Automated Surface Roughness Prediction from Scanning Electron Microscope Images using Convolutional Neural Network

Md. Ashikur Rahman Khan[1,*], Md. Shariar Kabir Asif[1], Ishtiaq Ahammad[2], M.M Rahman[3]

[1]Department of Information and Communication Engineering, Noakhali Science and Technology University, Noakhali 3814, Bangladesh, [1], ashik@nstu.edu.bd, [1]asif1113@student.nstu.edu.bd
[2]Department of Computer Science and Engineering, Northern University Bangladesh, Dhaka, Bangladesh, [3]ahammadishtiaq27@gmail.com
[3]Faculty of Mechanical and Automotive Engineering Technology, University Malaysia Pahang, Malaysia, [2]mustafizur@ump.edu.my
*Corresponding Author, E-mail: ashik@nstu.edu.bd

Abstract

In recent years, the advent of Convolutional Neural Networks (CNNs) has opened up new avenues for advancing Surface Roughness (SR) prediction methodologies, particularly through the analysis of Scanning Electron Microscope (SEM) images. However, notable gaps existed in the literature regarding the application of CNNs to SEM images for SR prediction. This research addresses these existing gaps by employing CNN to analyze SEM images for SR prediction, particularly focusing on the comparative analysis of different magnification levels. Three distinct datasets, magnified at 150X, 250X, and 500X, were utilized, comprising 2097, 2103, and 2102 images respectively. These images undergo preprocessing techniques to enhance the CNN model's ability to generalize to new images. Subsequently, a sequential CNN model, comprising 27 layers including convolutional, max pooling, batch normalization, flatten, and fully connected dense layers, is developed and trained on the datasets. The study provides detailed comparative analyses of accuracy, precision, recall, and F1-score across magnification levels. Results indicate that the dataset magnified at 500X consistently outperforms the others, exhibiting superior accuracy (75.7%), precision (0.65), recall (0.72), and F1-score (0.72). This suggests that higher magnification levels provide finer details and clearer images, enabling the model to discern subtle features with increased accuracy. Additionally, the 500X dataset exhibits a better balance between minimizing false positives and false negatives, making it more suitable for real-world applications requiring detailed analysis of microscopic structures. These findings underscore the importance of selecting appropriate magnification levels in SEM imaging for accurate SR prediction.

Keywords: Surface Roughness; Surface Roughness Prediction; Scanning Electron Microscope (SEM); Convolutional Neural Network (CNN); Magnified SEM Images

1. Introduction

1.1 Introduction

In contemporary research across various fields, surface roughness (SR) stands as a pivotal parameter, offering invaluable insights into surface topography and texture [1]. The applications of SR measurements span a multitude of scientific and engineering domains, each harnessing its potential to optimize materials, explore biological systems, manufacture microdevices, and elevate product quality. Notably, in the realm of engineering, SR assumes critical significance in tribology, where the study of friction, lubrication, and wear demands meticulous attention [2]. Moreover, materials science and mechanical design are deeply intertwined with SR considerations, as engineers' factor in SR when selecting materials for components that interact with other surfaces, such as bearings and gears. In the manufacturing landscape, SR plays a pivotal role in ensuring the overall quality and functionality of end products [3].

The aerospace industry places paramount importance on SR, as it directly influences aerodynamic performance, durability, and safety in aircraft and spacecraft. SR measurements are instrumental in guaranteeing that surfaces adhere to stringent specifications, thereby optimizing performance and ensuring safety standards are met [4]. Furthermore, the structural integrity of aerospace materials hinges on SR assessments, as minute cracks and defects originating at the surface can culminate in catastrophic failures over time. Hence, SR measurements serve as a vigilant means of monitoring and controlling these imperfections, assuring that components align with requisite specifications and uphold safety standards in aerospace applications [5].

In the domain of nanotechnology, SR assumes a pivotal role in the design and fabrication of microdevices. Here, SR measurements furnish vital insights into the quality and uniformity of nanoscale structures, facilitating the fine-tuning of fabrication processes for utmost precision and control [6]. Beyond this, SR measurements are indispensable in medical devices and implants, where they certify adherence to safety, reliability, and performance standards [7]. Notably, in orthopedics, SR measurements facilitate the exploration of interactions between implants and bone tissue, shedding light on crucial aspects of implant performance [8]. The field of ophthalmology also leverages SR measurements to delve into the interplay between contact lenses and the human eye [9]. Given the profound impact of SR on contact lens comfort and safety, its measurement and control are pivotal for enhancing contact lens performance and safety. Moreover, SR measurements play a significant role in biosensors, where they illuminate the intricacies of sensor interactions with biological analytes [10]. These measurements influence the sensitivity and selectivity of biosensors, underscoring their significance in enhancing biosensor performance. In summary, the scope of measuring surface roughness is vast, and it has applications in various prominent industries.

A Scanning Electron Microscope (SEM) represents a specialized microscopy technique utilized for obtaining high-resolution surface images at the nanoscale. Distinguished from conventional optical microscopes reliant on light for imaging, SEMs employ a focused electron beam to scan the surface of a specimen, ultimately generating an image [11]. The fundamental components integral to the SEM apparatus encompass an electron source, an array of electromagnetic lenses, a sample chamber, and a detector. Typically, the electron source incorporates either a heated filament or a cathode, serving as the emitter of electrons.

These emitted electrons undergo acceleration and precise focusing through a sequence of electromagnetic lenses, directing them towards the specimen of interest. To facilitate optimal electron emission and mitigate charging effects, the specimen is positioned within a vacuum chamber. Furthermore, it is customary to apply a thin layer of conductive material, such as gold or carbon, to the specimen's surface, enhancing electron emission characteristics. This description encapsulates the key operational aspects of an SEM, elucidating its role in achieving high-resolution nanoscale surface imaging [12].

1.2 Problem Statement

As already discussed above, SR measurement holds paramount importance across various industries. Numerous methods have been employed to predict SR, including Taguchi-based regression models [13], [14], statistical regression models [15], computational modeling (e.g., the Finite Element Method and Discrete Element Method) [16], and Machine Learning (ML) methods [17]. However, conventional statistical methods for SR prediction come with a set of limitations. They tend to lack flexibility, are often time-consuming, incur high costs, offer limited accuracy, and exhibit restricted applicability [18]. These methods struggle to account for interactions among process parameters and possess a limited understanding of the underlying mechanisms [19]. Moreover, traditional SR measurement techniques face drawbacks related to their intrusive nature [20]. The requirement for physical contact with the surface being measured can potentially distort the surface, leading to inaccuracies, especially when dealing with delicate or soft materials. Additionally, these methods typically allow for offline measurements but lack the crucial capability for real-time assessment, which is increasingly critical in modern manufacturing settings [21]. Furthermore, these techniques are typically one-dimensional, providing limited insights into the complex three-dimensional surface topography. Consequently, they may fall short in capturing fine-scale variations and subtle nuances in SR. The reliance on skilled operators for data collection and interpretation introduces subjectivity and the potential for human error, impacting measurement consistency. These limitations underscore the growing demand for modern SR measurement techniques that are non-contact, online, and automated, aiming to overcome these challenges and enhance the precision of surface quality assessment in various industrial applications.

The emergence of Artificial Intelligence (AI) approaches, encompassing both Machine Learning (ML) and deep learning (DL), has paved the way for addressing these limitations effectively. AI approaches offer the promise of delivering more precise and adaptable SR predictions. ML methods have demonstrated exceptional capabilities in deciphering intricate patterns and relationships within diverse datasets [22], [23]. Their strength lies in their capacity to capture non-linear interactions among various process parameters, a task that conventional methods often struggle with. By harnessing extensive datasets and sophisticated algorithms, ML models can uncover subtle correlations that might otherwise remain concealed. This not only enhances prediction accuracy but also fosters a profound understanding of the underlying factors influencing SR. DL, a subset of ML, takes predictive modeling to the next level [24]. Neural networks (NNs) within DL architectures automatically extract hierarchical features from raw data, empowering them to capture intricate patterns with remarkable precision. This capability proves invaluable when dealing with complex structures and intricate relationships. The depth and complexity of DL models enable them to unearth nuanced relationships that may elude traditional methodologies. Recently, researchers have placed significant emphasis ML, DL, and Convolutional Neural Network (CNN) for SR prediction. However, our extensive review

of the literature in the "Related Work" section has revealed a notable scarcity of research focusing on the utilization of CNN to analyze SEM images, particularly in the context of predicting SR. Moreover, previous studies in this domain have predominantly relied on single datasets, often overlooking variations in magnification levels of the images. Additionally, there is absence of investigations into the optimal magnification level dataset for SR prediction. Consequently, a lack of guidance exists regarding whether higher or lower magnification levels offer better performance. In light of these gaps, this paper introduces the application of CNN as a novel approach to predict SR based on SEM images. Thereby this research makes contributions to the advancement of the field of SR measurement.

1.3 Novelty of the Research

The research introduces a novel approach to SR prediction by harnessing the capabilities of CNN to analyze SEM images of titanium alloy (Ti-5Al-2.5Sn). The novelty of this study lies in its development of a sequential CNN model specifically tailored to classify SR from SEM images with notable accuracy. To carry out the research, initially we have collected data from referenced papers [25], [26], [27] with proper permission. The collected images vary at different magnification levels ranging from 25X to 4000X. In our study, we focus on three magnification levels of 150X, 250X, and 500X. That means, we utilized three distinct datasets of three magnification levels for SR prediction, encompassing 2097, 2103, and 2102 images respectively. After that, we preprocess the images by resizing, normalizing, and augmenting them to improve ability of our CNN model to generalize to new images. Such preprocessing simplifies the analysis and interpretation of images, ultimately enhancing the accuracy of subsequent analyses. We then build a CNN model which will carry out the SR prediction. Our specified CNN model is composed of 27 layers, encompassing 4 convolutional layers, 4 max pooling layers, 4 batch normalization layer, 1 flatten layer, and 8 fully connected dense layers with an additional 6 dropout layers. These integrated layers collaborate to extract and accentuate diverse features present in the surface texture, including texture patterns and roughness characteristics. The model is then trained utilizing the training dataset. Following the training phase, we assess the model's performance on the validation set. Finally, the test set is employed to evaluate the ultimate performance of the model and analyze any misclassified examples to identify areas for improvement.

The performance evaluation of our CNN model is done using multiple performance metrics to get a comprehensive understanding of the model's strengths and weaknesses. In our study, we utilized three distinct image datasets magnified at 150X, 250X, and 500X levels. For each of the datasets, we provide a comprehensive analysis of the training, validation, and testing results. The training and validation results are presented in terms of training and validation accuracy, as well as training and validation loss. These offers insights into the model's performance during the training phase and its ability to generalize to unseen data. Subsequently, the testing results are presents through confusion matrix, accuracy, precision, recall, and F1-score values. These provides a detailed assessment of the model's predictive capabilities on the test set. Following the presentation of individual dataset results, we conduct a comparative analysis across the three datasets to facilitate a deeper understanding of the performance variations observed. Across the magnification levels, the 500X dataset exhibited the highest accuracy (75.7%), followed by 150X (69.5%) and 250X (61.9%), indicating superior overall classification performance at higher magnification levels. On the other hand, the precision varies, with the 150X dataset achieving the highest

precision (0.71), followed by 500X (0.65), and then 250X (0.60). The 150X dataset demonstrates superior precision, suggesting a better ability to correctly identify positive instances. While the 250X dataset shows the lowest recall (0.60), the 150X and 500X datasets exhibit almost similar recall values (0.69 and 0.72, respectively). The 500X dataset stands out with the highest recall, indicating a better ability to capture all positive values. Among the three datasets, the highest F1-score is observed in the 500X magnified images dataset (0.72), while the lowest F1-score is found in the 250X magnified images dataset (0.60). That means, the 500X dataset demonstrating superior performance in achieving a balance between precision and recall, while the 250X dataset shows relatively weaker performance in this aspect. Based on our results, the 500X magnified images dataset delivers the best results due to its higher accuracy, competitive precision and recall, balanced F1-score, and suitability for real-world applications requiring detailed analysis of microscopic structures. Thereby, the impact of measuring SR from SEM image using CNN is vast, and it has the potential to revolutionize the way we measure SR in various industries and research fields.

The subsequent sections of this article are structured as follows. Section 2 offers an overview of previous studies concerning the prediction of SR, encompassing a discussion on the methodologies employed and the outcomes obtained. In Section 3, we present the methodology adopted in our study, highlighting the data collection, data pre-processing techniques, and the implementation of CNN model for SR prediction. The results derived from our investigation, along with detailed analysis, are presented in Section 4, shedding light on the performance and efficacy of the CNN model developed. Finally, Section 5 serves as the concluding segment of the article, summarizing key findings and delineating avenues for future research and exploration in this domain.

2. Related Works and Research Gap

Surface roughness prediction has been a significant area of research due to its impact on product quality, performance, and aesthetic appeal in manufacturing processes. Recent advancements in predictive modeling, ML, and data-driven approaches have notably enhanced the accuracy and efficiency of SR predictions. Below we studied some of these recent relevant works. After that we presents the research gap of these studies from which we found the motivation to carry out our research.

2.1 Related Works

At first, we discussed the papers which predicts SR without employing specific ML techniques. The review work conducted in reference [28] provides a comprehensive overview of methodologies employed for SR prediction. By categorizing approaches into those grounded in machining theory, experimental investigation, and designed experiments, the paper offers a nuanced understanding of the diverse techniques utilized in this domain. Each approach is meticulously analyzed, delineating its respective advantages and disadvantages, while also exploring current trends and potential future directions. In parallel, reference [29] introduces a comparative study focusing on regression methods for predicting surface roughness across various cutting parameters, such as spindle revolution, cutting speed, feeding speed, depth of cut (DOC), length of cut (LOC), helix angle, and nose radius of cut. This study aims to enhance comprehension of SR evolution through an analytical framework that accounts for multiple influencing factors. The study [30] emphasized Gaussian Process Regression's (GRP's) suitability for predicting SR levels,

based on superior R-squared values compared to other models, marking it as an effective training method. Authors concluded that GRP is the optimal model for datasets of small to medium size with numerous variables, owing to its high prediction accuracy and fit. The work [31] proposed the SPBI method for in-process SR measurement, leveraging the effects of interference and scattering of He–Ne laser light. The technique effectively utilizes bright and dark area counts for SR up to Ra = 3 λ, presenting a simple yet promising approach for automated surface measurement. The research [32] proposes a novel SR prediction model that integrates fused signal features from force and vibration signals generated during milling operations on P20 die steel. By leveraging variational modal decomposition and adopting Deep Belief Networks (DBN) for prediction, this research underscores the importance of feature selection and model optimization in achieving accurate SR estimation. The paper [33] presents an optimized model for SR prediction in shop floor machining operations. Through difference analysis augmented with feedback control mechanisms, this study offers a robust solution capable of adapting to varying machining conditions and generating reliable predictions. The research outcomes underscore the efficacy of the proposed methodologies across a spectrum of experimental scenarios, demonstrating their applicability to both simple and complex datasets with consistent performance.

