Proceedings of Engineering and Technology Innovation, vol. 29, 2025, pp. 11-25

Modeling and Suppression of Inhomogeneous Intensity Edible Bird Nest Images for Impurity Segmentation Using β-Variational Autoencoder

Khairul Firdaus Mohd Talib[1], Hasnida Ab-Samat[1], Lai Hoong Cheng[2], Kin Sam Yen[1,*]

[1]School of Mechanical Engineering, Universiti Sains Malaysia, Penang, Malaysia

[2]School of Industrial Technology, Universiti Sains Malaysia, Penang, Malaysia

Received 12 July 2024; received in revised form 10 November 2024; accepted 11 November 2024

https://doi.org/10.46604/peti.2024.13998

Abstract

This study proposes a β-variational autoencoder (β-VAE) method to address intensity inhomogeneity (IIH) in edible bird's nest (EBN) images, which creates an uneven intensity that obscures fine impurity details and reduces segmentation accuracy. First, the β-VAE is used to learn the feature distribution of EBN images by mapping them into a latent space. This latent space is then disentangled through selective filtering and penalization of specific latent dimensions. This unsupervised learning approach effectively captures and isolates IIH in EBN images. Additionally, enabling precise segmentation of EBN and impurities without requiring annotated datasets. It also enhances robustness in handling unseen image instances. The proposed method achieves an intersection over union of 73.08% (equivalent to a Dice coefficient of 84.44%), surpassing existing segmentation techniques. By resolving IIH, this method improves the reliability and adaptability of automated EBN inspection systems for practical applications.

Keywords: edible bird's nest (EBN), impurity segmentation, intensity inhomogeneity (IIH), β-variational autoencoder (β-VAE), anomaly detection

1. Introduction

Edible bird's nest (EBN) is a nutrient-rich animal bioproduct secreted by swiftlets during the breeding season [1]. Its health benefits such as immunity boost, cell growth promotion, and bone strengthening, have popularized its consumption in Southeast Asia [2-3]. However, beyond its therapeutic advantages, the constituents of EBN which are primarily composed of solidified swiftlet saliva, exhibit diverse characteristics, including variations in thickness, density, and strand shapes. These variations contribute to complex light interactions when illuminated, impacting its optical properties, particularly when captured as a digital image. It thrives the occurrence of intensity inhomogeneity (IIH) which refers to a smooth varying bias field in an image, leading to fluctuations in pixel intensity that affect both contrast and semantic information [4]. IIH or bias field is a multiplicative component of an image as depicted in Eq. (1), which exemplifies $v(x, y)$ as the intensity inhomogeneity-corrupted image, $u(x, y)$ as the intensity inhomogeneity-free image, $b(x, y)$ represents the bias field (IIH) and $n(x, y)$ are the noise components.

$$v(x,y)=u(x,y)\times b(x,y)+n(x,y) \tag{1}$$

In this context, when a different image with different object properties is examined, the multiplicative bias field will also vary. Thus, modulating the effects of IIH demands a robust modeling method to either filter its components independently or approximate its distribution in the original image to identify regions corresponding to the focused object. In either case,

* Corresponding author. E-mail address: meyks@usm.my

English language proofreader: Si-Yu Lin

determining an appropriate modeling method depends on the complexity and severity of the IIH component. In the EBN context, its overall constituents and optical properties can be considered as an IIH component whereby filtering it independently may result in omitting important features in the EBN region. Therefore, modeling IIH in EBN is a complex yet unresolved topic. It is essential to have a deeper understanding of the EBN inhomogeneous properties before applying any further analysis to the image.

As discussed in [5-6], the surge in studies on EBN is notable due to its rarity and widely recognized health benefits. In a more particular area, many researchers have started to delve into resolving the automation of impurity inspection and cleaning for raw EBN. This is driven by the challenges associated with the current manual process, which requires staggering hours for skilled operators to inspect and remove the impurities [7-8], it is also susceptible to inconsistencies and heavily relies on the subjective judgment of human operators [9]. Impurities like feathers, eggshell debris, and droppings are commonly found in raw EBN, which need to be removed [10].

The need to automate this process led to studies such as in [11] that apply color plane extraction and thresholding while the study in [12] employed fuzzy clustering (fuzzy C-mean, also known as K-mean) segmentation to detect the impurities respectively. These methods experienced significant drawbacks in locating the impurities that involved regions with IIH leading to a high misdetection rate. Work by [13-14] extended the impurity detection study by manipulating lighting setups, such as altering wavelength and projection angle to mitigate IIH artifacts. This approach enhanced the contrast ratio but the IIH persisted in corrupting the image and disrupting the detection performance. Additionally, these methods struggle as they rely solely on the pixel intensity values, without accounting for features and semantic information.

Recent studies by [15-16] have employed deep learning models, such as U-net (a type of convolutional neural network) and hybrid autoencoder to extract semantic information from EBN images. In [15], the U-net model is trained in a supervised manner to output binary probability tensors for classifying the EBN and impurity regions. Meanwhile, in [16], an autoencoder stacked with a single convolutional layer is used to interpret residual maps generated by subtracting original annotated images from autoencoder-reconstructed images, which effectively localizes the impurity regions in EBN images. Both methods showcase the superior potential of deep learning in resolving the IIH issue as compared to conventional image processing methods. However, they require meticulously crafted annotated datasets, requiring significant effort to construct, and limiting adaptability to unseen impurity types in the diverse nature of EBN. Additionally, the learned parameters of both models are in a deterministic form whereby it does not have a continuous representation that may cluster related features together.

Hence, employing a deep learning model that can interpret unseen examples and cluster similar features is crucial for modeling the specific distribution of EBN. This topic has become a research gap that is yet to be studied and unravel its potential. One of the prominent generative models which is the variational autoencoder (VAE) has the ability to undertake this task. Works in [17-20] demonstrated the capability of VAE in modeling the image distribution under undesired conditions and successfully handled tasks, including image denoising, low-light compensation, and image restoration. Therefore, modeling the IIH properties of EBN can be an additional prospect for VAE to solve. VAE can be leveraged for this purpose as it can approximate the image feature distribution, embed them in a simpler representation of latent space, and disentangle this space to identify specific features corresponding to impurities in EBN.

Leveraging VAE's ability to model complex distributions, β-VAE stands out as a compelling choice for impurity segmentation in EBN. Introducing a β parameter to the VAE objective function minimally modifies the architecture, enhancing its tunability while maintaining structural simplicity and interpretability. The β parameter controls the trade-off between reconstruction quality and disentanglement of the latent variable, promoting a more interpretable latent space where impurities can be separated from other features. This enables β-VAE to generalize effectively to unseen impurities, addressing the limitations of complex models like U-net [15] and hybrid autoencoders [16]. This makes it an ideal option for automating the

impurity segmentation process, offering both flexibility in learning from diverse impurity types and robustness against imaging challenges.

