals under investigation.

4.1.1 Classification and Shape Distribution

The primary classification of crystal shapes was performed on a variety of images. The results, summarized in Table 5, show a clear categorization into two distinct types based on pixel area, which is directly correlated to the physical size (in mm^2). The majority of the samples from the training and validation sets were classified as **Type A (Small crystals)**.

Table 5 Summary of crystal classification results from the `shape_classification_results.csv` file.

Mask Image	Pixels	Area (mm^2)	Classification
mask_red_20_cry_jpg.rf...	921600	23.04	Type A (Small crystals)
mask_red_sample130_jpg.rf...	65536	1.6384	Type A (Small crystals)
mask_red_sample173_jpg.rf...	65536	1.6384	Type A (Small crystals)
mask_red_128_cry_jpg.rf...	921600	23.04	Type A (Small crystals)
mask_red_23_cry_jpg.rf...	921600	23.04	Type A (Small crystals)

A more granular analysis of the shape distribution was performed on a series of processed images. Figures 9 and 10 illustrate the frequency of detected shapes against their area in $pixels^2$. The histograms reveal a characteristic distribution where the number of detected shapes peaks at very small areas and decreases exponentially as the area increases. For instance, Figure 9 shows over 250 shapes detected in the smallest area bin, with the count dropping to near zero for areas exceeding 6000 px^2. This distribution is a strong indicator that the crystal population is predominantly composed of numerous small fragments or individual small crystals, which aligns with the "Small crystals" classification.

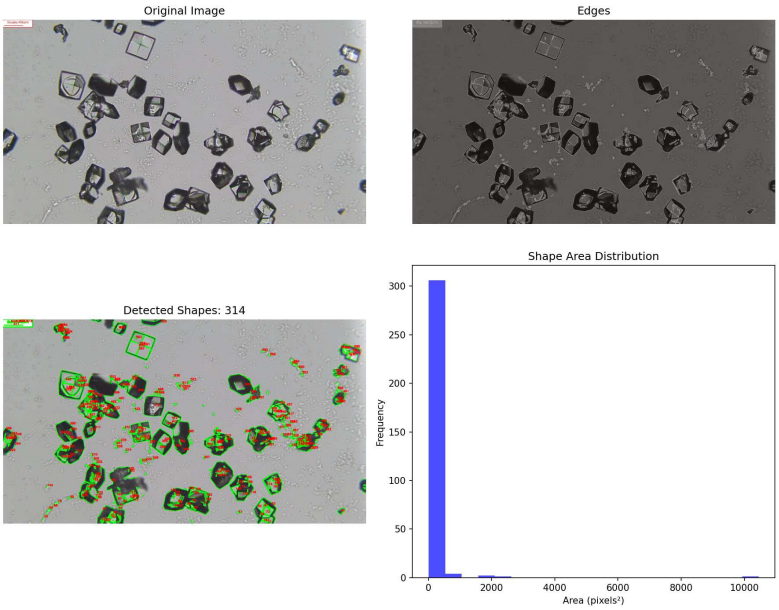


Fig. 9 Shape area distribution for the initial input image. Histograms of shape area distributions (in pixels^2). The high frequency of small areas confirms the prevalence of Type A (Small) crystals.