Now, we discuss the papers which predicts SR by employing various ML approaches. The authors [34] investigated SR prediction of Al6061 using ML models including linear regression (LR), random forest (RF), neural network (NN), and decision tree (DT). The NN emerged as the most effective, achieving low error metrics (MAE: 0.00279, RMSE: 0.00770). In [35], researchers introduced a genetic algorithm-based EHW chip for inspecting SR in milling processes. The technique, which preprocesses images to eliminate noise, showed improved correlation in SR evaluation post-image enhancement with the EHW system, paving the way for future applications using ANN for prediction based on image features. The authors [36] developed a non-contact method for measuring SR, processing surface images of machined materials through ANNs. Despite varying accuracies across different cutting parameters, the study highlighted the potential for enhanced outcomes through further experiments and learning algorithm improvements. In [17], authors demonstrated the predictive power of ANN and classic ML methods (RF, DT, AdaBoost, SVM) for SR in diamond turning of Ge and Cu. The study underscored the efficiency of these methods in overcoming the limitations of traditional models, especially for materials exhibiting brittle behaviour like Ge, thus enhancing predictive accuracy for SR during machining processes. The study [37] explored the prediction of SR values for aluminium alloys machined with Wire Electrical Discharge Machining (WEDM) using ML. The study introduced models including ELM and SVR, highlighting ELM's rapid training times and SVR's iterative risk minimization training. The models demonstrated high accuracy, suggesting their applicability in manufacturing industries utilizing WEDM, with SR values ranging between 2.490 and 3.177. Another similar paper referenced as [38] introduces novel methods for predicting WEDM SR, leveraging both manufacturing parameters established pre-production and machining conditions observed during production. Specifically, the study explores the application of deep neural networks (DNN) and Markov chains integrated with deep neural networks (MC-DNN). Evaluation of prediction errors, conducted in terms of Mean Absolute Percentage Error (MAPE), indicates that the proposed methods exhibit favourable performance compared to existing approaches. Additionally, the utilization of Long Short-Term Memory (LSTM) in SR forecasting, as discussed in [39], demonstrates its adaptability to varying data lengths and ability to capture long-term series features. Meanwhile, the integration of ML with machine

vision, as detailed in [40], presents an innovative approach to predicting machined component roughness. By utilizing a combination of CNC milling processes, SR measurements, and machine vision analysis of surface images, the proposed model showcases promising results in accurately predicting SR values. Furthermore, [41] delves into the prediction of milled SR using regression models, integrating feature extraction techniques such as Grey-Level Co-occurrence Matrix (GLCM) and Discrete Wavelet Transform (DWT). The proposed multiple linear regression model demonstrates promising accuracy in predicting SR values without requiring actual measurements. Additionally, [42] investigates the influence of minimum quantity lubrication on SR in turning operations, employing ML models including LR, RF, and SVM. Notably, RF emerges as the top-performing model, offering accurate predictions of SR under varying cutting conditions. The research [43] focuses on predicting lathe-turned SR across different materials and cooling conditions, employing Gaussian Process Regression (GPR) within MATLAB. Furthermore, [44] explores SR prediction through multiple regression analysis, utilizing cutting parameters, tool wear, and statistical parameters extracted from vibration signals. These efforts underscore the growing significance of ML techniques in enhancing SR prediction accuracy and efficiency in manufacturing processes.

Finally, we focus our study on the research works which employs CNN to predict SR. The study [45] introduces a method combining a CNN classification with electrical discharge-assisted post-processing to enhance the surface quality of metal additive manufacturing (AM) components. The research demonstrates substantial improvements in surface finish by polishing under a low-energy regime, with a notable 74% enhancement in surface quality. The paper [46] proposes a theoretical and deep learning hybrid model for predicting the SR of diamond-turned polycrystalline materials. The model incorporates a kinematic-dynamic roughness component and a material-defect roughness component, the latter of which is predicted using a cascade forward NN. The study successfully enhances prediction accuracy for SR, achieving a remarkably flat surface finish with Sa 1.314 nm. This model provides insights into the influence of grain boundaries and processing parameters on the SR of polycrystalline materials. The research [47] focuses on quality assurance in metal additive manufacturing by employing an area-scan hyperspectral camera coupled with a CNN to predict the SR Rz of magnesium alloy WE43. The study sets up an efficient data acquisition and processing methodology, allowing the network to predict SR within a MAE of 4.1 μm. The research [48] investigates the effects of various machining parameters on the SR of high-strength carbon fiber composite plates during dry end milling. To achieve precise SR estimation, the study introduces a hybrid approach combining an ANN and Genetic Algorithm (GA). The hybrid ANN-GA model demonstrates exceptional performance, with a high prediction correlation ratio (R = 0.96177) indicating strong accuracy. In reference [49], a novel approach combining physical knowledge and deep learning is introduced, termed Physics-Informed Deep Learning (PIDL), aimed at predicting milling SR while adhering to physical laws. During training, a physically guided loss function is formulated to steer the model training process with the aid of physical principles. Leveraging the superior feature extraction capabilities of CNNS and Gated Recurrent Units (GRUs) across spatial and temporal scales, a CNN–GRU model is adopted as the primary model for SR predictions. Experimental validation conducted on the open-source datasets S45C and GAMHE 5.0 demonstrates that the proposed model achieves the highest prediction accuracy compared to state-of-the-art methods, reducing the mean absolute percentage error on the test set by an average of 3.029% compared to the best comparison method. In contrast, reference [50] presents the construction of an "optical image-surface roughness" dataset and designs SSEResNet, a regression prediction model

for SR utilizing a feature fusion approach. SSEResNet exhibits effective extraction of detailed features from optical images, with optimization facilitated through the Adam method. Experimental evaluation, comparing SSEResNet with seven other CNN backbone networks, highlights its superior performance. These findings collectively contribute to advancing the field of SR prediction and underscore the significance of integrating physical principles with CNN techniques for enhanced accuracy and reliability. In research [51], LineNet1 and LineNet2, two deep CNNs, are introduced to address the challenges of denoising and edge image prediction in low-dose SEM images. To train LineNet1 and LineNet2, supervised learning datasets comprising single-line and multiple-line SEM images, along with edge positions information, are used. The results of this approach demonstrate improved edge estimation in multiple-line images and a reduction in memory requirements for single-line images, all while maintaining accuracy in edge prediction. In the study [52], authors explored the application of CNNs for classifying SEM images of pharmaceutical raw material powders to assess their particle morphology. Authors investigated ten pharmaceutical excipients, each exhibiting distinct particle morphologies. The findings, obtained through 5-fold cross-validation, revealed that the CNN models achieved high classification accuracy using either pretrained model, effectively discerning the type of excipient with accuracy. These results indicate that CNN models possess the capability to detect nuances in particle morphology, including differences in particle size, shape, and surface condition.

2.2 Research Gap and our Study's Solution

The existing literature highlights a scarcity of research focusing on the application of CNN to analyze SEM images, particularly in the context of predicting surface roughness. Addressing this gap, our study aims to explore surface roughness prediction using SEM images and CNNs. Additionally, prior studies in this field have predominantly utilized single datasets without considering variations in magnification levels. To address this limitation, we have employed three distinct datasets for SR prediction: a 150X magnification set comprising 2097 images, a 250X magnification set containing 2103 images, and a 500X magnification set comprising 2102 images. Furthermore, the literature lacks investigations into the optimal magnification level dataset for SR prediction. Consequently, there is a lack of guidance regarding whether higher or lower magnification levels yield better performance. To address this gap, we conducted a comparative study among the three magnification level datasets (i.e., 150X, 250X, and 500X). Our findings indicate that higher magnification levels, (i.e., 500X in our research), offer finer details and clearer images, enabling the model to discern subtle features with increased accuracy. This leads to more precise classification and improved overall performance.

3. Methodology

This section provides a comprehensive overview of the methodologies adopted throughout the study. It describes important elements such as dataset details, terminologies, data collection, data processing methods, model building, training, testing, and the evaluation parameters of the model performance. At first, we preprocess the images by resizing, normalizing, and augmenting them to improve ability of the model to generalize to new examples. We then build a CNN model with multiple convolutional layers, pooling layers, and fully connected layers. The model is then trained utilizing the training dataset.

Following the training phase, we assess the model's performance on the validation set. Finally, the test set is employed to evaluate the ultimate performance of the model and analyze any misclassified examples to identify areas for improvement.

3.1 Workflow of the Methodology

This research is divided into five main parts: collecting data, preparing data, creating a CNN model, training and testing the model, and evaluating its performance. We follow each part step by step, breaking down the main parts into smaller steps. The basic steps for classifying surface roughness are shown in Figure 1, giving a clear picture of how we approached the study.

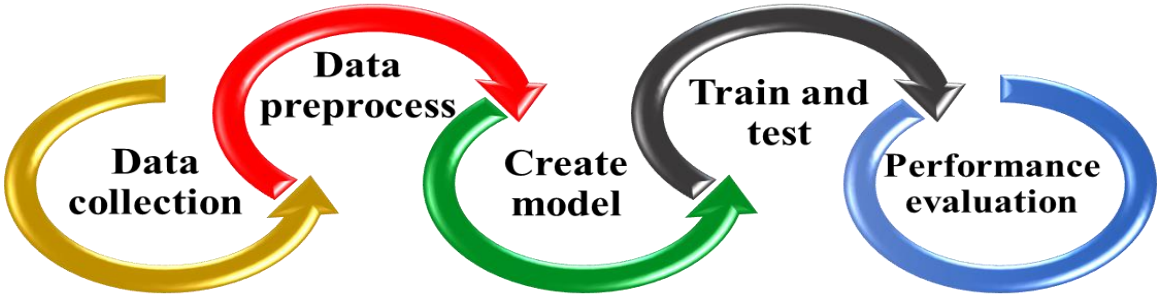

Figure 1: Workflow diagram of the methodology

3.2 Dataset Generation

The initial step in the CNN workflow is acquiring image data, and its quality significantly influences the performance of the trained models. In this study, surface characteristic images of a titanium alloy material (Ti-5Al-2.5Sn) were generated using SEM. These specific images were sourced from the referenced papers [25], [26], [27], with due permission from relevant authorities. Figure 2 outlines the procedural steps involved in generating raw image data along with corresponding surface roughness values.

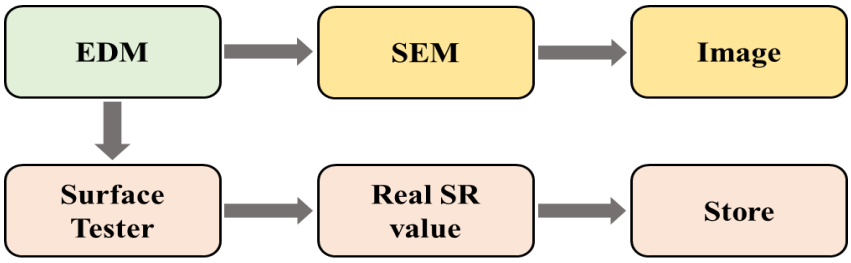

Figure 2: Raw image generation procedure

3.3 Surface Characteristic Image

Utilizing both positive and negative polarities, along with varying peak currents and pulse on times for three electrodes (Cu, CuW, Gr), we obtained multiple surface roughness (SR) values for the Ti-5Al-2.5Sn alloy. This was achieved through 18 sets of images, each

classified by different parameters applied during image generation. The specific machining parameters employed in generating all raw images are presented in Table 1.

Table 1: Machining parameters and their corresponding SR values

SL No.	Discharge (Peak current mA × Pulse on time µs)	Electric Polarity	Electrode	Surface Roughness (SR) value
229	2×95	Positive	Cu	0.9862
230	15×180	Positive	Cu	2.7543
241	29×320	Positive	Cu	5.9965
239	2×95	Negative	Cu	1.0593
231	15×180	Negative	Cu	4.728
238	29×320	Negative	Cu	6.7334
232	2×95	Positive	CuW	0.8315
233	15×180	Positive	CuW	3.0652
234	29×320	Positive	CuW	6.3067
235	2×95	Negative	CuW	1.069
240	15×180	Negative	CuW	4.3732
236	29×320	Negative	CuW	6.7838
244	2×95	Positive	Gr	2.0119
242	15×180	Positive	Gr	4.553
237	29×320	Positive	Gr	6.1157
243	2×95	Negative	Gr	3.4227
245	15×180	Negative	Gr	9.9842
246	29×320	Negative	Gr	15.6117

Each set comprises 10 images taken at different magnification levels ranging from 25X to 4000X. For this study, we focused on three specific magnification ranges: 150X, 250X, and 500X. This resulted in a total of 16 sets of images, each containing three magnified versions, thus yielding 16 X 3 images as our original dataset. The raw images collected

during this phase are stored for subsequent operations. A representative excerpt of an image within the 150X magnification range is illustrated in Figure 3.