In brief, the challenge of impurity segmentation in EBN images is compounded by IIH, which distorts pixel intensity and hampers segmentation performance. Current methods are struggling to model IIH effectively while relying on labor-intensive annotated datasets and lacking generalization to unseen data. Thus, the purpose of this study is to utilize the β-VAE with a specific latent disentanglement method in an attempt to model EBN's IIH properties and identify the impurity features to segment them from the overall EBN component. To the best of current knowledge, this is the first attempt to model the EBN feature distribution in an unsupervised manner. The contributions of this study can be summarized as follows:

- Propose and implement an unsupervised approach for modeling EBN images that mitigates intensity inhomogeneity, reducing reliance on annotated datasets.
- Introduce an effective disentanglement technique that improves the interpretability of latent space feature distributions, allowing the model to adapt to various impurity types.

Section 2 provides a detailed explanation of the proposed methodology and framework, followed by results and discussion in Section 3. The study concludes with Section 4, summarizing the findings and making recommendations for future works.

2. Background and Proposed Methodology

In this chapter, the background and proposed methodology for impurity segmentation in EBN images are presented. The chapter begins with an overview of the β-VAE, which discusses its core principles and components. The image acquisition process and dataset preparation are then described to ensure the creation of a well-structured dataset for model training. Following this, the development and training procedures of the model are described in detail. The methodology is then focused on modeling EBN and segmenting impurities using the β-VAE. Finally, the performance evaluation metrics used to validate the model's reconstruction capability and segmentation performance are outlined.

2.1. β-Variational autoencoder (β-VAE)

VAE is composed of an encoder-decoder architecture with convolutional feature extractors that learn and embed image features into a simpler representation in its latent space. Unlike traditional autoencoders that deterministically map input data to fixed points in the latent space, VAEs introduce stochasticity in their latent representations [21]. This allows them to capture the underlying data distribution more effectively, leading to a more expressive and flexible latent space with powerful modeling capabilities.

As demonstrated in Eq. (2), the loss function in VAE governs two major components which are the log-likelihood, which controls image reconstruction quality, and the Kullback-Leibler (KL) divergence, which acts as a regularizer to measure the divergence between the posterior prior distributions [22]. Based on the equation, $p(z)$ denotes the prior distribution, $p_\theta(x|z)$ is the probabilistic decoder parametrized by the neural network that generates data x based on the latent variable, z,$q_\varnothing(z|x)$ is the variational posterior which approximates the posterior distribution of the latent variable z given the data x and is parameterized by an encoder network. The training objective of VAE is to maximize L_{VAE} over the training data, which optimizes the log-likelihood and the KL divergence.

$$L_{VAE}(ELBO) = \underbrace{E_{q_{\varnothing}(z|x)}\left[\log p_\theta(x \mid z)\right]}_{Log\ \text{likelihood}} - \underbrace{\beta \cdot D_{KL}\left[q_{\varnothing}(z \mid x) \| p(z)\right]}_{KL\ divergence} \tag{2}$$

Prior distribution $p(z)$, is often referenced with normal Gaussian distribution as in $G(0, 1)$ where it expected the mean (μ) and standard deviation (σ) to be zero and one respectively. Thus, the posterior distribution $q_\varnothing(z|x)$, which can be represented

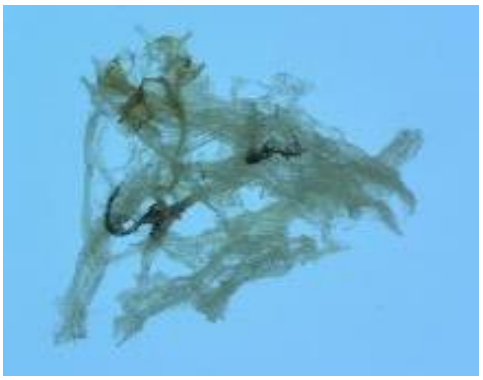
as $G\ (\mu_\theta\ (x),\sigma_\theta^2(x))$, is regularized to be as close as $G\ (0,\ 1)$ by the KL divergence. This regularization clusters related features in the latent space, making it more disentangled and interpretable. However, in the EBN scenario, the complexity of its optical properties poses significant challenges for reconstruction.

While VAEs aim to learn disentangled representations by separating features in the latent space, reconstructing images from such representations becomes challenging when the clusters are widely separated [23]. This can result in excessive noise or smoothening effects which disregards the intricate details of EBN structure. According to [24], applying a smaller weightage, known as β value, to the KL divergence as outlined in Eq. (2) can control this behavior. This adjustment transforms the VAE into what is known as β-VAE. This reduces the emphasis on enforcing a strict prior distribution in the latent space, allowing for more flexibility and potentially capturing a wider range of features during reconstruction. However, this comes with the trade-off of difficulties in disentangling the latent space which can be a challenge in interpreting the learned feature distributions. Thus, this will be addressed through a novel disentanglement approach involving the filtering and penalizing of latent dimensions, as explained in subsequent sections. The latent dimension in this context refers to the compressed representations of individual data points within the latent space, where each element corresponds to a specific latent dimension.

2.2. Image acquisition and dataset preparation

The raw EBN used in this study is made into "chips" with a size of less than 4 mm in thickness and a diameter of approximately 25mm. It is based on the common sizing of the EBN products for packaging [14]. Two categories of EBN are used which are raw clean (RC) which is free from impurities, and raw unclean (RUC) which contains impurities, as shown in Fig. 1.

(a) Raw clean (RC) category which contains no impurities



(b) Raw Unclean (RUC) category which contains multiple impurities

Fig. 1 Chip-sized EBN used in this study

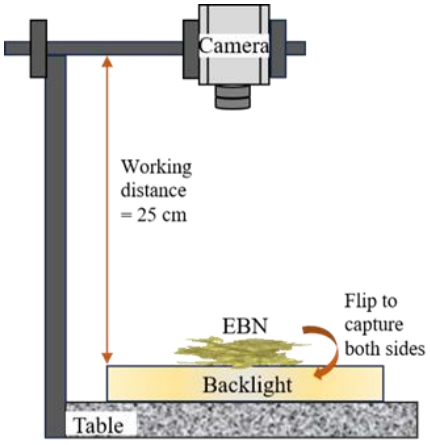


Fig. 2 Image acquisition setup with a fixed working distance of 25 cm

An industrial-grade Basler acA2500-14uc camera with a resolution of 1942 pixels × 2590 pixels is positioned 25 cm above the EBN sample as shown in Fig. 2 for image acquisition, making one pixel equivalent to 0.019 mm × 0.019 mm. This setup suffices to match the minimum observable impurity by human inspectors which is approximately 0.075 mm [13]. Besides laboratory general lighting, backlighting with a parallel angle to the camera is used to enhance the contrast between EBN and the background. A total of 150 EBN samples are utilized, capturing both front and back parts for each sample, resulting in 300 images for the overall dataset. Capturing both sides of the sample is a physical augmentation performed to cater to variations in structure, light penetration, and impurities on both sides. This increases the learning instances and facilitates the deep learning model learning a broader range of spatial information. This will also reduce the dependency on placement orientation during actual inspection.