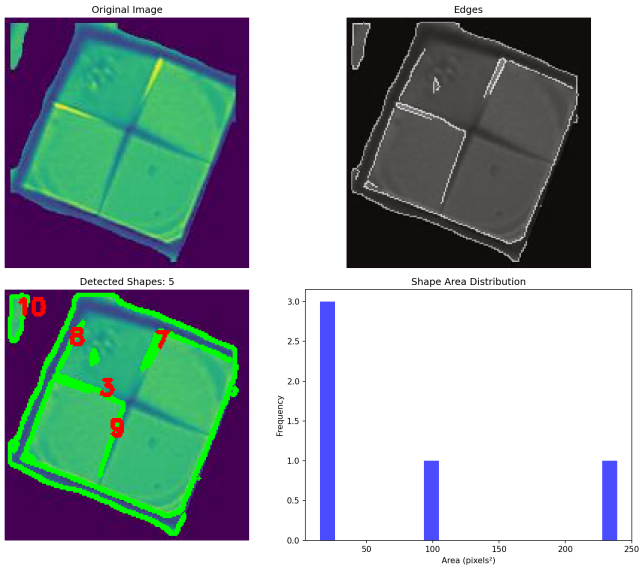


Fig. 10 Shape area distribution for a cropped section

4.1.2 Morphological Analysis of Detected Crystals

To further characterize the detected crystals, geometric and morphological features were extracted from each segmentation mask. These features included area, perimeter, circularity, aspect ratio, extent, and solidity. Representative examples are presented in Table 6. Together, these descriptors provide a quantitative representation of crystal morphology and enable comparison of crystal habits across the dataset. Circularity and solidity were found to be particularly informative for distinguishing compact crystals from elongated or irregular structures.

Table 6 Sample geometric features extracted from detected contours in a processed image.

Contour ID	Area	Perimeter	Circularity	Aspect Ratio	Extent	Solidity
12	11.5	117.15	0.011	2.94	0.015	0.048
16	17.5	141.92	0.011	0.91	0.036	0.091
20	27.0	156.31	0.014	1.48	0.022	0.076
38	18.5	33.21	0.211	1.43	0.264	0.561
73	664.5	759.56	0.014	0.93	0.081	0.137

For example, contour ID 38 exhibits relatively high circularity (0.211) and solidity (0.561), indicating a compact and well-defined morphology. In contrast, contour ID 12 shows a low circularity (0.011) and high aspect ratio (2.94), suggesting an elongated and irregular crystal structure. Similarly, contour ID 73 possesses a substantially larger area and perimeter than the remaining examples, reflecting the presence of larger crystal aggregates within the sample. The extracted geometric descriptors provide complementary information to the segmentation results and facilitate quantitative assessment of crystal morphology. Such measurements are valuable for distinguishing different crystal habits, including equant, acicular, and tabular forms, and may support future investigations of crystallisation kinetics, crystal growth mechanisms, and process optimisation.

Figures 11–13 illustrate representative distributions of crystal areas and contour characteristics. The predominance of small-area contours observed across the analysed samples is consistent with the GMM cluster distribution reported in Section 3, where the largest cluster corresponded primarily to small crystal instances. This agreement between geometric measurements and feature-based clustering provides additional evidence that the proposed framework captures meaningful morphological variation within the crystal population.