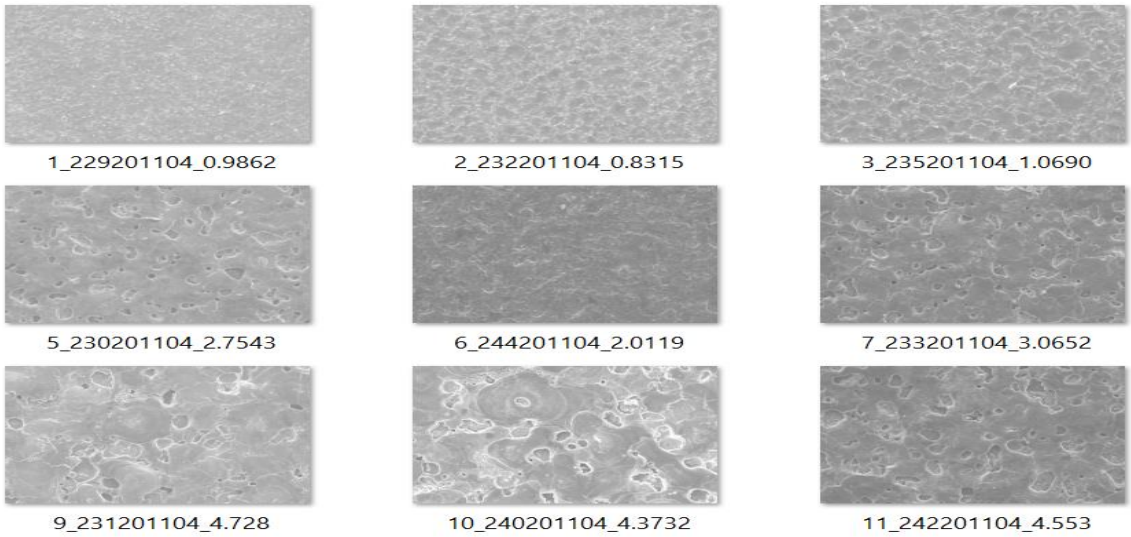

Figure 3: First nine collected images of 150X magnification

3.4 Raw Image Pre-processing

Image data preprocessing involves employing a set of techniques to ready digital images for analysis or other processing. This typically includes removing irrelevant information to ensure accurate input into the CNN model. Such preprocessing simplifies the analysis and interpretation of images, ultimately enhancing the accuracy of subsequent analyses. The process encompasses various steps, such as cropping images, augmenting them, storing the augmented images, and ultimately splitting all the images. Figure 4 visually represents the sequential steps involved in our image data preprocessing.

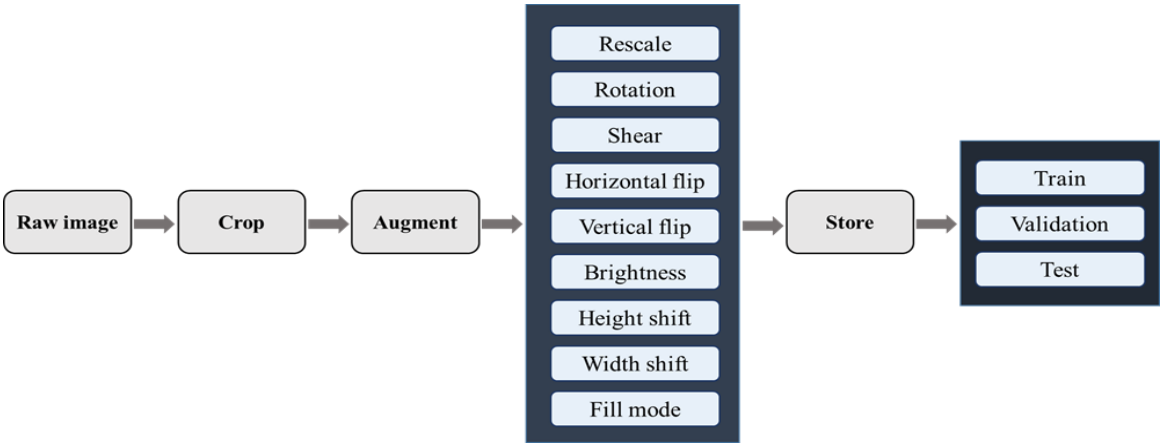

Figure 4: Image preprocessing steps

Cropping

Cropping is a process of eliminating undesired portions of an image to concentrate on a particular area of interest. This technique is commonly applied to eliminate backgrounds or zoom in on specific objects. In this context, our dataset images undergo cropping, resulting in a dimension change from 724×1024 to 690×1024. This eliminates unnecessary sections from the image (including SEM parameters), aiding the CNN model in extracting the most crucial features. The cropping operation for our dataset is visually depicted in Figure 5.

Augmentation

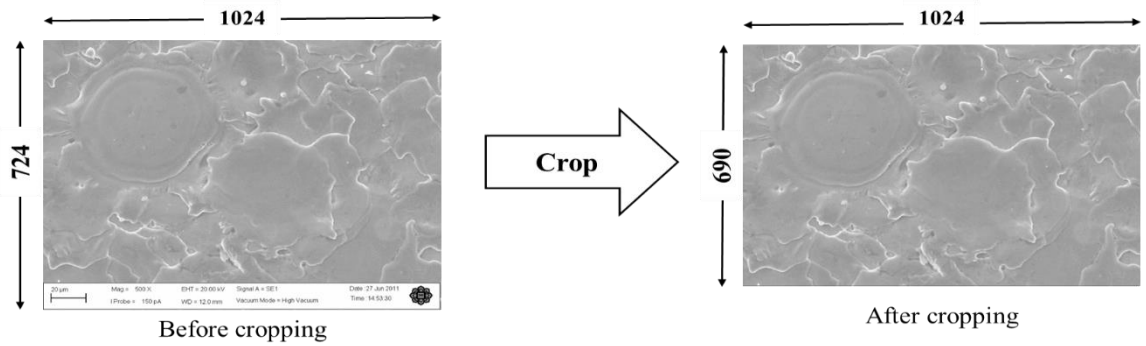

Figure 5: Cropping process of an image

Data augmentation involves creating additional images by applying various transformations, such as rotation, flipping, or shearing, to the original images. This process aims to expand the dataset and enhance the model's robustness. In our dataset, we augment images separately at three magnification levels: 150X, 250X, and 500X. This results in a total of 6,302 augmented images derived from the initial 48 images across the three magnifications.

To execute data augmentation, we employ the *Image Data Generator* class from Keras. This class facilitates real-time augmentation during the training of our CNN model by generating batches of image data. The applied properties include:

- Rotation Range: Images are randomly rotated within a range of 180 degrees.
- Width Shift Range and Height Shift Range: Random horizontal and vertical translation of images within a fraction (0.2) of the total width or height. In our model, width_shift_range is set to 0.2, and height_shift_range is set to 0.2.
- Shear Range: Random application of shearing transformations within a range of 0.2.
- Zoom Range: Random application of zooming transformations within a range of 0.1.
- Horizontal Flip and Vertical Flip: Boolean values indicating whether to randomly flip the images horizontally or vertically. Both are set to "True."
- Fill Mode: To fill the pixel value in augmented images after shifting and flipping, we set fill_mode to 'reflect' for better results.

- Brightness Range: Random adjustment of brightness within a specified range (0.3, 1.4).
- Rescale: A scaling factor applied to the pixel values of the images. In our model, rescale is set to 1/255.
- Data Format: The format of the input data, specified as 'channels_last' in our model.

A portion of augmented images of our dataset are shown in Figure 6.

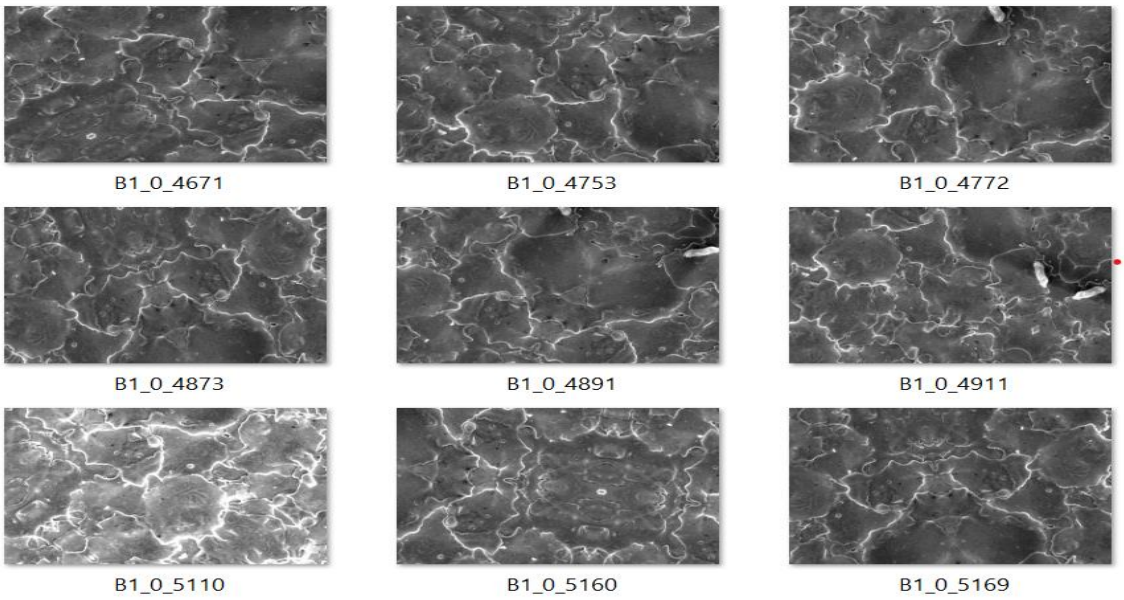

Figure 6: A portion of augmented images

Data Storage and Dataset Frequency

The image datasets at 150X, 250X, and 500X magnifications are divided into training, validation, and test sets, each comprising seven classes labeled A, B, C, D, E, F, and G. Class A contains images with SR values ranging from 0 to 0.9, while Classes B to G encompass images with SR values in the corresponding range from 1 to 6.9. These classes are defined as keys in a dictionary. Table 2 shows the class corresponding to the SR range with dictionary values**.**

Table 2: Classes corresponding to SR value

Class	Values	Surface roughness value (µm)
A	0	0-0.9
B	1	1-1.9
C	2	2-2.9
D	3	3-3.9
E	4	4-4.9

F	5	5-5.9
G	6	6-6.9

All images, totaling 6,302 augmented images across the three magnification sets, are categorized into seven classes. The distribution includes 2,097 images in the 150X magnification set, 2,103 images in the 250X magnification set, and 2,102 images in the 500X magnification set. Figure 7 visually represents the data distribution among the three magnification sets. The dataset is further utilized in three subsets: the training dataset trains the model, the validation dataset fine-tunes the model's hyperparameters, and the test dataset evaluates the model's performance on unseen data. This division ensures a comprehensive assessment of the model's capabilities across different datasets.

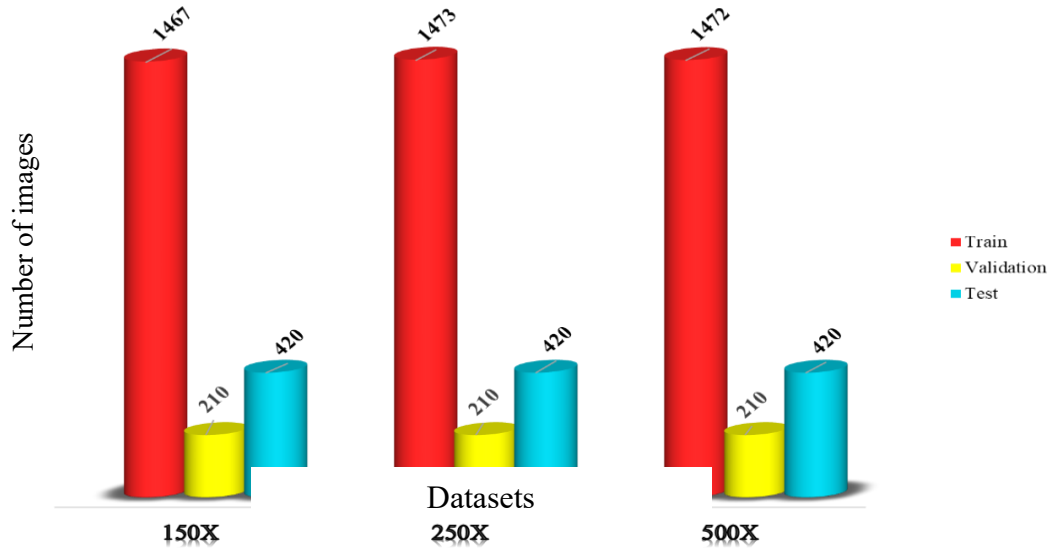

Figure 7: Graphical representation of dataset frequency

3.5 Convolutional Neural Network for Surface Roughness Prediction

The sequential CNN model for SR classification from image is basically a deep learning model that is designed to classify different levels of SR from images. Our specified CNN model for SR is composed of 27 layers, encompassing 4 convolutional layers, 4 max pooling layers, 4 batch normalization layer, 1 flatten layer, and 8 fully connected dense layers with an additional 6 dropout layers. These integrated layers collaborate to extract and accentuate diverse features present in the surface texture, including texture patterns and roughness characteristics. Figure 8 visually illustrates the architecture of the CNN model, providing a comprehensive depiction of all layers involved in feature extraction from images and subsequent output generation. All the layers of the CNN model are briefly described in bellow.