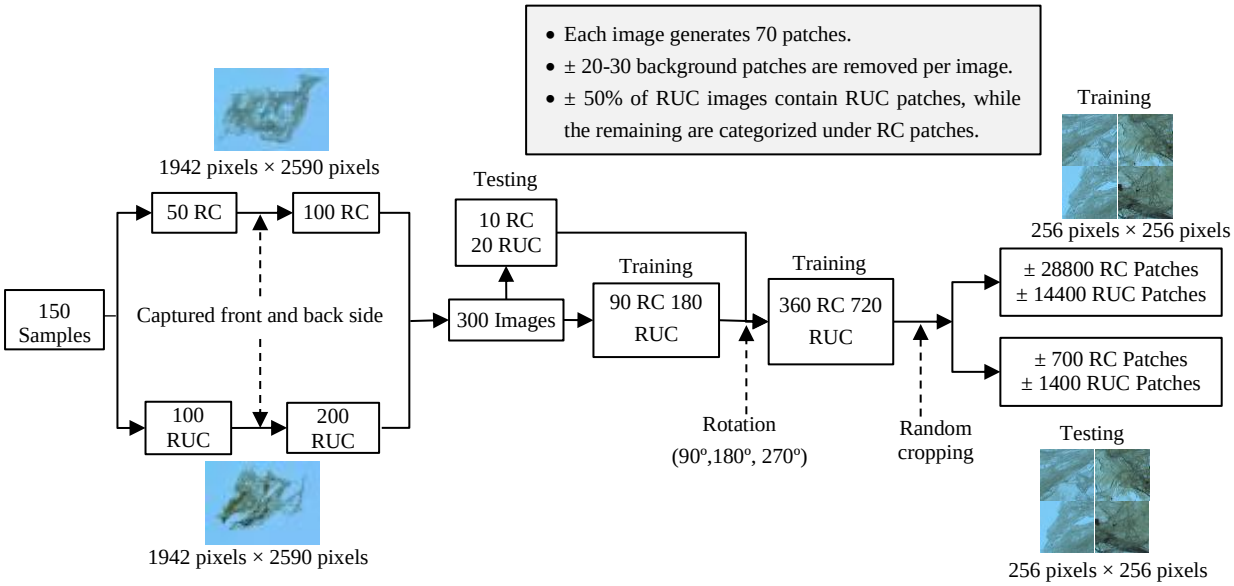


Fig. 3 The augmentation and division of the dataset

The preparation of the entire dataset and an estimated data quantity are illustrated in Fig. 3. The ratio of RC to RUC samples is established at 1:2, signifying 50 samples for RC and 100 samples for RUC. This ratio arrangement is chosen to balance the instances between both categories since most of the EBN regions in RUC images that are free from impurities will be categorized as RC when the image is turned into patches. Even with this ratio, the instances of RC still outnumber RUC, but it is still acceptable as the primary focus is to determine the EBN distribution, which requires more RC EBN instances compared to the impurities. A total of 30 images are reserved as test datasets while the remaining are allocated for training.

The raw training images will undergo further pre-processing, including rotational transformations at 90º, 180º, and 270º angles, making the quantity increase to 1080 samples in total for RC and RUC. This is performed to make the model rotationally invariant. Then, random cropping is applied to both the training and testing datasets to simplify the data loading process for model training and prediction. Eventually, the overall dataset will be in the form of patches with 256 x 256 pixels size. Out of all generated patches, 20 to 30 patches for each image containing only background without any EBN or impurities will be excluded. This results in approximately 28,800 and 14,400 patches for RC and RUC respectively in the training dataset. Meanwhile, the testing sample consists of approximately 700 and 1400 patches for RC and RUC respectively.

2.3. Model development and training

In this study, a computer with Intel® Xeon® W-1370P at 3.60GHz, 8 Core(s) CPU, 16 GB NVIDIA RTX A4000 GPU, and 32 GB DDR4 RAM module is utilized to develop the deep learning model. The VAE model architecture, shown in Fig. 4, is configured based on the results of the hyperparameter tuning which are explained later in this section. The model is employed throughout this study to learn the EBN and impurity features. In brief, the model architecture consists of three main components

which are the encoder, latent space, and decoder. The encoder consists of a sequence of convolutional (Conv) layers that progressively reduce the spatial dimensions of the input while extracting its feature representations. It maps the input data to a compact representation in the latent space, which is parameterized by two convolutional layers: one for the mean (μ) and another for the standard deviation (σ) of the latent variables. The latent space represents compressed data formed by sampling latent variables using the reparameterization trick, $z = \mu + \sigma \cdot \epsilon$, where ϵ is random noise sampled from a standard normal distribution. This step enforces the probabilistic nature of latent space. The decoder, in turn, reconstructs the input data from this latent representation using transposed convolutional (TConv) layers and upsampling operations. This division of the model into encoder, latent space, and decoder components aligns with the general structure of VAE, where the latent space serves as the probabilistic bottleneck, enforcing a distributional prior during training. Based on the model architecture, $\text{Conv}_{a,b,c}$ and $\text{TConv}_{a,b,c}$ represent the convolutional and the transposed convolutional layers, respectively. In this context, "a" represents the number of input channels, "b" denotes the number of channels after convolution, and "c" specifies the dimensions of the c×c (width×height) of the convolution kernel, which determines the spatial extent of the filter applied to the input.