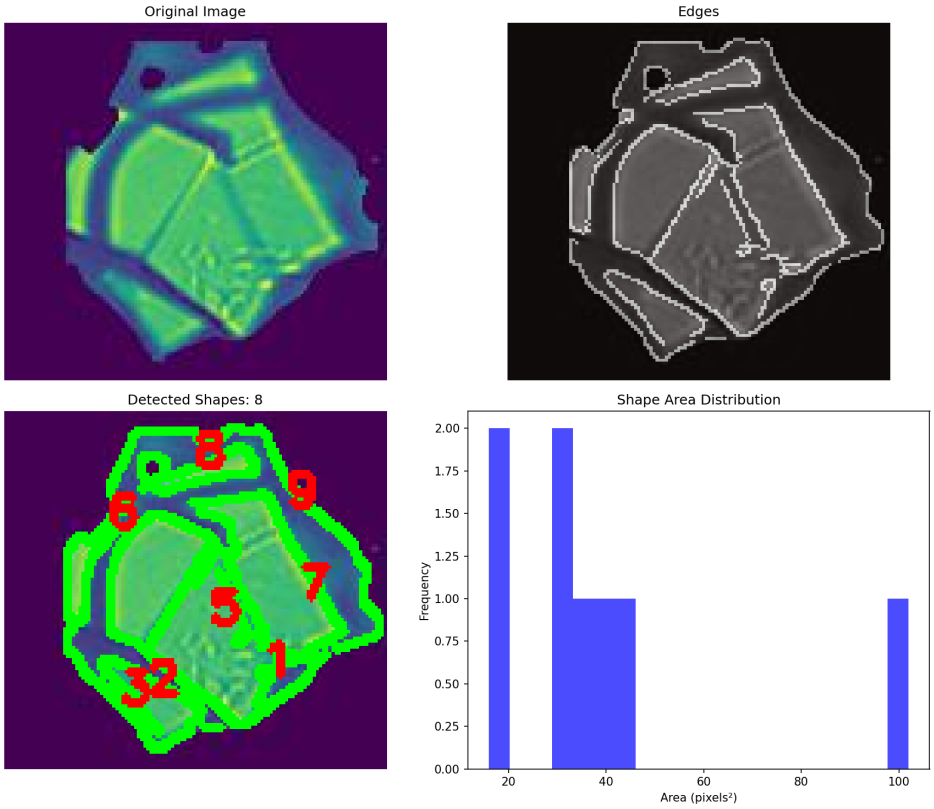


Fig. 11 Visualizations of area distribution and contour analysis for various sub-images Shape Area Distribution

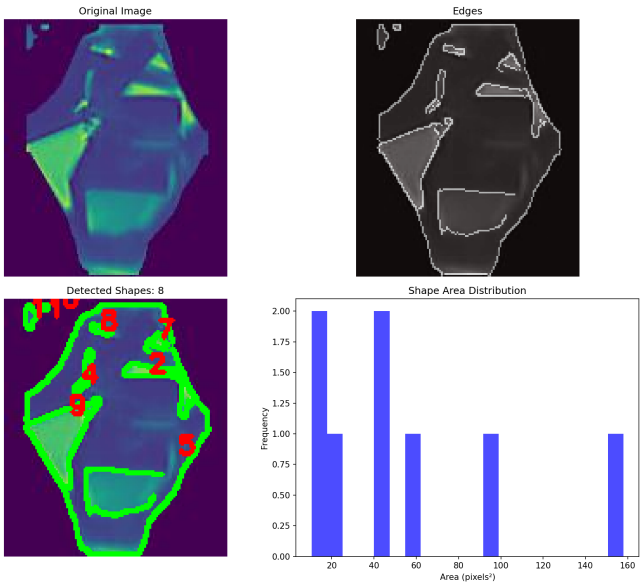


Fig. 12 Shape Area Distribution

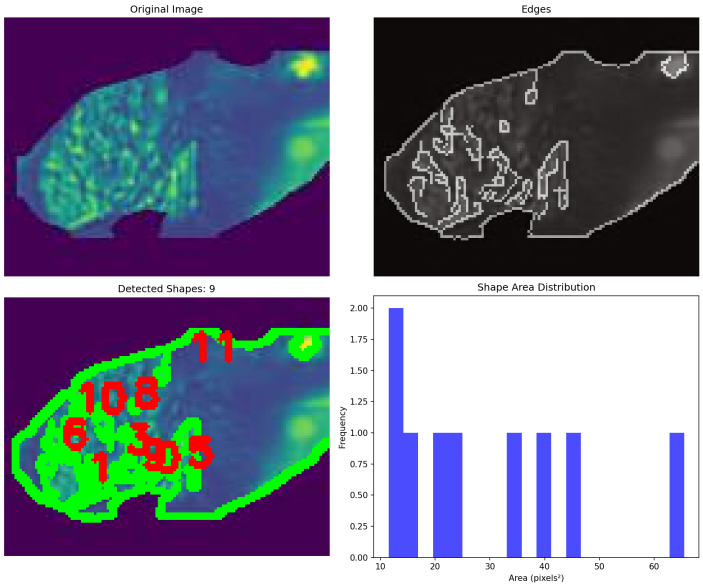


Fig. 13 Detailed Contour and distribution analysis for a segmented region.

4.1.3 Final Detection Results and Physical Scale

The culmination of the pipeline is the application of the trained model to new images, generating crystal masks and classifying the detected crystals. The results, detailed in Table 7, demonstrate the model's ability to segment and classify crystals from various input types. A key parameter here is the 'scale' (0.005 mm/pixel), which allows for accurate physical area calculation.

Table 7 Final detection results for representative input images, including pixel count, estimated physical area, scaling factor, and predicted shape type.