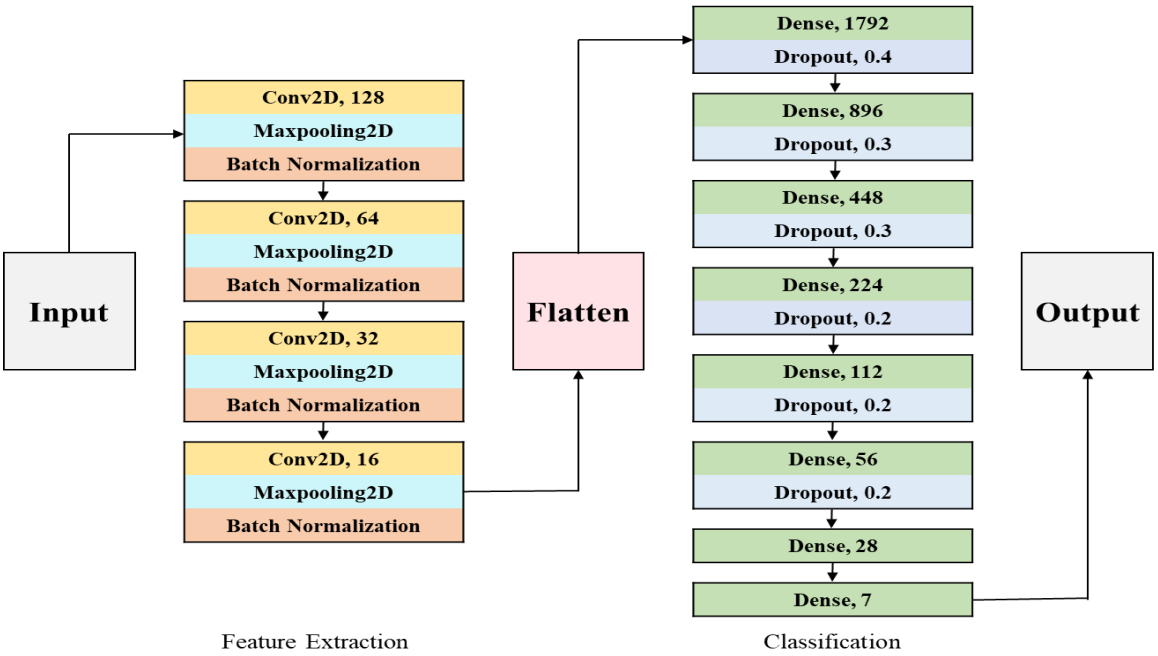

Figure 8: Proposed CNN model architecture

Convolutional Layer

This layer executes the convolution operation on the input image, involving the systematic movement of a small filter (referred to as a kernel) across the input image to extract distinctive features. In our CNN model, the convolutional layer incorporates 128, 64, 32, and 16 filters, each with dimensions of (5, 5), (3, 3), (3, 3), and (3, 3), respectively. These filters traverse the input tensor, generating 128, 64, 32, and 16 output feature maps. The 'valid' padding scheme is employed, implying that no additional padding is applied to the input image or feature map. Consequently, the size of the output feature maps is dictated by the dimensions of the input tensor and the filter sizes. The Activation layer introduces a Rectified Linear Unit (ReLU) activation function to the convolutional layer's output, incorporating non-linearity into the model.

Pooling Layer

This layer reduces the spatial dimensions of the feature maps by performing a pooling operation (e.g., max-pooling or average pooling). This helps to reduce the computational complexity of the model and also helps to make the model more robust to small variations in the input data. In our CNN model, we feed the input feature map to the MaxPooling2D layer. The MaxPooling2D layer employs a 2x2 max pooling operation on the input feature map, thereby reducing its spatial dimensions by a factor of 2 in both height and width directions. During this max pooling operation, the input feature map is partitioned into non-overlapping regions of size (2, 2), and the highest value in each region becomes the output value for that region. This operation effectively trims down the spatial size of the feature map while preserving the most crucial information.

Batch Normalization Layer

For normalizing the input data to our training, we use batch normalization. Normalization preprocessed the numerical data and bring them into a common scale without changing the data shape. The batch normalization processes a mini batch at a time and makes our CNN more stable and faster. A general formula for batch normalization can be defined as follow;

$$z^N = \left(\frac{z - m_z}{s_z}\right) \quad (1)$$

where m_z = mean of the neurons' output and s_z = the standard deviation of the neurons' output.

Flatten Layer

A Flatten layer is a type of layer in a NN that flattens the input into a one-dimensional array or vector. It is commonly used in between convolutional layers and fully connected layers in a CNN model to convert the output of the convolutional layers into a format that can be passed to the fully connected layers.

Dropout Layer

The dropout layer is employed to randomly exclude selected neurons during training. This means that the weights and biases for certain neurons are not updated in the backward pass because the forward pass momentarily ignores the contribution of neuron activations. In our CNN model, six dropout layers are incorporated after the dense layers with dropout rates set at 0.4, 0.3, 0.3, 0.2, 0.2, and 0.2. This means that 40%, 30%, 30%, 20%, 20%, and 20% of neuron activations in the forward pass are temporarily omitted. This regularization technique helps prevent overfitting during the training of the CNN.

Dense Layer

A Dense layer, also referred to as a fully connected layer, is a type of NN layer where each neuron in the layer is linked to every neuron in the preceding layer. In a Dense layer, each neuron receives input from all neurons in the prior layer and produces a singular value transmitted to each neuron in the subsequent layer. In our CNN model, eight dense layers are present with a varying number of units: 1792, 896, 448, 224, 112, 56, 28, and 7 neurons. The first seven dense layers employ the "ReLU" activation function, which eliminates negative data by providing a direct output if the input is true; otherwise, it yields zero. The final dense layer constitutes a fully connected layer with 7 neurons and utilizes the "softmax" activation function for multi-class classification. The unit parameter specifies the seven classes to be predicted, such as A, B, C, D, E, F, or G.

3.6 Training and Testing

Training

Training a CNN model encompasses several steps, including defining the model architecture, compiling the model, preparing the training data, fitting the model to the data, and evaluating its performance. Our designated CNN model was fitted to the training data

for 80 epochs with a batch size of 16. Figure 9 illustrates some training images from various classes in the 150X magnification set.

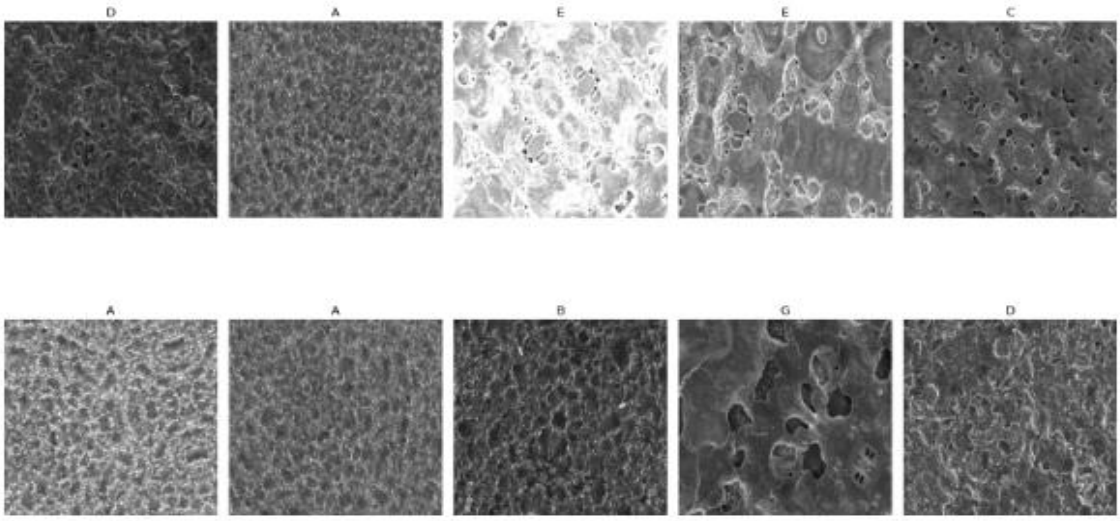

Figure 9. Random training images of different class

In the training of our CNN model, the compile method is employed to configure the learning process. Specifically, it establishes the optimizer, loss function, and evaluation metrics, which are discussed below.

Optimizer

To enhance the performance of our CNN model, we employ an optimization algorithm that contribute to reducing the error rate. In this context, the "Adam" optimizer is utilized with a learning rate of 0.0001. Adam is an extension of stochastic gradient descent (SGD) designed to update network weights more efficiently than conventional SGD. It demonstrates computational efficiency and demands less memory than alternative algorithms. The adaptive learning rate feature allows it to dynamically adjust the learning rate during training. The weights in Adam optimizer are updated using the following equation.

$$w_t = w_{t-1} - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t + \epsilon}} \tag{2}$$

where, w = model weights and η = Step size (depend on iteration),

Loss Function and Metrics

To assess the performance of our multiclass classification model, we employ the "categorical cross-entropy" loss function. In the context of multi-class classification problems, where an example belongs to only one of several potential categories, categorical cross-entropy serves as a suitable loss function. This function determines the loss through a specific equation which is presented below.

$$CE = -\sum_i^C t_i \, log \, (f(s)_i) \tag{3}$$

Where, t_i and s_i = ground truth.

The metrics parameter is set to evaluate the model using the accuracy metric, which gauges the proportion of correctly classified examples. Throughout training, these metric monitors progress and guides decisions regarding when to conclude the training process. It provides both accuracy and validation during the model training phase.

Testing

Testing constitutes a pivotal phase in the development of the CNN model, ensuring its capacity to generalize effectively to new data and deliver robust performance in real-world scenarios. In our study, distinct test datasets corresponding to 150X, 250X, and 500X magnifications are employed, representing the type of data the model is expected to handle and ensuring the reliability of the evaluation process. Figure 10 showcases some test images from various classes within the 150X magnification set. To conduct the testing of our CNN model, we utilize the evaluate method from the model object in Keras. This method takes input from the test images and their corresponding labels, providing the accuracy of the model on the test set as the output.

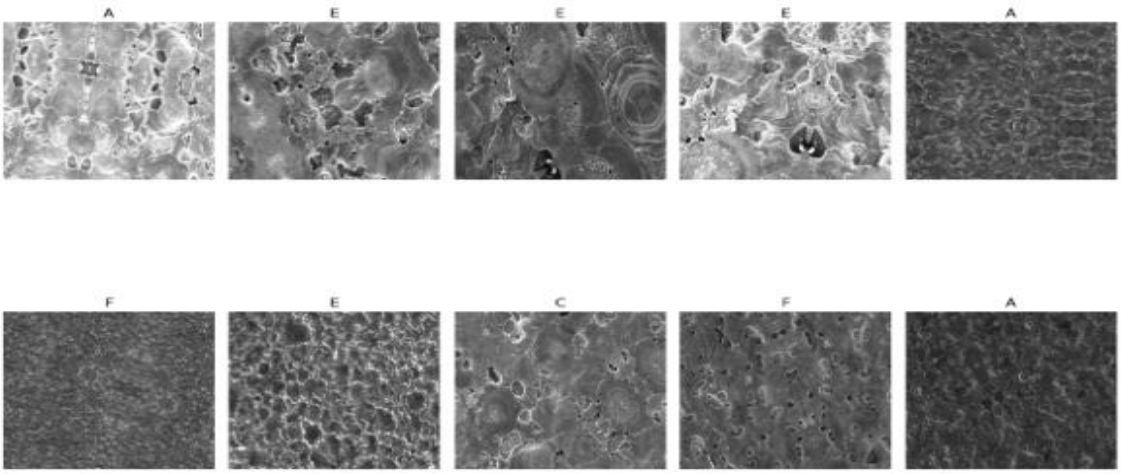

Figure 10 Some test images of different class

3.7 Performance Evaluation Metrices

Evaluating a CNN model's accuracy and efficiency is necessary to determine how well it performs. Several performance metrices must be used in order to fully comprehend the capabilities of the model. Furthermore, assessing the effectiveness of the model and highlighting areas for development is made easier by comparing its performance with that of other models. Our CNN model's performance was evaluated using a range of widely used performance metrics. These are described in details below.

Confusion Matrix

A confusion matrix is a table used to evaluate the performance of a classification model, including CNNs. The table compares the actual values of the target variable (true labels) with the predicted values generated by the model. The confusion matrix is usually represented as a table with four entries, as shown in Figure 3.11. True positive (TP)

represents the number of positive samples correctly classified as positive by the model, false positive (FP) represents the number of negative samples incorrectly classified as positive by the model, true negative (TN) represents the number of negative samples correctly classified as negative by the model, and false negative (FN) represents the number of positive samples incorrectly classified as negative by the model. A confusion matrix provides a visual representation of the model's performance, showing where the model is making correct and incorrect predictions. It can be used to calculate various performance metrics such as accuracy, precision, recall, F1 score, and others.

Accuracy

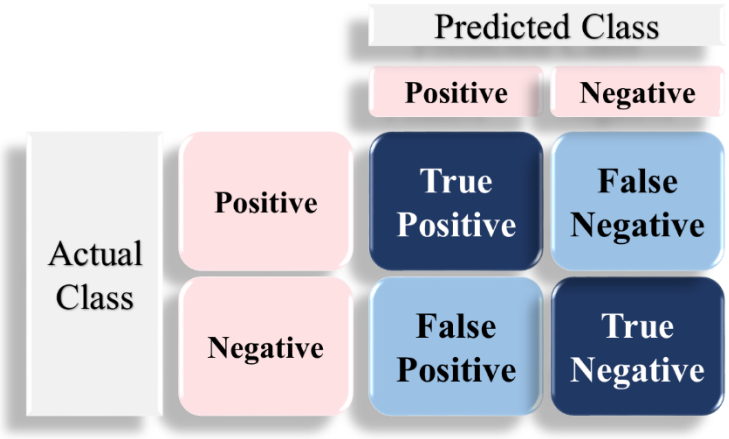

Figure 11: Confusion matrix structure

The most used performance factor in DL is accuracy. It is very easy to interpret the performance of CNNs using accuracy as it gives the percentage value. Accuracy is the ratio of correct prediction and the total number of classes. It gives better results with the symmetric dataset, where false positive and false negative are the same. The basic equation for calculating the accuracy is as follows:

$$\text{Accuracy} = \frac{True\ Positive + True\ Negative}{True\ positive + True\ Negative + False\ Positive + False\ Negative} \quad (4)$$

Precision

Precision is a performance metric used in classification problems to measure the percentage of true positive samples out of the total number of positive samples predicted by the model. In the context of CNNs, precision measures how many of the predicted positive samples are actually positive. Precision is an important metric when the cost of a false positive is high. The basic equation for calculating the precision is as follows:

$$\text{Precision} = \frac{\text{True Positive}}{\text{True positive} + \text{False Positive}} \quad (5)$$

Recall

Recall is a performance metric used in classification problems to measure the percentage of true positive samples out of the total number of actual positive samples in the test dataset. In the context of CNNs, recall measures how many of the positive samples in the test dataset the model correctly identifies. It is defined as:

$$\text{Recall} = \frac{\text{True Positive}}{\text{True positive+False Negative}} \tag{6}$$

High recall indicates that the model has a low false negative rate, meaning that the model is correctly identifying positive samples and not missing any positive samples. A low recall indicates a high false negative rate, meaning that the model is missing positive samples, and there are many positive samples being classified as negative.