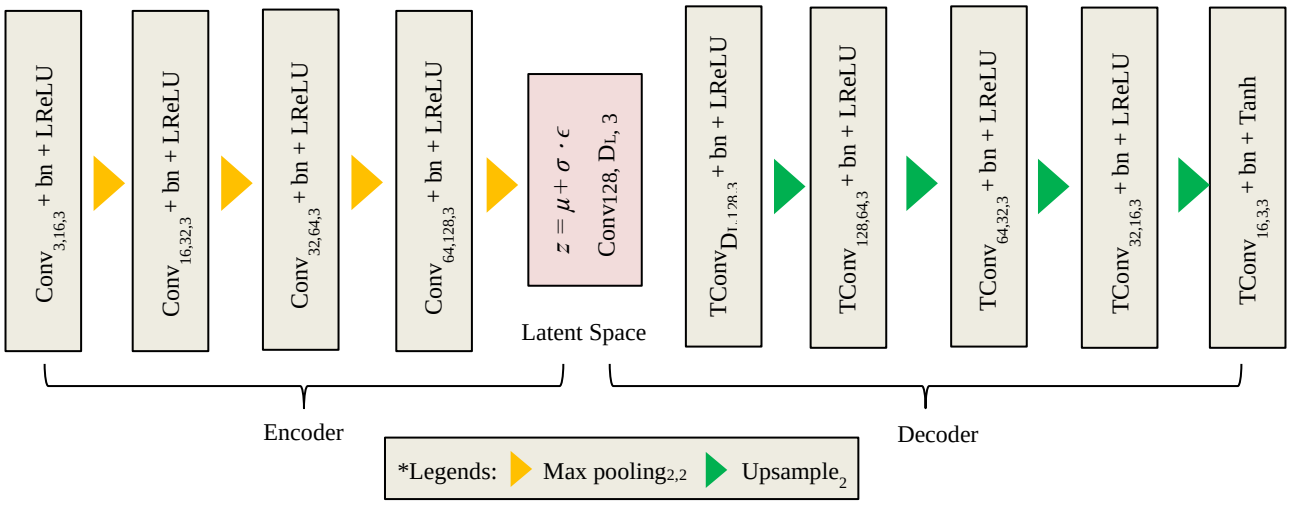


Fig. 4 VAE model architecture used in this study

Batch normalization is represented as bn, while the leaky rectified linear unit (LReLU) and hyperbolic tangent activation functions are denoted as LReLU and TanH respectively. The "Legends" in Figure 4 refer to the graphical symbols used to represent different operations in the model architecture: the yellow triangle indicates $\text{Max pooling}_{2,2}$, which is a downsampling operation that reduces the spatial dimensions of feature maps by taking the maximum value within each 2×2 window, with a stride of 2 to ensure non-overlapping regions. The green triangle represents Upsample_2, which is an upsampling operation that increases the spatial dimensions of feature maps using bicubic interpolation, scaling them by a factor of 2 to restore the resolution. The latent dimension (D_L) value will be determined based on the hyperparameter study. Throughout the training, early stopping is applied with the stopping criterion set to 10% of total epochs to avoid overfitting. Other general hyperparameters used are listed in Table 1.

Table 1 General hyperparameter used throughout the study

Hyperparameter	Value
Batch size	16
Max epochs	100
Initial learning rate	$1e^{-4}$
Optimizer	Adam
Weight decay	$1e^{-4}$

To further optimize the model architecture, the hyperparameter tuning is conducted to examine a certain range of hyperparameter values namely β value, latent dimension, and layer depth. Table 2 demonstrates the range for these values which is experimentally examined to obtain the best combination. The best hyperparameter combination is evaluated based on its effect on the model in minimizing time and computational resources, optimizing VAE loss, and the ability to preserve the reconstruction quality. These factors are essential for selecting an effective feature extractor model.

Table 2 Range values and variations for hyperparameter tuning

Layer Depth	Latent Dimension, D_L	β Value
5 Layers	5, 10, 20, 30	3.0, 1.0, 0.1, 0.01, 0.001, 0.0001
4 Layers	20	0.001
6 Layers	20	0.001

As explained in Section 2.1, the β value in VAE signifies the weightage of KL divergence from the overall VAE equation. A higher value forces the model to create a more disentangled distribution, improving representation [25]. However, this comes with the trade-off of the reconstruction quality, where lower values are preferable. The latent dimension value affects the capacity of latent space. Higher values allow for more learned features to be stored but it may reduce clarity when it is too large, making feature clustering less efficient [26]. This may cause the range of learned distribution to become wider and more difficult to modulate specific features during reconstruction. Layer depth influences the learning capability. A deeper network captures more underlying information contained in the image. However, it may consume significant computational resources and time and is more difficult to optimize [27]. Thus, the hyperparameter tuning is used to evaluate all these hyperparameter effects on the model and determine the optimal combination.

2.4. EBN modeling and impurity segmentation

After training and identifying the optimal hyperparameter, the model's latent space comprising mean (μ) and standard deviation (σ) values in each latent dimension is flattened into a 1-D tensor to impose thresholds on each of the individual dimensions in the latent space. The threshold is initially set at approximately two-thirds of the highest mean value and is experimentally adjusted to optimize reconstruction quality and segmentation performance.

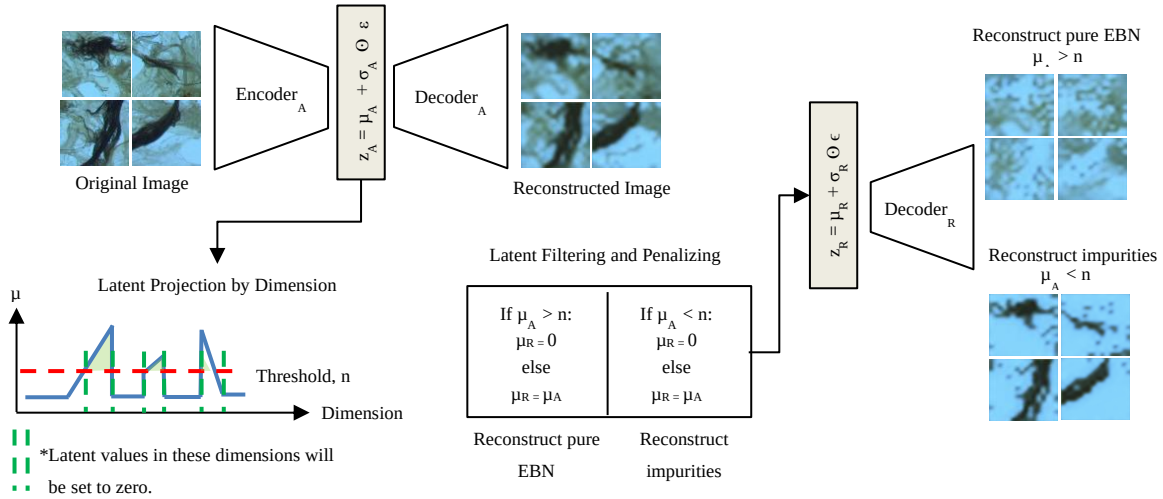


Fig. 5 Proposed disentanglement learning which utilized threshold filtering of individual dimension in β-VAE latent space

The two-thirds value follows the empirical rule which classifies normal data within the range of the overall distribution thus the abnormal data or anomalies lie further from this region [28]. It also aligns with the nature of VAE which follows the normal distribution which forces the mean value in each dimension to be close to zero [25]. Thus, the dimensions with higher values especially exceeding two-thirds of overall latent values, are considered anomalies. Therefore, this threshold is set to

categorize latent dimensions into two groups: one with mean values exceeding the threshold and another with values below it. To obtain a clean EBN without impurities, a binary condition is established. Dimensions with mean values exceeding the threshold are penalized and set to 0, while those with values lower than the threshold retain their original values. The process is reversed if the goal is to exclusively reconstruct impurities, as illustrated in Fig. 5.

Meanwhile, Eq. (3) below illustrates how the mean value for each dimension is checked with the threshold value *n*, which is made as a condition to determine the resultant mean value, μ_R that will be used to reconstruct impurities only:

$$\begin{cases} \mu_R = 0, \mu_A < n \\ \mu_R = \mu_A, \mu_A \geq n \end{cases} \tag{3}$$

A suitable threshold value is needed because setting it too low the model may fail to suppress some of the EBN regions during reconstruction. However, if the value is too high, impurities may not be able to be reconstructed and localized. Therefore, a range of thresholds is tested at 0.5 intervals (both increasing and decreasing) from the initial value to identify the optimal threshold based on performance evaluation metrics. Once the threshold is determined, the values in the resulting latent dimension, Z_R, are obtained with a targeted distribution representing the impurity features. The resultant decoder, $Decoder_{R,}$ utilizes the filtered latent space to reconstruct and locate impurities. The reconstructed impurities then undergo adaptive thresholding with a window size of 11 to turn it into a binary mask before overlaying it with the ground truth to assess detection performance.