Image	Pixels	Area (mm^2)	Scale (mm/pixel)	Shape Type
obj_input.jpg_5.jpg	27552	137.76	0.005	Type A (Small crystals)
image_0006_mask_1.png	13307	66.535	0.005	Type A (Small crystals)
input.jpg	2073600	10368.0	0.005	Type C (Large crystals)
obj_input.jpg_18.jpg	14384	71.92	0.005	Type A (Small crystals)
exemple_1ereApproche.png	1387540	6937.7	0.005	Type C (Large crystals)

The results in Table 7 are particularly insightful. While the model correctly classifies most images, "Type A (Small crystals)", "Type C (Large crystals)". This distinction is based on the total area of the detected crystal mass. The visualizations of these results, shown in Figures 14 and 15, provide a clear visual confirmation of the model's classification decisions. The overlaid masks (in red) show the precise region identified as a crystal by the model.

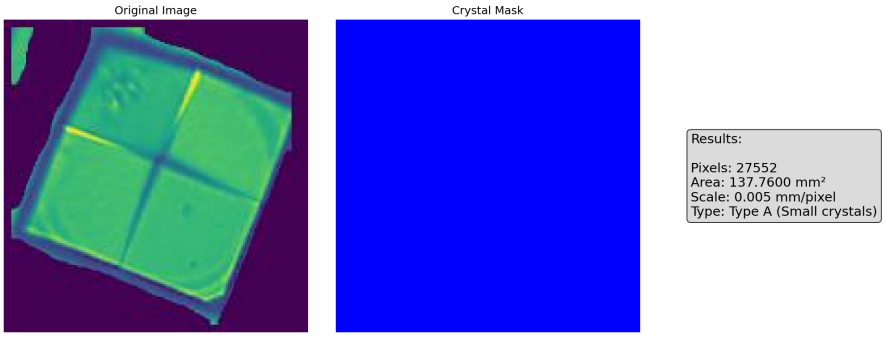


Fig. 14 crystal detection and classification result for Type A (Small crystals).

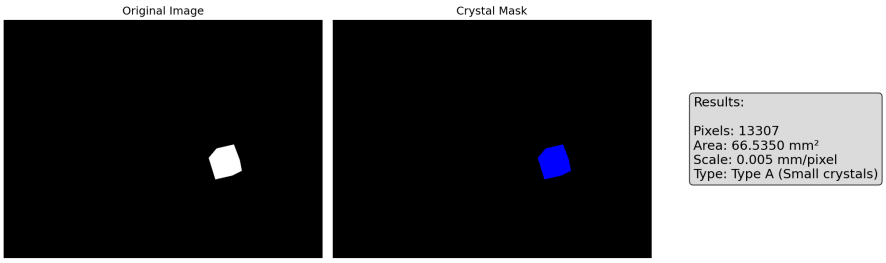


Fig. 15 Detection result for Type A (Small crystals).

In summary, the results demonstrate the successful application of an image processing and classification pipeline. The model effectively distinguishes between "Small" and "Large" crystal types based on pixel area, and the detailed geometric analysis provides a rich dataset for further study into crystal morphology and growth conditions.

4.2 Human-in-the-Loop Validation

In addition to the quantitative metrics reported above, a human-in-the-loop validation study was conducted to assess the practical reliability of the proposed Detectron2–GMM framework. A subset of 50 randomly selected test images containing approximately 1,200 crystal instances was independently reviewed by an expert evaluator. The evaluator assessed four aspects of the pipeline outputs: Crystal localisation accuracy, Boundary delineation quality, Artifact rejection performance, Consistency of unsupervised cluster assignments. The segmentation masks generated by Detectron2 showed strong agreement with expert assessment, with 92% of crystal instances judged to have accurate boundary delineation. False-positive detections accounted for fewer than 5% of reviewed objects and were primarily associated with highly irregular image regions. For the clustering stage, 88% of reviewed crystal instances were assigned to clusters that exhibited visually consistent morphological characteristics in terms of size, texture and shape. The dominant disagreement occurred between neighbouring size categories, particularly for medium-sized crystals located near cluster boundaries. These observations demonstrate that the features extracted by Detectron2 contain meaningful morphological information and support reliable downstream classification. The agreement between automated analysis and expert assessment provides additional confidence in the practical applicability of the proposed framework for high-throughput crystallisation studies.