F1 Score

F1 score is a performance metric used in classification problems that combines precision and recall to provide a balanced measure of the model's performance. F1 score provides a way to balance the trade-off between precision and recall. In the context of CNNs, F1 score is the harmonic mean of precision and recall and is defined as:

$$\text{F1 score} = 2 \times \frac{\text{Precission} \times \text{Recall}}{\text{Precission+Recall}} \tag{7}$$

A high F1 score indicates a model with both high precision and high recall, signifying accurate predictions of positive samples without missing any or falsely categorizing negative samples. Conversely, a low F1 score suggests either low precision or low recall, indicating potential issues with missing positive samples or erroneously classifying negative samples as positive. F1 score is an important metric when the class distribution is imbalanced, meaning that there are more samples in one class than the other. In such cases, accuracy alone may not be a good performance metric as a high accuracy could be achieved by always predicting the majority class. In these instances, the F1 score provides a more comprehensive measure of the model's performance by considering both precision and recall for each class.

4. Results and Discussions

In this chapter, our research's results are presented and discussed regarding the aim of the study, which was able to classify the surface roughness of Ti-5Al-2.5Sn from an image using a CNN model. We utilized three distinct datasets to predict surface roughness: a 150X magnification set comprising 2097 images, a 250X magnification set with 2103 images, and a 500X magnification set containing 2102 images. Here, the classification performance was assessed separately for each of these datasets, and a comparative analysis of their performances is provided at the end of the chapter. The performance evaluation of a CNN model should be done using multiple performance metrics to get a comprehensive understanding of the model's strengths and weaknesses. Therefore, for performance evaluation; Accuracy, Confusion Matrix, Precision, Recall, and F1 Score metrices are considered in our research. The basic of these metrices are already described in the previous Methodology chapter.

4.1 SR Prediction Results using 150x Magnified Images Dataset

The result analysis of our CNN model involves evaluating the model's performance on the given image dataset. Specifically, here, we focused on the results of 150X magnified dataset, comprising 2097 images. This dataset was divided into training (1467), testing (420), and validation (210) sets, each categorized into seven classes. The following metrices are used for evaluating the performance of our CNN model for this 150X magnified dataset.

4.1.1 Training and Validation Result

The training results in classification reflects how well the model has learned to make accurate predictions on the training data. Here we considered a total of 1467 images for training and 210 images for validation in 150X magnified dataset. During the training process, the model is exposed to a set of labeled examples with 150X magnification, and it learns to map the input features to the output class labels. Figure 12 shows the outputs of the last five epochs during the training process.

```
Epoch 76/80
92/92 - 12s - loss: 0.5632 - accuracy: 0.7737 - val_loss: 0.7922 - val_accuracy: 0.7238 - lr: 3.1623e-10 - 12s/epoch - 126ms/step
Epoch 77/80
92/92 - 12s - loss: 0.5515 - accuracy: 0.7689 - val_loss: 0.7908 - val_accuracy: 0.7143 - lr: 3.1623e-10 - 12s/epoch - 129ms/step
Epoch 78/80
92/92 - 12s - loss: 0.5277 - accuracy: 0.7866 - val_loss: 0.7907 - val_accuracy: 0.7190 - lr: 3.1623e-10 - 12s/epoch - 125ms/step
Epoch 79/80
92/92 - 12s - loss: 0.5261 - accuracy: 0.7771 - val_loss: 0.7922 - val_accuracy: 0.7190 - lr: 1.0000e-10 - 12s/epoch - 126ms/step
Epoch 80/80
92/92 - 12s - loss: 0.5073 - accuracy: 0.8044 - val_loss: 0.7903 - val_accuracy: 0.7190 - lr: 1.0000e-10 - 12s/epoch - 128ms/step
```

Figure 122: Code Snippet of outputs of the last five epochs of the training process

Training and Validation Accuracy

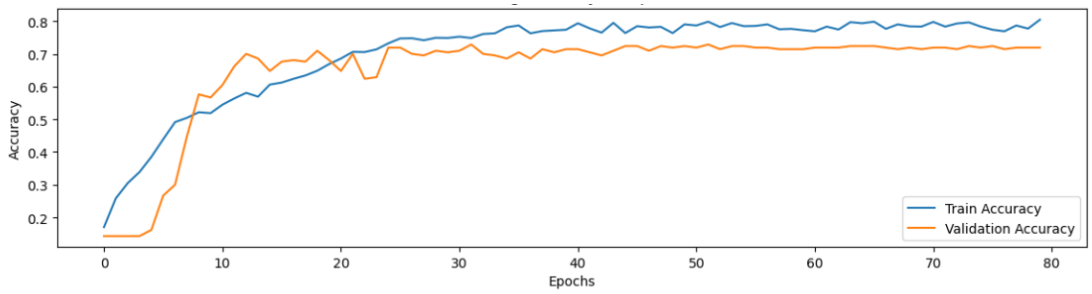

Figure 13: Training and validation accuracy for the magnification factor 150X

Accuracy is the most common evaluation metric for classification problems. A high training accuracy means that the model can fit the training data well. The validation accuracy is the performance metric that measures how well the model generalizes to new, unseen data. Figure 13 shows the accuracy graph for both training and validation set. The

graph shows that the validation accuracy starts to grow from the 5th epoch considerably and continues to increase until the 15th. After that it maintains a stable value until the 80th epoch. After 80th epochs, the training accuracy is 0.8044, and the validation accuracy is 0.7190. Here, the difference between training accuracy and validation accuracy are lower. This suggests that the model is generalizing well to new, unseen data and is balanced with the training data. This situation is desirable as it indicates that the model has learned the underlying patterns in the data without simply memorizing the training set.

Training and Validation Loss

The training loss measures how well the model fits the training data, while validation loss measures how well the model generalizes to new, unseen data. Both metrics are essential in evaluating the performance of our CNN model and preventing overfitting. Figure 14 shows the loss graph for both training and validation. During the training of our 150X magnified dataset, the loss is 0.5073 at the last epoch, and the validation loss is 0.7903 at the last epoch. The loss curves for both training and validation are substantially more stable. Training loss and validation loss are highly correlated. Therefore, there is no overfitting concern.

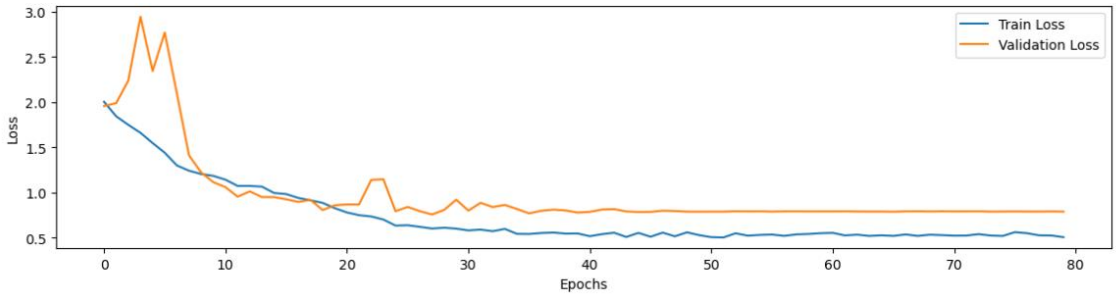

Figure 14: Training loss and validation loss curve

4.1.2 Testing Result

After training our model with training set and optimized to minimize the loss function, we assess its performance on a separate test set containing 420 new images across seven categories. This evaluation helps us understand how well our CNN model generalizes to fresh, unseen data, measuring its error rate on the new test dataset. The test set estimates the model's generalization error, which is the error rate on new data. The test result is reported using evaluation metrics, like accuracy, confusion matrix, precision, recall, and F1 score.

Testing Accuracy

Test accuracy measures the percentage of correctly classified instances in the test set. Table 3 shows the testing accuracy and error for 420 images of 7 classes, where each class contains 60 images for finding testing accuracy. As seen from the table, the overall testing accuracy for the 150X magnification factor is 69.5%, and the error is 30.47%. That means the model predicts 69 images accurately out of 100 images.

Table 3: Testing accuracy and error for 150X magnification factor

Class	No. of images	Correctly predicted	Incorrectly predicted	Accuracy (%)
A	60	41	19	68.3
B	60	43	17	71.6
C	60	39	21	65.0
D	60	43	17	71.6
E	60	43	17	71.6
F	60	41	19	68.3
G	60	42	18	70.0
Average	420	292	128	69.5

Confusion Matrix

The confusion matrix gives the summary of our prediction results. The accuracy metric alone may not be able to accurately interpret the actual result for variation in the number of observations in each class.

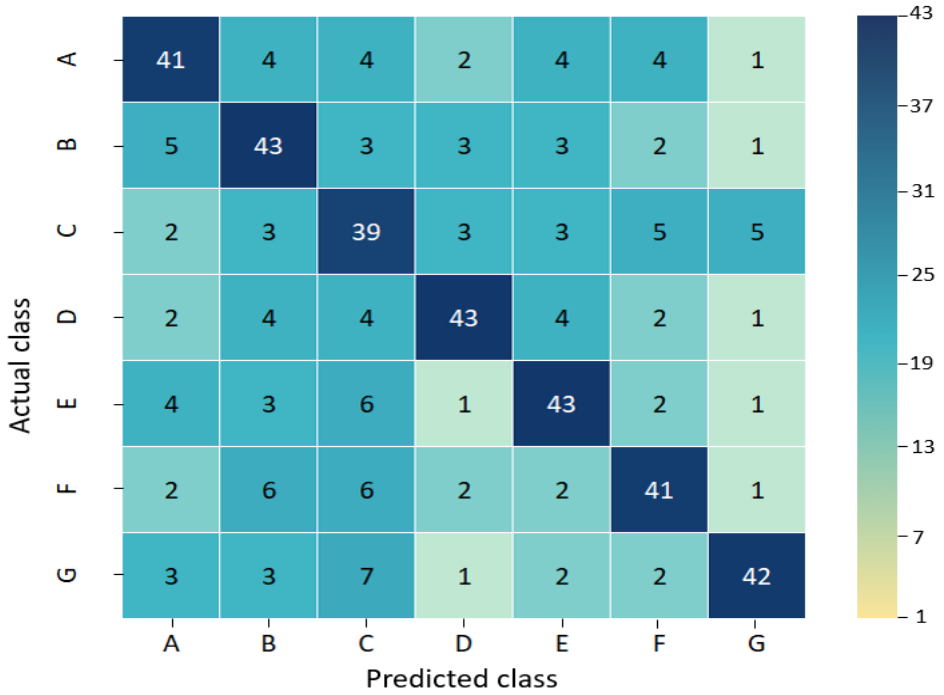

Figure 15: Confusion matrix for 150X magnification

In that case, confusion matrix gives the number of correct and incorrect predictions together, which helps us to interpret the model more easily. Figure 15 shows the confusion matrixes for our 150X magnification dataset.

Precision, Recall, and F1 Score

We now examine the precision, recall, and F1 score performances in order to present a thorough analysis of our test results. These metrics are calculated individually for all seven classes, and the outcomes are summarized in Table 4, showing the precision, recall, and F1 scores for each class in the 150X magnification dataset.

The weighted average precision, calculated for 420 test images, is 0.71, indicating a high valid positive rate. This implies that the model accurately identifies negative samples. However, there are instances where negative samples are incorrectly classified as positive. The average recall score, determined for 420 test images, stands at 0.69. A high recall score suggests a low false negative rate, indicating that the model correctly identifies positive samples, with few instances of misclassifying positive samples as negative. The average F1-score, computed for 420 test images, is 0.70. This signifies that the model achieves a balance between precision and recall, effectively predicting positive samples without significant misses.

Table 4: Precision, recall, and F1 score for 150X magnification

Class	Precision	Recall	F-1 score	Test Images
A	0.69	0.68	0.69	60
B	0.67	0.71	0.69	60
C	0.57	0.65	0.60	60
D	0.76	0.71	0.73	60
E	0.72	0.71	0.72	60
F	0.73	0.68	0.70	60
G	0.80	0.70	0.75	60
Average	0.71	0.69	0.70	420

4.2 SR Prediction Results using 250x Magnified Images Dataset

Now, here, we focus on the results of 250X magnified dataset, comprising 2103 images. This dataset was divided into training (1473), testing (420), and validation (210) sets, each categorized into seven classes. The following metrices are used for evaluating the performance of our CNN model for this 250X magnified dataset.

4.2.1 Training and Validation Result

The training results in classification reflects how well the model has learned to make accurate predictions on the training data. Here we considered a total of 1473 images for training and 210 images for validation in 250X magnified dataset. During the training process, the model is exposed to a set of labeled examples with 250X magnification, and it learns to map the input features to the output class labels. Figure 16 shows the outputs of the last five epochs during the training process.

```
93/93 - 11s - loss: 0.4348 - accuracy: 0.7855 - val_loss: 0.9901 - val_accuracy: 0.7190 - lr: 1.0000e-09 - 11s/epoch - 123ms/step
Epoch 76/80
93/93 - 11s - loss: 0.4350 - accuracy: 0.7889 - val_loss: 0.9894 - val_accuracy: 0.7190 - lr: 1.0000e-09 - 11s/epoch - 123ms/step
Epoch 77/80
93/93 - 12s - loss: 0.4347 - accuracy: 0.7923 - val_loss: 0.9888 - val_accuracy: 0.7190 - lr: 3.1623e-10 - 12s/epoch - 124ms/step
Epoch 78/80
93/93 - 12s - loss: 0.4445 - accuracy: 0.7889 - val_loss: 0.9898 - val_accuracy: 0.7143 - lr: 3.1623e-10 - 12s/epoch - 125ms/step
Epoch 79/80
93/93 - 11s - loss: 0.4369 - accuracy: 0.7848 - val_loss: 0.9887 - val_accuracy: 0.7190 - lr: 3.1623e-10 - 11s/epoch - 124ms/step
Epoch 80/80
93/93 - 11s - loss: 0.4364 - accuracy: 0.7923 - val_loss: 0.9892 - val_accuracy: 0.7238 - lr: 3.1623e-10 - 11s/epoch - 123ms/step
```

Figure 16: Code Snippet of outputs of the last five epochs of the training process

Training and Validation Accuracy

Figure 17 shows the accuracy graph for both training and validation dataset. The graph shows that the validation accuracy starts to grow from the 8th epoch considerably and continues to increase until the 12th. After that it follows a decrease and increase pattern until the 25th epoch. After that it maintain a stable pattern until the 80th epoch. After 80th epochs, the training accuracy is 0.7923, and the validation accuracy is 0.7328. Here, the difference between training accuracy and validation accuracy are lower. This suggests that the model is generalizing well to new, unseen data and is balanced with the training data. This situation is desirable as it indicates that the model has learned the underlying patterns in the data without simply memorizing the training set.