2.5. Performance evaluation

The performance measurement for this study consists of both qualitative and quantitative evaluations. During training, qualitative evaluation assesses the visual quality of the image reconstruction to identify the best hyperparameter combination. It is complemented with the value of VAE loss to support the hyperparameter tuning process. Meanwhile, for testing, quantitative evaluation employs several metrics such as IoU and pixel-based classification derived from the confusion matrix are used to examine the impurities segmentation performance. The impurities reconstructed by the $Decoder_R$ are overlaid with the ground truth to identify intersecting, missed, or falsely detected regions. Fig. 6 shows how the IoU is computed for this study.

Comparing reconstructed impurities with the ground truth determines the confusion matrix components. Based on this, the true positive values (TP) indicate the model predicting the correct impurities, the false positives (FP) specify the model misclassifies the false impurities in its detection and the false negatives (FN) point out the impurities have not been detected.

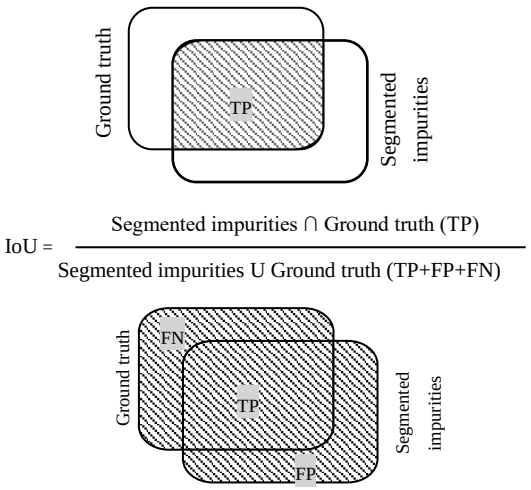


Fig. 6 IoU metric derivation. It evaluates the intersection of predicted impurities with ground truth

by the model. The true negative value (TN) is not used in this study as no specific false impurities are annotated. The precision, recall, F1 score, and misclassification rate can be computed from the confusion matrix, as presented below

$$Precision = \frac{TP}{TP+FP} \tag{4}$$

$$Recall = \frac{TP}{TP+FN} \tag{5}$$

$$F1-Score = 2\frac{Precision \times Recall}{Precision+Recall} \tag{6}$$

$$Misclassification\ rate = \frac{FP+FN}{TP+FP+FN} \tag{7}$$

3. Results and Discussion

This chapter presents the results and discussion of the impurity segmentation in EBN images. Section 3.1 details the process of hyperparameter tuning for the VAE architecture, examining the configurations that optimize model performance. The segmentation results for EBN impurities are then presented in Section 3.2, discussing the model performance in modeling the IIH and segmenting the impurities.

3.1. Hyperparameter tuning for VAE architecture

The hyperparameter combinations outlined in Table 2 of Section 2.3 are evaluated to see how they affect the model performance, especially the visual quality of reconstructed images and VAE loss value. The β value played a major role in tuning the image reconstruction quality. As depicted in Fig. 7, when the β value is set to 3.0, the image becomes smoothened and reconstructs only the background because the dominance of KL divergence suppresses the impact of the reconstruction loss in comparing the reconstructed image with the original image. As a result, the latent space becomes smoothly distributed and disentangled but possesses minimal capacity for interrelation among latent dimensions to be sampled, thus rendering it to reconstruct a meaningful image.

As the β value decreases, the reconstruction quality improves, preserving more image details. A significant change in reconstruction quality from β equals 1.0 until 0.01 is observed while the difference is getting discrete between 0.001 to 0.0001. Moreover, the VAE loss as summarised in Table 3 shows a slight difference, decreasing from 0.032 at β = 0.001 to 0.028 at β = 0.0001. Thus, a β value of 0.001 is chosen to obtain a high reconstruction quality and at the same time, the latent disentanglement effect is still being preserved.

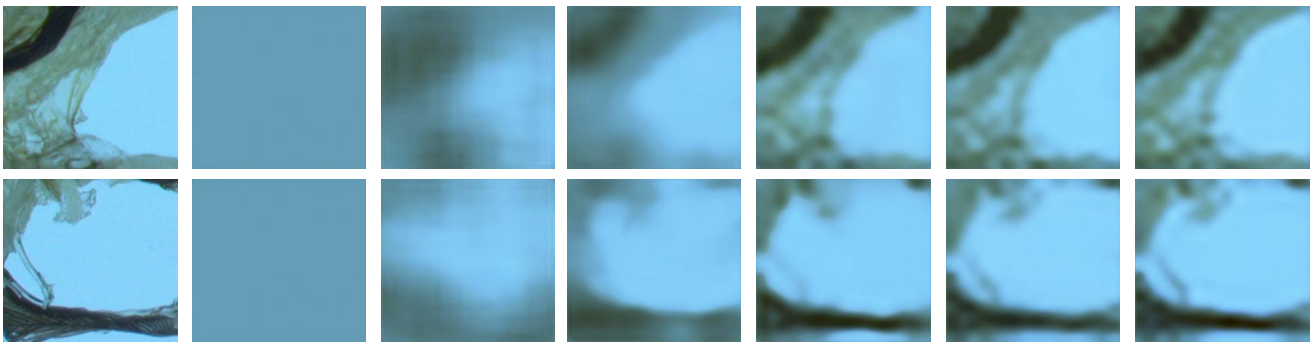


Fig. 7 Different reconstructed image based on β values

Table 3 VAE loss values in varying latent dimensions and β values

Latent dimension, D_L	β =3.0	β =1.0	β =0.1	β =0.01	β =0.001	β =0.0001
5	0.370	0.250	0.135	0.074	0.041	0.030
10	0.370	0.216	0.113	0.061	0.035	0.029
20	0.370	0.179	0.095	0.051	0.032	0.028
30	0.370	0.160	0.085	0.045	0.031	0.027

In terms of latent dimension, the changes in value which indicate the overall size of latent space, have minimal effect on reconstruction details except that higher dimensions exhibit better color contrast as presented in Fig. 8. A latent dimension of 20 already shows the optimal details, sufficient to distinguish the contrast between EBN and impurities. The VAE loss also shows a minimal difference of only 0.001 between latent dimensions of 20 and 30. Thus, a latent dimension with a size of 20 is chosen to avoid the complexity of interpreting higher latent dimensions.