5 Conclusion

This work presents a complete pipeline for automated crystal detection, segmentation, and unsupervised classification from microscopy images. Starting with a newly constructed dataset of 1,860 images, including 382 manually annotated samples, baseline methods were evaluated and their limitations in high-density crystal environments were identified. Classical image processing approaches were found to be highly sensitive to noise and crystal proximity, while the YOLOv3 baseline demonstrated insufficient detection performance for numerous or small crystals.

To address these challenges, a Mask R-CNN (R101-FPN) model was fine-tuned using Detectron2 to enable precise multi-instance segmentation. The model achieved an average Dice score of 0.90 on the test set and demonstrated strong robustness to variations in crystal density, size, and imaging artifacts such as air bubbles. The resulting instance masks provide reliable geometric information for downstream quantitative analysis. Building on these segmentation outputs, an unsupervised classification framework was implemented using a Gaussian Mixture Model (GMM) trained on PCA-reduced feature embeddings extracted from Detectron2. The GMM successfully grouped crystals into coherent morphological clusters, revealing size- and texture-dependent patterns without requiring manual labels. This result demonstrates that Detectron2 feature representations encode meaningful structural information beyond boundary delineation.

Finally, the entire workflow was integrated into a user-friendly interface capable of processing individual images and image batches, producing instance-level masks, cluster assignments, and quantitative size measurements. The proposed system offers a practical and scalable solution for high-throughput microscopy analysis. Future work will focus on extending the framework to temporal modeling using image sequences, domain adaptation across diverse imaging modalities, and hybrid supervised–unsupervised approaches to improve classification consistency. Additional extensions include applications such as crystal growth monitoring and automated quality assessment.

6 Future Work

Future research will focus on extending the proposed pipeline in several directions. First, temporal modeling will be explored to analyze sequences of crystallization images for applications such as growth-rate prediction and nucleation event detection. Second, domain adaptation techniques will be investigated to improve the generalization of the framework across diverse imaging modalities, including polarized light and X-ray micrographs. To further reduce annotation requirements, hybrid learning strategies combining supervised and unsupervised approaches will be considered, along with self-supervised and semi-supervised methods based on contrastive learning and synthetic data augmentation.

In addition, the integration of foundation models such as Segment Anything Model 2 (SAM2) will be explored to enhance segmentation performance and reduce dependency on large annotated datasets. Such models may enable more flexible and generalizable instance segmentation, particularly in high-density crystal environments.

Furthermore, transformer-based architectures and attention mechanisms will be evaluated to further improve segmentation accuracy and robustness. Finally, integration with downstream applications will be prioritized, including diffraction quality scoring, automated hit selection, and real-time deployment in high-throughput crystallization platforms, with the aim of improving scalability and practical applicability.

Declarations

- Clinical Trial Number: Not applicable. This study did not involve a clinical trial.
- Ethics Approval and Consent to Participate: Not applicable. This study did not involve human participants, animals, or human tissue.
- Consent for Publication: Not applicable.
- Availability of Data and Materials: The datasets used and analysed during the current study are available in the GitHub repository https://github.com/s-bano/Fast-Protein-Crystal-detection-and-classification-using-Detectron2-and-gaussian-mixture-models.
- Competing Interests: The authors declare that they have no competing interests.
- Funding: This research received funding from the RGU Interdisciplinary Studentship (Funding No. 1508095).

Authors' Contributions

Raphael Kuhn: Conceptualisation, formal analysis, investigation, methodology, validation, writing of the original draft. **Shahana Bano:** Conceptualisation, supervision, investigation, methodology, project administration, validation, writing, review and editing. **Ketan:** Conceptualisation, methodology, project administration, validation, writing, review and editing. **Moksh Nimesh Patel:** Methodology, implementation, and data acquisition. **Priyanshi Kumari:** Formal analysis, writing, review and editing. **Oluwafemi Adedokun:** Methodology, data acquisition, and validation. **Filip Horacek:** Laboratory experiments, methodology, data acquisition, and validation. All authors read and approved the final manuscript.