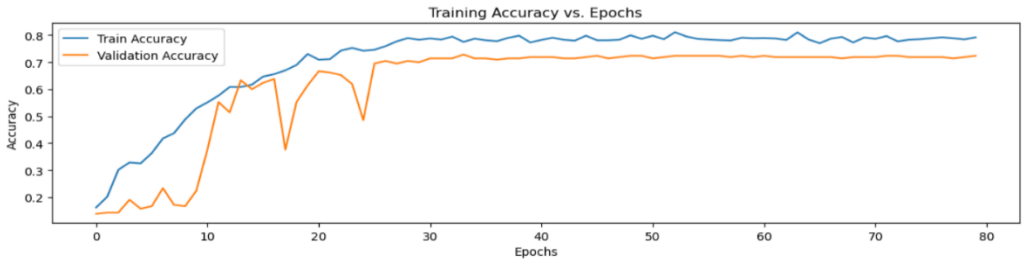

Figure 17: Train and validation accuracy for 250X magnified dataset

Training and Validation Loss

Figure 18 shows the loss graph for both training and validation. During the training of our 250X magnified dataset, the loss is 0.4364 at the last epoch, and the validation loss is 0.9892 at the last epoch. The loss curves for both training and validation are substantially more

stable. Training loss and validation loss are highly correlated. Therefore, there is no overfitting concern.

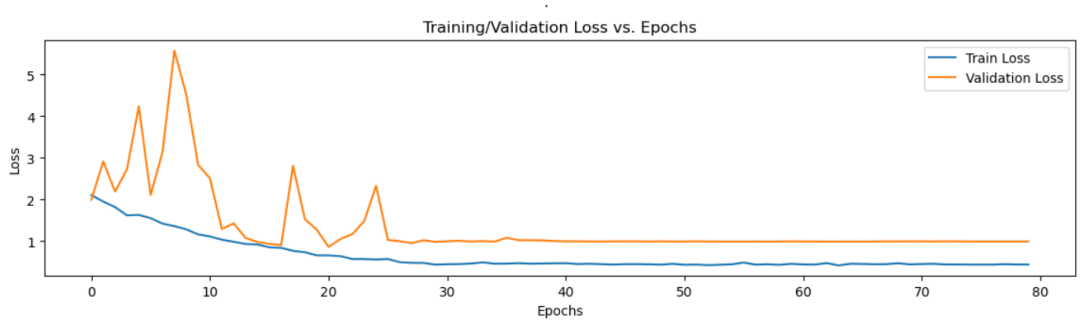

Figure 18: Train loss and validation loss curve for 250X magnified dataset

4.2.2 Testing Result

After training our model with training set of 250X magnification dataset and optimized to minimize the loss function, we assess its performance on a separate testing dataset containing 420 new images across seven categories. This evaluation helps us understand how well our CNN model generalizes to fresh, unseen data, measuring its error rate on the new test dataset. The test set estimates the model's generalization error, which is the error rate on new data. The test result is reported using evaluation metrics, like accuracy, confusion matrix, precision, recall, and F1 score.

Testing Accuracy

Table 5 shows the testing accuracy and error for 420 images of 7 classes of 250X magnification dataset, where each class contains 60 images for finding testing accuracy. As seen from the table, the overall testing accuracy for the 250X magnification factor is 61.90%, and the error is 38.10%. That means the model predicts almost 62 images accurately out of 100 images.

Table 5: Testing accuracy and error for 250X magnification factor

Class	No. of images	Correctly predicted	Incorrectly predicted	Accuracy (%)
A	60	37	23	61.66
B	60	37	23	61.66
C	60	34	26	56.67
D	60	43	17	71.60
E	60	39	21	65.00

F	60	33	27	55.00
G	60	37	23	61.66
Average	420	260	160	61.90

Confusion Matrix

The confusion matrix gives the summary of our prediction results. The accuracy metric alone may not be able to accurately interpret the actual result for variation in the number of observations in each class. In that case, confusion matrix gives the number of correct and incorrect predictions together, which helps us to interpret the model more easily. Figure 19 shows the confusion matrixes for our 250X magnification dataset.

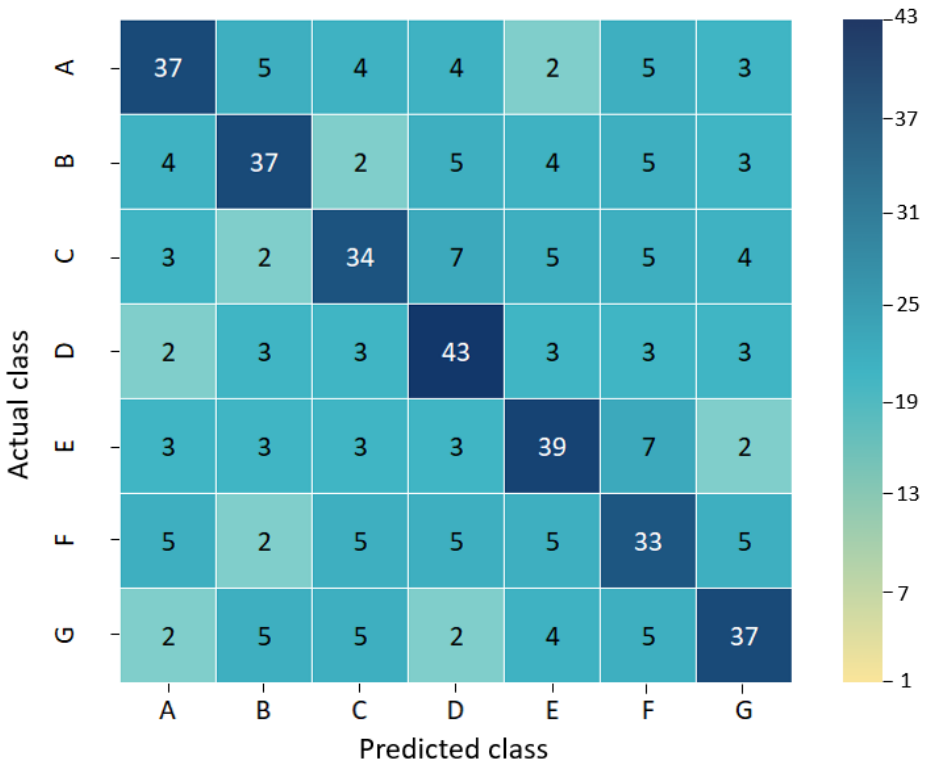

Figure 19: Confusion matrix for 250X magnified dataset

Precision, Recall, and F1 Score

We now examine the precision, recall, and F1 score performances in order to present a thorough analysis of our test results. These metrics are calculated individually for all seven classes, and the outcomes are summarized in Table 6, showing the precision, recall, and F1

scores for each class in the 250X magnification dataset. The weighted average precision, calculated for 420 test images, is 0.60, indicating a high actual positive rate. This implies that the model accurately identifies negative samples. However, there are instances where negative samples are incorrectly classified as positive. The average recall score, determined for 420 test images, stands at 0.60. A high recall score suggests a low false negative rate, indicating that the model correctly identifies positive samples, with few instances of misclassifying positive samples as negative. The average F1-score, computed for 420 test images, is 0.60. This signifies that the model achieves a balance between precision and recall, effectively predicting positive samples without significant misses. However, all the precision, recall, and F1 score values are comparatively much lower than 150X magnification dataset.

Table 6: Precision, recall, and F1 score for 250X magnification

Class	Precision	Recall	F-1 score	support
A	0.63	0.61	0.62	60
B	0.68	0.61	0.64	60
C	0.59	0.56	0.57	60
D	0.66	0.71	0.69	60
E	0.62	0.65	0.63	60
F	0.49	0.55	0.52	60
G	0.59	0.61	0.60	60
Average	0.60	0.60	0.60	420

4.3 SR Prediction Results using 500x Magnified Images Dataset

Now, here, we focus on the results of 500X magnified dataset, comprising 2102 images. This dataset was divided into training (1472), testing (420), and validation (210) sets, each categorized into seven classes. The following metrices are used for evaluating the performance of our CNN model for this 500X magnified dataset. Result analysis for 500X magnified images is segmented into two parts: training and validation result analysis and testing result analysis.

4.3.1 Training and Validation Result

The training results in classification reflects how well the model has learned to make accurate predictions on the training data. Here we considered a total of 1472 images for training and 210 images for validation in 500X magnified dataset. During the training process, the model is exposed to a set of labeled examples with 500X magnification, and it

learns to map the input features to the output class labels. Figure 20 shows the outputs of the last five epochs during the training process.

```
92/92 - 12s - loss: 0.3667 - accuracy: 0.8764 - val_loss: 0.6882 - val_accuracy: 0.7810 - lr: 3.1623e-09 - 12s/epoch - 128ms/step
Epoch 76/80
92/92 - 12s - loss: 0.3749 - accuracy: 0.8770 - val_loss: 0.6883 - val_accuracy: 0.7857 - lr: 3.1623e-09 - 12s/epoch - 130ms/step
Epoch 77/80
92/92 - 12s - loss: 0.3619 - accuracy: 0.8764 - val_loss: 0.6897 - val_accuracy: 0.7762 - lr: 3.1623e-09 - 12s/epoch - 128ms/step
Epoch 78/80
92/92 - 12s - loss: 0.3345 - accuracy: 0.8770 - val_loss: 0.6894 - val_accuracy: 0.7762 - lr: 3.1623e-09 - 12s/epoch - 125ms/step
Epoch 79/80
92/92 - 12s - loss: 0.3620 - accuracy: 0.8736 - val_loss: 0.6886 - val_accuracy: 0.7762 - lr: 1.0000e-09 - 12s/epoch - 127ms/step
Epoch 80/80
92/92 - 12s - loss: 0.3390 - accuracy: 0.8818 - val_loss: 0.6884 - val_accuracy: 0.7762 - lr: 1.0000e-09 - 12s/epoch - 126ms/step
```

Figure 20: Code Snippet of outputs of the last five epochs of the training process

Training and Validation Accuracy

Figure 21 shows the accuracy graph for both training and validation dataset. The graph shows that the validation accuracy starts to grow from the 5th epoch considerably and continues to increase until the 35th. After that it maintains a stable value until the 80th epoch. After 80th epochs, the training accuracy is 0.8818, and the validation accuracy is 0.7762. Here, the difference between training accuracy and validation accuracy are lower. This suggests that the model is generalizing well to new, unseen data and is balanced with the training data. This situation is desirable as it indicates that the model has learned the underlying patterns in the data without simply memorizing the training set.

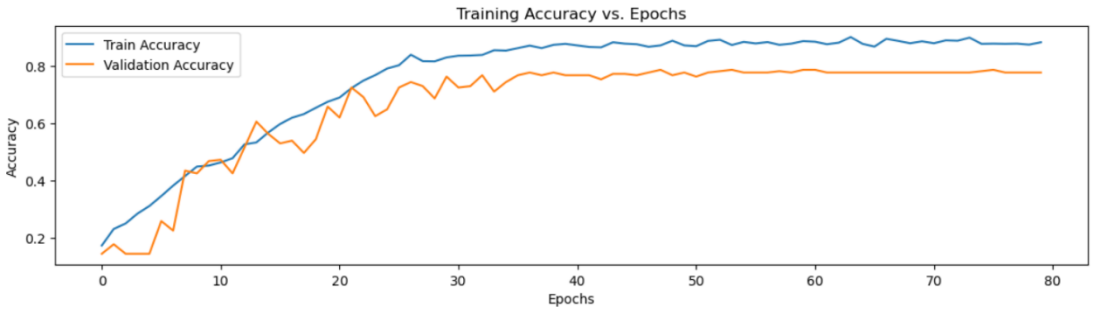

Figure 21: Train and validation accuracy for 500X magnified dataset

Training and Validation Loss

Figure 22 shows the loss graph for both training and validation. During the training of our 500X magnified dataset, the loss is 0.3390 at the last epoch, and the validation loss is 0.6884 at the last epoch. The loss curves for both training and validation are substantially stable. Training loss and validation loss are highly correlated. Therefore, there is no overfitting concern.

Figure 22: Train loss and validation loss curve for 500X magnified dataset

4.3.2 Testing Result

After training our model with training set of 500X magnification dataset and optimized to minimize the loss function, we assess its performance on a separate testing dataset containing 420 new images across seven categories. This evaluation helps us understand how well our CNN model generalizes to fresh, unseen data, measuring its error rate on the new test dataset. The test set estimates the model's generalization error, which is the error rate on new data. The test result is reported using evaluation metrics, like accuracy, confusion matrix, precision, recall, and F1 score.

Testing Accuracy

Table 7 shows the testing accuracy and error for 420 images of 7 classes of 500X magnification dataset, where each class contains 60 images for finding testing accuracy. As seen from the table, the overall testing accuracy for the 500X magnification factor is 75.71%, and the error is 24.29%. That means the model predicts almost 75 images accurately out of 100 images.

Table 7: Testing accuracy and error for 500X magnification factor

Class	No. of images	Correctly predicted	Incorrectly predicted	Accuracy (%)
A	60	44	16	73.33
B	60	47	13	78.33
C	60	44	16	73.33
D	60	45	15	75.0
E	60	51	9	85.0
F	60	43	17	71.66
G	60	44	16	73.33
Average	420	318	102	75.71

Confusion Matrix

The confusion matrix gives the summary of our prediction results. Figure 23 shows the confusion matrixes for our 500X magnification dataset.