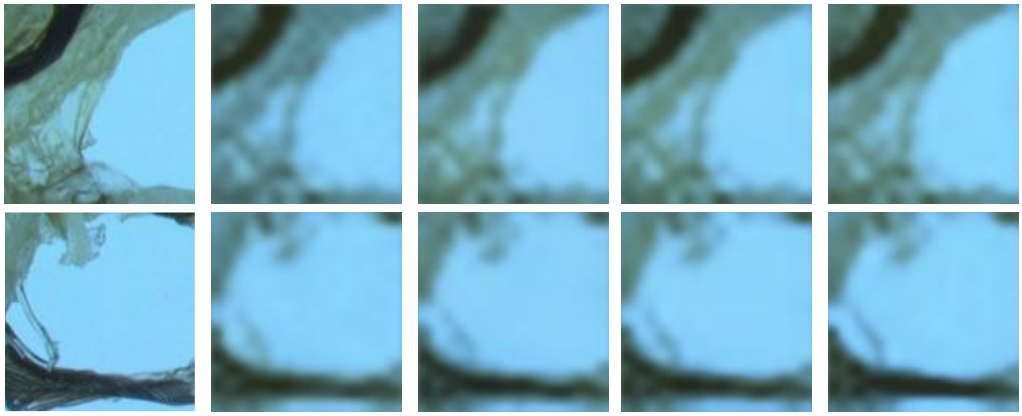


Fig. 8 Different reconstructed image based on latent dimension values

The layer depth effect does not contribute significant changes in model performance, especially the reconstruction quality and the loss value. As shown in Fig. 9, the reconstruction quality is relatively similar and the VAE loss as in Table 4 indicates a minor difference between all of the layer configurations. The model with four convolutional layers achieves a slightly better image reconstruction quality and a lower VAE loss value than the five convolutional layers model. Meanwhile, the model with six convolutional layers obtains the lowest VAE loss and shows the most refined reconstruction image in terms of contrast and details. However, the training time is greatly affected by the changes in layer configuration. The model which consists of five convolutional layers required approximately 3.8 hours to finish training, while the models with four convolutional layers and six convolutional layers took around 3.6 hours and 12 hours to finish training, respectively.

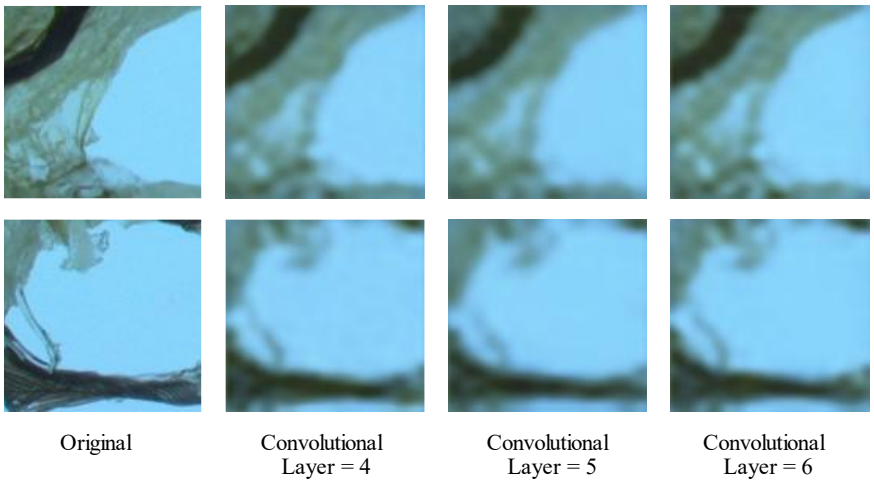


Fig.9 Different reconstructed images based on the depth of the convolutional layer

Table 4 VAE loss values and training time in different convolutional layers depth

Convolutional layers	VAE loss	Training time
4	0.030	3.62 hours
5	0.032	3.83 hours
6	0.029	12.08 hours

With minimal differences between each other in terms of reconstruction quality and VAE loss, the four-convolutional-layers configuration is the optimal choice as it requires less time and computational resources. The VAE model is finalized using these hyperparameters.

3.2. EBN impurity segmentation result

To examine the impurity segmentation performance, the testing datasets that are not presented in the training and validation are used. The reconstructed impurities are compared with the ground truth to calculate several metrics including IoU, precision, recall, F1 score, and misclassification rate. This indicates how effectively the model reconstructs and pinpoints impurities based on the learned features. A range of threshold values is examined to analyze their impact on segmentation performance. The overall result is summarized in Table 5.

Table 5 Overall result of the impurities segmentation based on different threshold values, *n*

Threshold, *n*	IoU	Precision	Recall	F1 score	Misclassification rate
1.5	46.65	47.32	97.86	62.97	53.35
2.0	68.05	73.84	90.08	80.64	31.95
2.5	73.08	85.64	83.22	84.01	26.92
3.0	62.62	94.14	65.75	76.16	37.38
3.5	28.71	95.72	29.71	41.88	71.29

In summary, the model attains its optimal performance with a threshold value set to 2.5 (two-thirds of the overall μ value), yielding scores of 73.08% IoU, 85.64% precision, 83.22% recall, 84.01% F1 score, and 26.92% misclassification rate. This model outperforms the highest IoU and F1 scores while exhibiting the lowest misclassification rate. The highest IoU value indicates the impurities segmented by this model have the highest frequency of correct overlapping with the ground truth. Moreover, the model's lowest misclassification rate signifies its minimal tendency to segment the FP or FN impurities. The precision and recall score of this model is not the highest. Higher threshold models achieve stricter filtering, reducing the FP value, while lower threshold value models reconstruct more regions, reducing FN values.

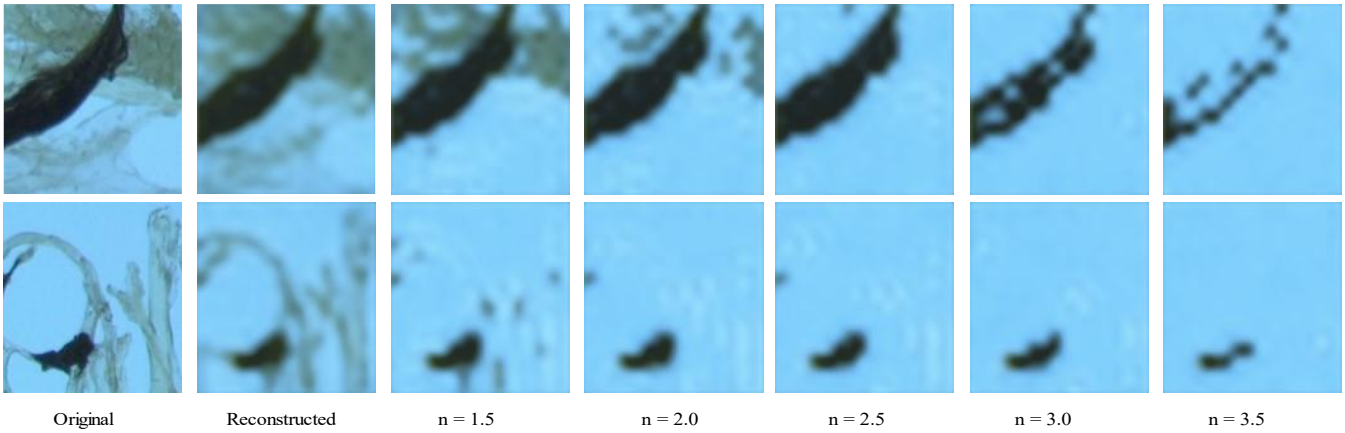


Fig. 10 Simple impurities structure which have a continuous and solid region

Nonetheless, this model strikes a good balance between precision and recall, resulting in the highest F1 score. The F1 score, the harmonic mean of precision and recall balances rigid and loose threshold tolerances. A deeper analysis revealed a few aspects that play a major role in these performance scores. The segmentation performs better when it is tested on a simple impurity type characterized by a continuous or solid region, as depicted in Fig. 10. The threshold value of 2.5 effectively segments the impurity features while suppressing the EBN region, resulting in fewer FP and FN in this impurity type.