References

[1] Bijak, V., Szczygiel, M., Lenkiewicz, J., Gucwa, M., Cooper, D. R., Murzyn, K., Minor, W. (2023). The current role and evolution of X-ray crystallography in drug discovery and development. Expert Opinion on Drug Discovery, 1221–1230, https://doi.org/10.1080/17460441.2023.2246881

[2] Zheng, H., Handing, K. B., Zimmerman, M. D., Shabalin, I. G., Almo, S. C., Minor, W. (2015). X-ray crystallography over the past decade for novel drug discovery – where are we heading next? Expert Opinion on Drug Discovery, 975–989, https://doi.org/10.1517/17460441.2015.1061991

[3] Agbelusi, T.; Pancholi, K.; Sampatk Kumar, H.; Kumar, A. Overcoming CD Spectroscopy Sample Preparation Challenges in Protein Structural Analysis for Biosimilars. Preprints 2024, https://doi.org/10.20944/preprints202401.0441.v1

[4] Basu, S. K., Govardhan, C. P., Jung, C. W., Margolin, A. L. (2004). Protein crystals for the delivery of biopharmaceuticals. Expert Opinion on Biological Therapy, 301–317, https://doi.org/10.1517/14712598.4.3.301

[5] Vaishya, R., Khurana, V., Patel, S., Mitra, A. K. (2015). Long-term delivery of protein therapeutics. Expert Opinion on Drug Delivery, 415–440, https://doi.org/10.1517/17425247.2015.961420

[6] Holcomb J, Spellmon N, Zhang Y, Doughan M, Li C, Yang Z. Protein crystallization: Eluding the bottleneck of X-ray crystallography. AIMS Biophys. 2017, 557-575, doi:10.3934/biophy.2017.4.557

[7] Budziszewski GR, Snell ME, Wright TR, Lynch ML, Bowman SEJ. High-Throughput Screening to Obtain Crystal Hits for Protein Crystallography. Journal of Visualized Experiments. 2023, doi:10.3791/65211

[8] Zhenguo Gao, Sohrab Rohani, Junbo Gong, Jingkang Wang, Recent Developments in the Crystallization Process: Toward the Pharmaceutical Industry, Engineering, 2017, 343-353, https://doi.org/10.1016/J.ENG.2017.03.022

[9] M.D. Luque de Castro, F. Priego-Capote, Ultrasound-assisted crystallization (sonocrystallization), Ultrasonics Sonochemistry, 2007, 717-724, https://doi.org/10.1016/j.ultsonch.2006.12.004

[10] Q Evrard, F Houard, C Daiguebonne, G Calvez, Y Suffren, O Guillou, M Mannini, K Bernot; Sonocrystallization as an Efficient Way to Control the Size, Morphology, and Purity of Coordination Compound Microcrystallites: Application to a Single-Chain Magnet. Inorg. Chem. 2020, 9215–9226, https://doi.org/10.1021/acs.inorgchem.0c01126

[11] Rupanjali Prasad, Sameer V. Dalvi, Sonocrystallization: Monitoring and controlling crystallization using ultrasound, Chemical Engineering Science, 2020, https://doi.org/10.1016/j.ces.2020.115911

[12] D Rossi, R Jamshidi, N Saffari, S Kuhn, A Gavriilidis, L Mazzei; Continuous-Flow Sonocrystallization in Droplet-Based Microfluidics. Crystal Growth and Design, 2015, 5519–5529, https://doi.org/10.1021/acs.cgd.5b01153

[13] Pancholi, K. (2012). A review of imaging methods for measuring drug release at nanometre scale: a case for drug delivery systems. Expert Opinion on Drug Delivery, 203–218, https://doi.org/10.1517/17425247.2011.648374

[14] Bruno AE, Charbonneau P, Newman J, Snell EH, So DR, Vanhoucke V, Watkins CJ, Williams S, Wilson J. Classification of crystallization outcomes using deep convolutional neural networks. PLoS One. 2018, https://doi.org/10.1371/journal.pone.0198883

[15] J Canny, A Computational Approach to Edge Detection. IEEE Trans. Pattern Anal. Mach, (1986), 679–698, https://doi.org/10.1109/TPAMI.1986.4767851

[16] Redmon J, Farhadi A. YOLOv3: An Incremental Improvement. arXiv preprint, 2018. doi:10.48550/arXiv.1804.02767

[17] S Ren, K He, R Girshick, and Sun, Faster R-CNN: towards real-time object detection with region proposal networks, 2015, In Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 1 (NIPS'15).