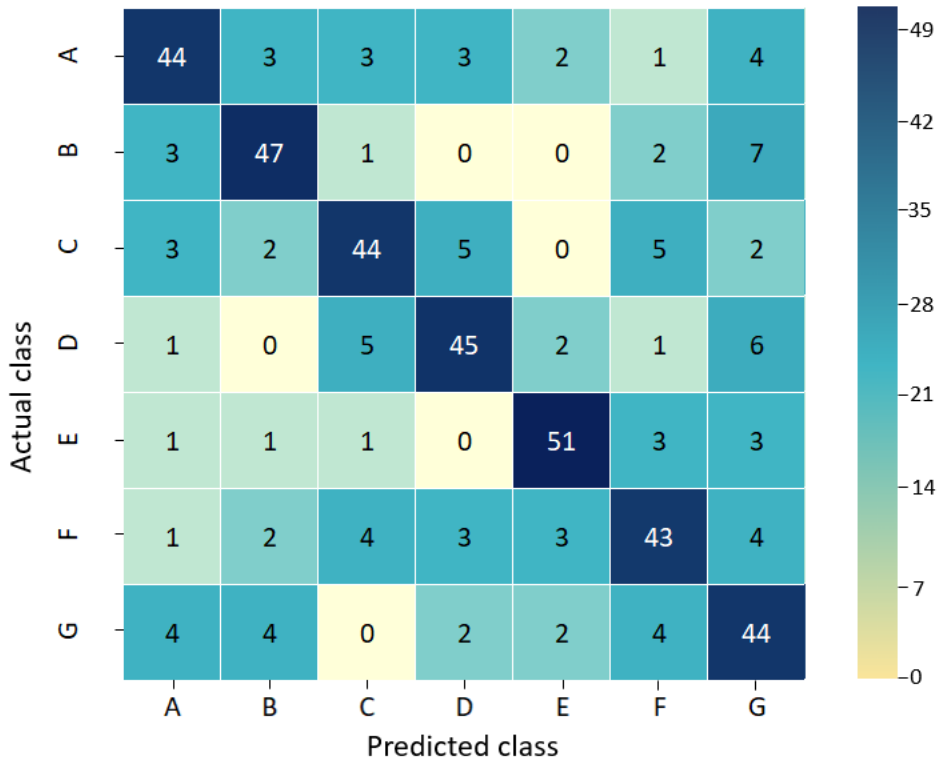

Figure 23: Confusion matrix for 500X magnified dataset

Precision, Recall, and F1 Score

Table 8 shows the precision, recall, and F1 scores for each class in the 500X magnification dataset. The weighted average precision, calculated for 420 test images, is 0.65, indicating a high actual positive rate. This implies that the model accurately identifies negative samples. However, there are instances where negative samples are incorrectly classified as positive. The average recall score, determined for 420 test images, stands at 0.72. A high recall score suggests a low false negative rate, indicating that the model correctly identifies positive samples, with few instances of misclassifying positive samples as negative. The average F1-score, computed for 420 test images, is 0.72. This signifies that the model achieves a balance between precision and recall, effectively predicting positive samples without significant misses.

Table 8: Precision, recall and F1 score for 500X magnification factor

Class	Precision	Recall	F-1 score	support
A	0.78	0.78	0.78	60
B	0.82	0.80	0.81	60
C	0.77	0.71	0.74	60
D	0.76	0.80	0.78	60
E	0.32	0.75	0.74	60
F	0.65	0.67	0.65	60
G	0.51	0.56	0.54	60
Average	0.65	0.72	0.72	420

4.4 Comparative Analysis among Performance Parameter Scores

Now, we conduct a comparative analysis of our results across three datasets. In our study, we utilized three distinct image datasets magnified at 150X, 250X, and 500X levels. Figure 24 illustrates the comparative performance, considering training, testing, and validation accuracy for the three datasets. The X-axis represents the three different datasets, while the Y-axis depicts the accuracy scores. Upon examining the comparison chart, it becomes evident that the dataset with images magnified at 500X provides superior accuracy compared to the 150X and 250X magnified datasets.

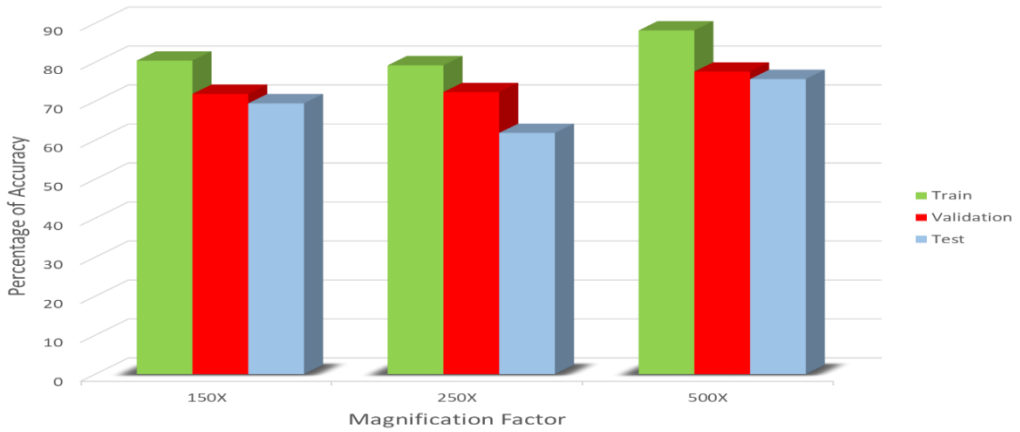

Figure 24: Accuracy comparison among three datasets

For a more detailed breakdown of the correctly predicted values and accuracy across all seven classes, we present Table 9. The accuracy percentages directly reflect the number of correctly predicted data points for each magnification level. As seen from the table, the 500X magnified images dataset exhibits the highest accuracy (75.7%), followed by 150X (69.5%) and 250X (61.9%). The higher accuracy in the 500X dataset implies better overall classification performance at higher magnification levels.

Table 9: Comparison Table of Accuracy and Correctly Predicted Data

Class	**No. of images**	**For 150X magnified images**		**For 250X magnified images**		**For 500X magnified images**	
		Correctly predicted	Accuracy (%)	Correctly predicted	Accuracy (%)	Correctly predicted	Accuracy (%)
A	60	41	68.3	37	61.66	44	73.33
B	60	43	71.6	37	61.66	47	78.33
C	60	39	65.0	34	56.67	44	73.33
D	60	43	71.6	43	71.60	45	75.0
E	60	43	71.6	39	65.00	51	85.0
F	60	41	68.3	33	55.00	43	71.66
G	60	42	70.0	37	61.66	44	73.33
Total	420	292	69.5	260	61.9	318	75.7

The Table 10 presents a comparative analysis of precision, recall, and F1-score across three magnification levels (150X, 250X, and 500X) in image datasets. This comparison allows for an in-depth evaluation of the model's efficacy at different levels of magnification. Across the magnification levels, precision varies, with the 150X dataset achieving the highest precision (0.71), followed by 500X (0.65), and then 250X (0.60). The 150X dataset demonstrates superior precision, suggesting a better ability to correctly identify positive instances. While the 250X dataset shows the lowest recall (0.60), the 150X and 500X datasets exhibit almost similar recall values (0.69 and 0.72, respectively). The 500X dataset stands out with the highest recall, indicating a better ability to capture all positive instances. Among the three datasets, the highest F1-score is observed in the 500X magnified images dataset (0.72), while the lowest F1-score is found in the 250X magnified images dataset (0.60). That means, the 500X dataset demonstrating superior performance in achieving a balance between precision and recall, while the 250X dataset shows relatively weaker performance in this aspect.

Table 10: Comparison Table of Precision, Recall, and F1-Score

Performance Parameter	150X Magnified Images Dataset	250X Magnified Images Dataset	500X Magnified Images Dataset
Average Precision	0.71	0.60	0.65
Average Recall	0.69	0.60	0.72
Average F1-Score	0.70	0.60	0.72

4.5 Optimal Dataset and its Scores

Based on the provided data, the 500X magnified images dataset delivers the best results among the three magnification levels evaluated. Here's a detailed description of why the 500X dataset stands out.

Higher Accuracy: The 500X magnified images dataset achieves the highest accuracy percentage of 75.7%. This indicates that the model trained on the 500X dataset correctly classified the highest proportion of instances compared to the other magnification levels.

Balanced Precision and Recall: Although precision and recall measures may vary across magnification levels, the 500X dataset demonstrates competitive values. It achieves an average precision of 0.65 and an average recall of 0.72, suggesting a good balance between minimizing false positives and false negatives.

Superior F1-Score: The F1-score, which combines precision and recall into a single metric, is consistent across the magnification levels. The 500X dataset achieves an average F1-score of 0.72, indicating a robust performance in terms of both precision and recall.

Effective Identification of Features: Higher magnification levels, such as 500X, typically provide finer details and clearer images, enabling the model to identify subtle features with greater accuracy. This results in more precise classification and higher overall performance.

Real-world Applicability: In many real-world scenarios, particularly in fields like medical imaging or materials science, where detailed examination of microscopic structures is crucial, higher magnification levels are preferred for accurate analysis and diagnosis. Therefore, a model trained on 500X magnified images is more likely to generalize well and perform effectively in practical applications.

In summary, the 500X magnified images dataset delivers the best results due to its higher accuracy, competitive precision and recall, balanced F1-score, and suitability for real-world applications requiring detailed analysis of microscopic structures.

5. Conclusion and Future Research Scope

5.1 Conclusion

Surface Roughness measurement stands as a pivotal parameter with wide-ranging implications across numerous industries, encompassing aerospace, materials science, and medical imaging. This research represents a significant advancement in the field of SR prediction using CNN applied to SEM images. By addressing the gap in existing literature regarding the utilization of CNNs for SR prediction and considering the impact of varying magnification levels, this study provides valuable insights into optimizing surface quality assessment across diverse industries. The findings underscore the critical importance of selecting appropriate magnification levels in SEM imaging for accurate SR prediction. The comparative analysis revealed that the dataset magnified at 500X consistently outperformed those at 150X and 250X, demonstrating superior accuracy, precision, recall, and F1-score. This indicates that higher magnification levels offer finer detail and clearer images, enabling the CNN model to discern subtle features with increased accuracy. The successful application of CNN in analyzing SEM images for SR prediction represents a significant step forward in surface quality assessment methodologies. By leveraging CNN (a DL technique), this research provides a robust framework for enhancing SR measurement precision and efficiency in various industrial and research domains.

5.2 Limitations and Future Research Scopes

This research has contributed valuable insights into optimizing SR measurement methodologies. However, as with any scientific endeavor, there are inherent limitations to consider, along with opportunities for future exploration and advancement. Below, we outline these aspects to provide a comprehensive understanding of the research landscape and potential avenues for further investigation.

- **Limited Dataset:** Despite efforts to collect diverse SEM images, the dataset used in this research may not fully capture the variability present in real-world scenarios. For this limited dataset, the accuracy and other performance metrices delivers comparatively lower scores. Future studies could benefit from larger and more diverse datasets to improve model generalization.
- **Single Material Focus:** This research specifically focused on titanium alloy (Ti-5Al-2.5Sn), potentially limiting the generalizability of findings to other materials. Exploring SR prediction across a wider range of materials could provide a more comprehensive understanding.
- **Magnification Levels:** While this study investigated three magnification levels, there may be additional magnification levels or imaging parameters that could impact SR prediction accuracy. Exploring a wider range of magnifications and imaging settings could provide further insights.
- **Evaluation Metrics:** While accuracy, precision, recall, and F1-score were utilized as performance metrics, other evaluation measures such as mean squared error or root mean squared error could provide additional insights into model performance.
- **Enhanced Model Architectures:** Future research could explore more advanced CNN architectures, such as attention mechanisms or capsule networks, to further improve SR prediction accuracy and robustness.

Author Declarations

Funding: The authors receive no funding from any organization for this research.
Competing Interest: The authors declare no competing interest.
Ethics Approval: Not applicable.
Consent to Participate: All the authors have voluntarily agreed to participate in this research study.
Consent to Publication: All the authors have given their consent for the publication of identifiable details, which can include image(s) and/or videos and/or case history and/or details within the text ("Material") to be published in this Article.
Availability of Data and Materials: Data sets generated during the current study are available from the corresponding author on reasonable request.
Code Availability: Code written and utilized during the current study are available from the corresponding author on reasonable request.
Author's Contribution: Md. Ashikur Rahman Khan: Supervision, conceptualization, investigation, data curation, model design, analysis and interpretation of results, Draft manuscript preparation, writing – review and editing. **Md. Shariar Kabir Asif:** Data collection, conceptualization, model design, model training, analysis and interpretation of results, investigation on challenges, writing – original draft. **M.M Rahman:** Data collection, analysis and interpretation of results. **Ishtiaq Ahammad:** Draft manuscript preparation, investigation, analysis and interpretation of results, writing – review and editing.

References

[1] K. Palová, T. Kelemenová, and M. Kelemen, "Measuring Procedures for Evaluating the Surface Roughness of Machined Parts," *Applied Sciences*, vol. 13, no. 16, p. 9385, Aug. 2023, doi: 10.3390/app13169385.

[2] K. J. Kubiak, T. W. Liskiewicz, and T. G. Mathia, "Surface morphology in engineering applications: Influence of roughness on sliding and wear in dry fretting," *Tribol Int*, vol. 44, no. 11, pp. 1427–1432, Oct. 2011, doi: 10.1016/j.triboint.2011.04.020.

[3] D. Obilanade, P. Törlind, and C. Dordlofva, "Surface Roughness and Design for Additive Manufacturing: A Design Artefact Investigation," *Proceedings of the Design Society*, vol. 2, pp. 1421–1430, May 2022, doi: 10.1017/pds.2022.144.

[4] S. Ramesh, L. Karunamoorthy, and K. Palanikumar, "Measurement and analysis of surface roughness in turning of aerospace titanium alloy (gr5)," *Measurement*, vol. 45, no. 5, pp. 1266–1276, Jun. 2012, doi: 10.1016/j.measurement.2012.01.010.

[5] R. Russell *et al.*, "Qualification and certification of metal additive manufactured hardware for aerospace applications," in *Additive Manufacturing for the Aerospace Industry*, Elsevier, 2019, pp. 33–66. doi: 10.1016/B978-0-12-814062-8.00003-0.

[6] M. Radmilović-Radjenović, B. Radjenović, and Z. L. J. Petrović, "Application of level set method in simulation of surface roughness in nanotechnologies," *Thin Solid Films*, vol. 517, no. 14, pp. 3954–3957, May 2009, doi: 10.1016/j.tsf.2009.01.123.