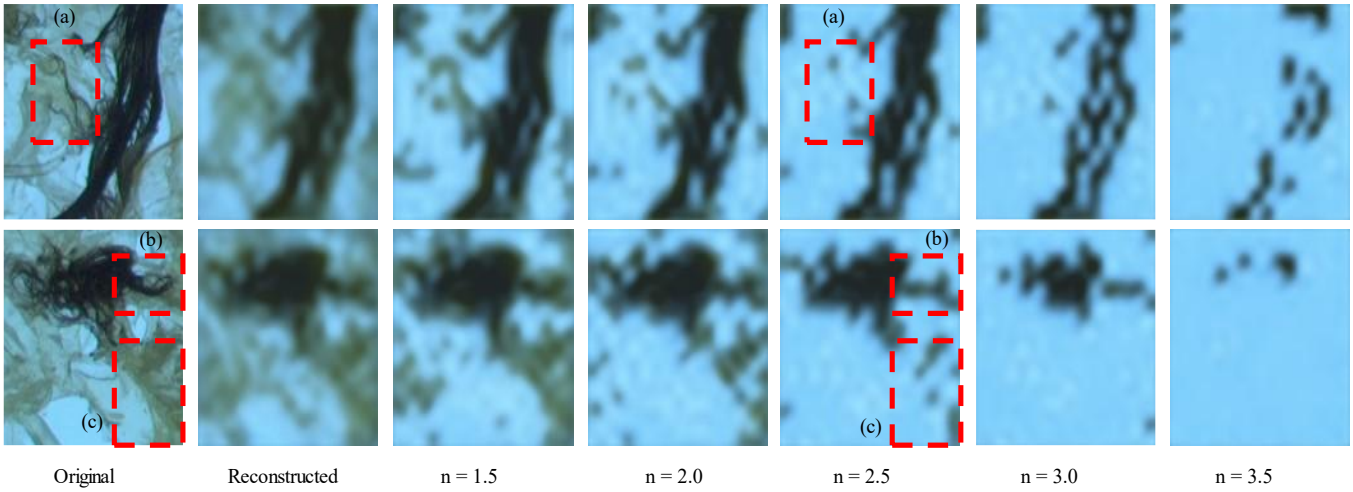


Fig. 11 Complex impurities structure which contained thin and interwoven branches
(a) Thin branches of impurities missed to be reconstructed
(b) The reconstruction of interwoven impurities becomes smoothen
(c) The thick EBN strand failed to be suppressed

Meanwhile, when testing on a complex impurity structure with intricate details such as thin and interwoven branches, as depicted in Fig. 11(a), the segmentation performance significantly dropped as it failed to reconstruct these particular details. This limitation is attributed to the probabilistic nature of VAE latent space, which tends to force the distribution to be normally distributed. This smoothens the image, causing blurring effects in the reconstruction. Moreover, the blurring effects in VAE contribute to an increased occurrence of FP, where more regions are reconstructed as impurities as in Fig. 11(b). Table 6 provides a detailed breakdown of impurity segmentation performance for respective impurity types, revealing higher performance metric scores in simple structure impurities compared to complex ones. The complex structure proved to be a pivotal factor influencing the overall model performance. Using a higher resolution camera could mitigate this issue as smaller details could be captured with a greater pixel representation.

Another crucial factor in impurity segmentation via selective latent filtering is the dimension with borderline feature values. In this context, the feature encompasses values marginally higher or lower than the threshold, influencing either the occurrence of false positives (FP) or false negatives (FN). This scenario commonly involves thick EBN strands, exhibiting intrinsic similarity with the features of impurities, as illustrated in Fig. 11(c).

Despite these drawbacks, the obtained score still demonstrates a noteworthy outcome. As indicated in [29], IoU and F1 score values exceeding 70% are considered satisfactory. This benchmark is surpassed with an overall IoU score of 73.08% and an F1 score of 84.01%. Additionally, when converting IoU to a Dice coefficient, as outlined in [30], the model attains a Dice coefficient of 84.44%, surpassing values reported in [15] of 76.46% ± 2.4% and in [16] at 61.80%. However, comparisons with other metrics, such as F1 score, precision, recall, and misclassification rate, are not directly viable, as existing methods primarily focus on impurity detection, whereas this study concentrates on impurity segmentation. The former calculates scores based on the count of detected impurities, while the latter employs a pixel-wise calculation. Apart from that, the proposed method effectively deals with small β values, particularly in disentangling the VAE latent space. This is crucial for studying complex object composition like EBN, where higher β values are less effective.

Table 6 Result by impurity category which consists of simple and complex types

Impurity type	IoU	Precision	Recall	F1 score	Misclassification rate
Simple	78.26	89.04	86.90	87.54	21.74
Complex	66.17	81.11	78.31	79.30	33.83

4. Conclusion

In this study, the feature distribution of the inhomogeneous properties of EBN and its impurities has been successfully modeled using β-VAE latent space in an unsupervised manner. Selective filtering and penalizing of latent dimensions were successfully developed to distinguish EBN features from their impurities, enabling targeted reconstruction or suppression to highlight pure EBN regions or pinpoint the impurities. From the results, the conclusion can be drawn as follows:

(1) Despite limitations such as blurring effects and smoothened reconstructions inherent in VAEs, this method surpasses the results of existing studies in [23] and [24], demonstrating its effectiveness.

(2) The misclassification rate is mainly caused by the complex-type impurities, which lose fine details from the simplification of the β-VAE reconstruction.

Therefore, future studies should focus on improving the β-VAE reconstruction quality which could have a high potential to improve the segmentation performance. Future improvements and major areas of focus are suggested as follows:

(1) Use images with higher resolution so that the VAE feature extractor captures finer details.

(2) Incorporation of more sophisticated encoding architecture to enable a more refined feature extraction and improve the segmentation performance.

(3) Exploration of a more complex latent disentanglement method to further adapt with several limitations addressed in this study.

(4) Implementation in other imaging domains to test the applicability and effectiveness of the method proposed in this study.

Acknowledgment

This research is primarily supported by the Ministry of Higher Education Malaysia for Fundamental Research Grant Scheme with project reference code: FRGS/1/2021/ICT02/USM/02/2. Appreciation also goes to the National Institute of Information and Communications Technology (NICT) (http://www.nict.go.jp/en/index.html) for supporting the project with GPU powered computer under an asset loan agreement. Lastly, Tian Ma Bird Nest Sdn. Bhd. is acknowledged for generously sharing valuable samples and insights in the EBN industry.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] T. H. Lee, S. Wong, C. H. Lee, N. A. Azmi, M. Darshini, S. Kavita, et al., “Identification of Malaysia’s Edible Bird’s Nest Geographical Origin Using Gel Electrophoresis Analysis,” Chiang Mai University Journal of Natural Sciences, vol. 19, no. 3, pp. 379-392, 2020.