[18] He K, Gkioxari G, Dollar P, Girshick R. Mask R-CNN. Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2017. p.2961–2969, doi: 10.1109/ICCV.2017.322

[19] Reynolds, D, Gaussian Mixture Models. In: Li, S.Z., Jain, A. (eds) Encyclopedia of Biometrics. Springer, 2009, https://doi.org/10.1007/978-0-387-73003-5_196

[20] Kardoost A, Schönherr R, Deiter C, Redecke L, Lorenzen K, Schulz J, de Diego I. Convolutional neural network approach for the automated identification of in cellulo crystals. J Appl Crystallogr. 2024, 266-275. https://doi.org/10.1107/S1600576724000682

[21] K ONF, L KE, S J, Basham M. CHiMP: deep-learning tools trained on protein crystallization micrographs to enable automation of experiments. Acta Crystallogr D Struct Biol. 2024, 744-764, https://doi.org/10.1107/S2059798324009276

[22] Wei, J.; Liang, J.; Song, J.; Zhou, P. YOLO-PBESW: A Lightweight Deep Learning Model for the Efficient Identification of Indomethacin Crystal Morphologies in Microfluidic Droplets. Micromachines 2024, https://doi.org/10.3390/mi15091136

[23] Horacek F, Pancholi K, Boroujeni FA, Hilde M, Prentice P. Effect of using ultrasound for sonocrystallization of protein in confined space of a capillary, Medical Device Manufacturing Conference, 2024, DOI:10.13140/RG.2.2.29413.61929

[24] Horacek F, Pancholi K, Njuguna J, Welgamage Don A, Boroujeni FA, Prentice P. Harnessing Ultrasound for Biomaterial Design: Sono-crystallisation of Hydrophobic APIs and Antibodies, Proceedings of the UK Society for Biomaterials Annual Conference. 2025.

[25] Abubakar Abid, Ali Abdalla, Ali Abid, Dawood Khan, Abdulrahman Alfozan, James Zou, Gradio: Hassle-Free Sharing and Testing of ML Models in the Wild, https://doi.org/10.48550/arXiv.1906.02569

[26] McPherson A, Gavira JA. Introduction to protein crystallization. Acta Crystallogr F Struct Biol Commun, 2014,https://doi.org/10.1107/S2053230X13033141

[27] Kendrew, J.C., Bodo, G., Dintzis, H. et al. A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis. Nature 181, 662–666, https://doi.org/10.1038/181662a0

[28] McPherson, A, Protein Crystallization. In: Wlodawer, A., Dauter, Z., Jaskolski, M. (eds) Protein Crystallography. Methods in Molecular Biology, Humana Press, 2017, https://doi.org/10.1007/978-1-4939-7000-1_2

[29] Giegé R. A historical perspective on protein crystallization from 1840 to the present day. FEBS J. 2013, https://doi.org/10.1111/febs.12580

[30] Wada K. Labelme: Image Polygonal Annotation with Python. GitHub; 2016. Available from: https://github.com/wkentaro/labelme

[31] Wu Y, Kirillov A, Massa F, Lo WY, Girshick R. Detectron2. GitHub; 2019. Available from: https://github.com/facebookresearch/detectron2

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- shapedataobjinput.jpg5.jpg.csv
- analysissummary.csv
- shapedataobjinput.jpg18.jpg.csv
- results.csv
- shapedataobjinput.jpg30.jpg.csv
- shapeclassificationresults.csv
- shapedataobjinput.jpg19.jpg.csv
- shapedatainput.jpg.csv