[7] B. C. Bovas, L. Karunamoorthy, and F. B. Chuan, "Effect of surface roughness and process parameters on mechanical properties of fabricated medical catheters," *Mater Res Express*, vol. 6, no. 12, p. 125420, Jan. 2020, doi: 10.1088/2053-1591/ab6652.

[8] T. J. Webster, R. W. Siegel, and R. Bizios, "Nanoceramic surface roughness enhances osteoblast and osteoclast functions for improved orthopaedic/dental implant efficacy," *Scr Mater*, vol. 44, no. 8–9, pp. 1639–1642, May 2001, doi: 10.1016/S1359-6462(01)00873-9.

[9] Y. W. Ji *et al.*, "Comparison of Surface Roughness and Bacterial Adhesion Between Cosmetic Contact Lenses and Conventional Contact Lenses," *Eye & Contact Lens: Science & Clinical Practice*, vol. 41, no. 1, pp. 25–33, Jan. 2015, doi: 10.1097/ICL.0000000000000054.

[10] A. Kurup, P. Dhatrak, and N. Khasnis, "Surface modification techniques of titanium and titanium alloys for biomedical dental applications: A review," *Mater Today Proc*, vol. 39, pp. 84–90, 2021, doi: 10.1016/j.matpr.2020.06.163.

[11] N. Erdman, D. C. Bell, and R. Reichelt, "Scanning Electron Microscopy," in *Springer Handbook of Microscopy*, Springer, 2019, pp. 229–318. doi: 10.1007/978-3-030-00069-1_5.

[12] N. A. Hameed, I. M. Ali, and H. K. Hassun, "Calculating Surface Roughness for a Large Scale SEM Images by Mean of Image Processing," *Energy Procedia*, vol. 157, pp. 84–89, Jan. 2019, doi: 10.1016/j.egypro.2018.11.167.

[13] L. Cao, J. Li, J. Hu, H. Liu, Y. Wu, and Q. Zhou, "Optimization of surface roughness and dimensional accuracy in LPBF additive manufacturing," *Opt Laser Technol*, vol. 142, p. 107246, Oct. 2021, doi: 10.1016/j.optlastec.2021.107246.

[14] F. Martín Fernández and M. J. Martín Sánchez, "Analysis of the Effect of the Surface Inclination Angle on the Roughness of Polymeric Parts Obtained with Fused Filament Fabrication Technology," *Polymers (Basel)*, vol. 15, no. 3, p. 585, Jan. 2023, doi: 10.3390/polym15030585.

[15] P. Mishra, S. Sood, V. Bharadwaj, A. Aggarwal, and P. Khanna, "Parametric Modeling and Optimization of Dimensional Error and Surface Roughness of Fused Deposition Modeling Printed Polyethylene Terephthalate Glycol Parts," *Polymers (Basel)*, vol. 15, no. 3, p. 546, Jan. 2023, doi: 10.3390/polym15030546.

[16] M. Kopp and E. Uhlmann, "Prediction of the Roughness Reduction in Centrifugal Disc Finishing of Additive Manufactured Parts Based on Discrete Element Method," *Machines*, vol. 10, no. 12, p. 1151, Dec. 2022, doi: 10.3390/machines10121151.

[17] N. E. Sizemore, M. L. Nogueira, N. P. Greis, and M. A. Davies, “Application of Machine Learning to the Prediction of Surface Roughness in Diamond Machining,” *Procedia Manuf*, vol. 48, pp. 1029–1040, 2020, doi: 10.1016/j.promfg.2020.05.142.

[18] Z. Li, Z. Zhang, J. Shi, and D. Wu, “Prediction of surface roughness in extrusion-based additive manufacturing with machine learning,” *Robot Comput Integr Manuf*, vol. 57, pp. 488–495, Jun. 2019, doi: 10.1016/j.rcim.2019.01.004.

[19] T. Batu, H. G. Lemu, and H. Shimels, “Application of Artificial Intelligence for Surface Roughness Prediction of Additively Manufactured Components,” *Materials*, vol. 16, no. 18, p. 6266, Sep. 2023, doi: 10.3390/ma16186266.

[20] V. Lyukshin, D. Shatko, and P. Strelnikov, “Methods and approaches to the surface roughness assessment,” *Mater Today Proc*, vol. 38, pp. 1441–1444, 2021, doi: 10.1016/j.matpr.2020.08.122.

[21] M.-H. Tsai, J.-N. Lee, H.-D. Tsai, M.-J. Shie, T.-L. Hsu, and H.-S. Chen, “Applying a Neural Network to Predict Surface Roughness and Machining Accuracy in the Milling of SUS304,” *Electronics (Basel)*, vol. 12, no. 4, p. 981, Feb. 2023, doi: 10.3390/electronics12040981.

[22] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, “Definitions, methods, and applications in interpretable machine learning,” *Proceedings of the National Academy of Sciences*, vol. 116, no. 44, pp. 22071–22080, Oct. 2019, doi: 10.1073/pnas.1900654116.

[23] K. Choudhary *et al.*, “Recent advances and applications of deep learning methods in materials science,” *NPJ Comput Mater*, vol. 8, no. 1, p. 59, Apr. 2022, doi: 10.1038/s41524-022-00734-6.

[24] W.-J. Lin, S.-H. Lo, H.-T. Young, and C.-L. Hung, “Evaluation of Deep Learning Neural Networks for Surface Roughness Prediction Using Vibration Signal Analysis,” *Applied Sciences*, vol. 9, no. 7, p. 1462, Apr. 2019, doi: 10.3390/app9071462.

[25] Md. A. R. Khan, M. M. Rahman, K. Kadirgama, M. A. Maleque, and M. Ishak, “Prediction of Surface Roughness of Ti-6Al-4V in Electrical Discharge Machining: A Regression Model,” *JOURNAL OF MECHANICAL ENGINEERING AND SCIENCES*, vol. 1, pp. 16–24, Dec. 2011, doi: 10.15282/jmes.1.2011.2.0002.

[26] Md. A. R. Khan, M. M. Rahman, K. Kadirgama, M. A. Maleque, and R. A. Bakar, “Artificial Intelligence Model to Predict Surface Roughness of Ti-15-3 Alloy in EDM Process,” *World Acad Sci Eng Technol*, pp. 121–125, 2011.

[27] M. M. Rahman, Md. A. R. Khan, M. M. Noor, K. Kadirgama, and R. A. Bakar, “Optimization of Machining Parameters on Surface Roughness in EDM of Ti-6Al-4V Using Response Surface Method,” *Adv Mat Res*, vol. 213, pp. 402–408, Feb. 2011, doi: 10.4028/www.scientific.net/AMR.213.402.

[28] P. G. Benardos and G.-C. Vosniakos, “Predicting surface roughness in machining: a review,” *Int J Mach Tools Manuf*, vol. 43, no. 8, pp. 833–844, Jun. 2003, doi: 10.1016/S0890-6955(03)00059-2.

[29] A. Sudianto, Z. Jamaludin, and A. Azwan Abdul Rahman, “Prediction of Surface Roughness for Development of Smart Milling Machine,” *J Phys Conf Ser*, vol. 1201, no. 1, p. 012008, May 2019, doi: 10.1088/1742-6596/1201/1/012008.

[30] M. Cheng *et al.*, “Prediction of surface residual stress in end milling with Gaussian process regression,” *Measurement*, vol. 178, p. 109333, Jun. 2021, doi: 10.1016/j.measurement.2021.109333.

[31] M. R. Narayanan, S. Gowri, and M. M. Krishna, “On Line Surface Roughness Measurement Using Image Processing and Machine Vision,” in *Proceedings of the World Congress on Engineering*, London, U.K: WCE, Jul. 2007.

[32] M. Guo, J. Zhou, X. Li, Z. Lin, and W. Guo, “Prediction of surface roughness based on fused features and ISSA-DBN in milling of die steel P20,” *Sci Rep*, vol. 13, no. 1, p. 15951, Sep. 2023, doi: 10.1038/s41598-023-42968-4.

[33] M. K. O. Ayomoh and K. A. Abou-El-Hossein, “Surface roughness prediction using a hybrid scheme of difference analysis and adaptive feedback weights,” *Heliyon*, vol. 7, no. 3, p. e06338, Mar. 2021, doi: 10.1016/j.heliyon.2021.e06338.

[34] W. Zhang, “Surface Roughness Prediction with Machine Learning,” *J Phys Conf Ser*, vol. 1856, no. 1, p. 012040, Apr. 2021, doi: 10.1088/1742-6596/1856/1/012040.

[35] A. Varun, M. S. Kumar, K. Murumulla, and T. Sathvik, “Surface Roughness Prediction using Machine Learning Algorithms while Turning under Different Lubrication Conditions,” *J Phys Conf Ser*, vol. 2070, no. 1, p. 012243, Nov. 2021, doi: 10.1088/1742-6596/2070/1/012243.

[36] E. Kayahan, H. Oktem, F. Hacizade, H. Nasibov, and O. Gundogdu, “Measurement of surface roughness of metals using binary speckle image analysis,” *Tribol Int*, vol. 43, no. 1–2, pp. 307–311, Jan. 2010, doi: 10.1016/j.triboint.2009.06.010.

[37] M. Ulas, O. Aydur, T. Gurgenc, and C. Ozel, “Surface roughness prediction of machined aluminum alloy with wire electrical discharge machining by different machine learning algorithms,” *Journal of Materials Research and Technology*, vol. 9, no. 6, pp. 12512–12524, Nov. 2020, doi: 10.1016/j.jmrt.2020.08.098.

[38] C.-L. Fan and J.-R. Jiang, “Surface Roughness Prediction Based on Markov Chain and Deep Neural Network for Wire Electrical Discharge Machining,” in *2019 IEEE Eurasia Conference on IOT, Communication and Engineering (ECICE)*, IEEE, Oct. 2019, pp. 191–194. doi: 10.1109/ECICE47484.2019.8942705.

[39] K. Manjunath, S. Tewary, and N. Khatri, “Surface roughness prediction in milling using long-short term memory modelling,” *Mater Today Proc*, vol. 64, pp. 1300–1304, 2022, doi: 10.1016/j.matpr.2022.04.126.

[40] A. R. Babu, “A Machine Learning Approach for Prediction of Surface Roughness from the Images of Machined Components,” 2023, pp. 405–416. doi: 10.1007/978-981-19-3866-5_34.

[41] C. Palande, R. Nadar, P. Ambadekar, K. Sridhar, and T. Vashistha, “Machine Learning Application for Prediction of Surface Roughness of Milled Surface,” Springer, Singapore, 2022, pp. 203–219. doi: 10.1007/978-981-16-9952-8_20.

[42] V. Dubey, A. K. Sharma, and D. Y. Pimenov, “Prediction of Surface Roughness Using Machine Learning Approach in MQL Turning of AISI 304 Steel by Varying Nanoparticle Size in the Cutting Fluid,” *Lubricants*, vol. 10, no. 5, p. 81, May 2022, doi: 10.3390/lubricants10050081.

[43] A. Varun, M. S. Kumar, K. Murumulla, and T. Sathvik, “Surface Roughness Prediction using Machine Learning Algorithms while Turning under Different Lubrication Conditions,” *J Phys Conf Ser*, vol. 2070, no. 1, p. 012243, Nov. 2021, doi: 10.1088/1742-6596/2070/1/012243.

[44] M. Elangovan, N. R. Sakthivel, S. Saravanamurugan, Binoy. B. Nair, and V. Sugumaran, “Machine Learning Approach to the Prediction of Surface Roughness Using Statistical Features of Vibration Signal Acquired in Turning,” *Procedia Comput Sci*, vol. 50, pp. 282–288, 2015, doi: 10.1016/j.procs.2015.04.047.

[45] P. M. Abhilash and A. Ahmed, “Convolutional neural network–based classification for improving the surface quality of metal additive manufactured components,” *The International Journal of Advanced Manufacturing Technology*, vol. 126, no. 9–10, pp. 3873–3885, Jun. 2023, doi: 10.1007/s00170-023-11388-z.

[46] C. He, J. Yan, S. Wang, S. Zhang, G. Chen, and C. Ren, “A theoretical and deep learning hybrid model for predicting surface roughness of diamond-turned polycrystalline materials,” *International Journal of Extreme Manufacturing*, vol. 5, no. 3, p. 035102, Sep. 2023, doi: 10.1088/2631-7990/acdb0a.

[47] N. Gerdes, C. Hoff, J. Hermsdorf, S. Kaierle, and L. Overmeyer, “Hyperspectral imaging for prediction of surface roughness in laser powder bed fusion,” *The International Journal of Advanced Manufacturing Technology*, vol. 115, no. 4, pp. 1249–1258, Jul. 2021, doi: 10.1007/s00170-021-07274-1.

[48] C. Boga and T. Koroglu, “Proper estimation of surface roughness using hybrid intelligence based on artificial neural network and genetic algorithm,” *J Manuf Process*, vol. 70, pp. 560–569, Oct. 2021, doi: 10.1016/j.jmapro.2021.08.062.

[49] S. Zeng and D. Pi, “Milling Surface Roughness Prediction Based on Physics-Informed Machine Learning,” *Sensors*, vol. 23, no. 10, p. 4969, May 2023, doi: 10.3390/s23104969.

[50] Q. Hu, H. Xu, and Y. Chang, “Surface roughness prediction of aircraft after coating removal based on optical image and deep learning,” *Sci Rep*, vol. 12, no. 1, p. 19407, Nov. 2022, doi: 10.1038/s41598-022-24125-5.

[51] N. Chaudhary and S. A. Savari, “Simultaneous Denoising and Edge Estimation from SEM Images using Deep Convolutional Neural Networks,” in *2019 30th Annual SEMI Advanced Semiconductor Manufacturing Conference (ASMC)*, IEEE, May 2019, pp. 1–6. doi: 10.1109/ASMC.2019.8791764.

[52] H. Iwata, Y. Hayashi, A. Hasegawa, K. Terayama, and Y. Okuno, “Classification of scanning electron microscope images of pharmaceutical excipients using deep convolutional neural networks with transfer learning,” *Int J Pharm X*, vol. 4, p. 100135, Dec. 2022, doi: 10.1016/j.ijpx.2022.100135.