[2] Q. Fan, X. Liu, Y. Wang, D. Xu, and B. Guo, “Recent Advances in Edible Bird's Nests and Edible Bird's Nest Hydrolysates,” Food Science and Technology, vol. 42, article no. 67422, 2022.

[3] A. F. El Sheikha, “Why the Importance of Geo-Origin Tracing of Edible Bird Nests is Arising ?,” Food Research International, vol. 150, no B, article no. 110806, 2021.

[4] V. Venkatesh, N. Sharma, and M. Singh, "Intensity Inhomogeneity Correction of MRI Images Using InhomoNet," Computerized Medical Imaging and Graphics, vol. 84, article no. 101748, 2020.

[5] N. A. S. Ahmad Shuyuti, E. Salami, M. Dahari, H. Arof, and H. Ramiah, "Application of Artificial Intelligence in Particle and Impurity Detection and Removal: A Survey," IEEE Access, vol. 12, pp. 31498-31514, 2024.

[6] K. M. Goh, L. L. Lim, S. Krishnamoorthy, W. K. Lai, T. Maul, and J. K. Chaw, "Recent Advancement of Intelligent-Systems in Edible Birds Nest: A Review from Production to Processing," Multimedia Tools and Applications, vol. 83, pp. 51159-51209, 2023.

[7] C. Acharya and N. Satheesh, "Edible Bird's Nest (EBN): Production, Processing, Food and Medicinal Importance," vol. 4, no. 3, pp. 20-25, 2023.

[8] T. K. Hong, C. F. Choy, and A. O. H. Kait., "Approach to Improve Edible Bird Nest Quality & Establishing Better Bird Nest Cleaning Process Facility through Best Value Approach," Journal for the Advancement of Performance Information and Value, vol. 10, no. 1, pp. 38-50, 2018.

[9] F. S. A. Saad, A. Y. M. Shakaff, A. Zakaria, M. Z. Abdullah, A. H. Adom, and A. A. M. Ezanuddin, "Edible Bird Nest Shape Quality Assessment Using Machine Vision System," Proceedings of 3rd International Conference on Intelligent Systems, Modelling and Simulation, pp. 325-329, 2012.

[10] K. C. Chok, M. G. Ng, K. Y. Ng, R. Y. Koh, Y. L. Tiong, and S. M. Chye, "Edible Bird's Nest: Recent Updates and Industry Insights Based On Laboratory Findings," Frontiers in Pharmacology, vol. 12, article no. 746656, 2021.

[11] Y. Subramaniam, Y. C. Fai, and E. S. L. Ming, "Edible Bird Nest Processing Using Machine Vision and Robotic Arm," Jurnal Teknologi, vol. 72, no. 2, pp. 85-88, 2015.

[12] K. M. Goh, W. K. Lai, P. H. Ting, K. Daniel, and J. K. R. Wong, "Size Characterisation of Edible Bird Nest Impurities: A Preliminary Study," Procedia Computer Science, vol. 112, pp. 1072-1081, 2017.

[13] C. K. Yee, Y. H. Yeo, L. H. Cheng, and K. S. Yen, "Impurities Detection in Edible Bird's Nest Using Optical Segmentation and Image Fusion," Machine Vision and Applications, vol. 31, article no. 68, 2020.

[14] K. L. Gwee, L. H. Cheng, and K. S. Yen, "Optimization of Lighting Parameters to Improve Visibility of Impurities in Edible Bird's Nest," Journal of Electronic Imaging, vol. 28, no.2, article no. 023014, 2019.

[15] Y. H. Yeo and K. S. Yen, "Impurities Detection in Intensity Inhomogeneous Edible Bird's Nest (EBN) Using a U-Net Deep Learning Model," International Journal of Engineering and Technology Innovation, vol. 11, no. 2, pp. 135-145, 2021.

[16] Y. H. Yeo and K. S. Yen, "Development of a Hybrid Autoencoder Model for Automated Edible Bird's Nest Impurities Inspection," Journal of Electronic Imaging, vol. 31, no 5, article no. 051603, 2022.

[17] F. Ulger, S. E. Yuksel, and A. Yilmaz, "Anomaly Detection for Solder Joints Using β-VAE," IEEE Transactions on Components, Packaging and Manufacturing Technology, vol. 11, no. 12, pp. 2214-2221, 2021.

[18] N. Ferreira and M. Silveira, "Ship Detection in SAR Images Using Convolutional Variational Autoencoders," Proceeding of IEEE International Geoscience and Remote Sensing Symposium, pp. 2503-2506, 2020.

[19] N. Kozamernik and D. Bracun, "Visual Inspection System for Anomaly Detection on KTL Coatings Using Variational Autoencoders," Procedia CIRP, vol. 93, pp. 1558-1563, 2020.

[20] P. Ma and H. Kuang, "Old Photos Restoration by Using VAE," Proceedings of 2nd International Conference on Science Education and Art Appreciation, vol. 174, article no. 02001, 2023.

[21] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," https://doi.org/10.48550/arXiv.1312.6114, 2022.

[22] E. Dupont, "Learning Disentangled Joint Continuous and Discrete Representations," Proceedings of 32nd International Conference on Neural Information Processing Systems, pp. 708-718, 2018.

[23] C. P. Burgess, I. Higgins, A. Pal, L. Matthey, N. Watters, G. Desjardins, et al., "Understanding Disentangling in β-VAE," https://arxiv.org/abs/1804.03599, 2018.

[24] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, et al., "β-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework," International Conference on Learning Representations, pp. 1-22, 2017.

[25] A. Asperti and M. Trentin, "Balancing Reconstruction Error and Kullback-Leibler Divergence in Variational Autoencoders," IEEE Access, vol. 8, pp. 199440-199448, 2020.

[26] L. Zhou, W. Deng, and X. Wu, "Unsupervised Anomaly Localization Using VAE and beta-VAE," https://arxiv.org/pdf/2005.10686, 2020.

[27] L. Cai, H. Gao, and S. Ji, "Multi-Stage Variational Auto-Encoders for Coarse-to-Fine Image Generation," Proceedings of SIAM International Conference on Data Mining, pp. 630-638, 2019.

[28] N. Asad, "Normal Distribution and Empirical rule," https://doi.org/10.13140/RG.2.2.17900.51841, 2019.

[29] B. Guindon and Y. Zhang, "Application of the Dice Coefficient to Accuracy Assessment of Object-Based Image Classification," Canadian Journal of Remote Sensing, vol. 43, no. 1, pp. 48-61, 2017.

[30] D. Duque-Arturo, S. Velasco-Forero, J. E. Deschaud, F. Goulette, A. Serna, E. Decencière, et al., "On Power Jaccard Losses for Semantic Segmentation," Proceedings of 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP), vol. 5, pp. 561-568, 2021.