

Semantic Segmentation of Tuberculosis Bacilli from Microscopic Sputum Smear Images Using TransUNet

Ashutosh Satapathy*

Department of Computer Science and Engineering, Siddhartha Academy of Higher Education, Deemed to be University, Vijayawada-520007, Andhra Pradesh, India

Email: ashutosh@vrsiddhartha.ac.in

ORCID iD: <https://orcid.org/0000-0002-2531-4079>

*Corresponding Author

Praneeth Vallabhaneni

Department of Computer Science and Engineering, Siddhartha Academy of Higher Education, Deemed to be University, Vijayawada-520007, Andhra Pradesh, India

Email: praneeth9204@gmail.com

ORCID iD: <https://orcid.org/0009-0005-6578-1201>

Manisha Indugula

Department of Computer Science and Engineering, Siddhartha Academy of Higher Education, Deemed to be University, Vijayawada-520007, Andhra Pradesh, India

Email: manishaindugula2k04@gmail.com

ORCID iD: <https://orcid.org/0009-0004-9591-9817>

Received: 05 November, 2024; Revised: 27 April, 2025; Accepted: 20 July, 2025; Published: 08 December, 2025

Abstract: According to the World Health Organization (WHO) touchstones of 2022, Tuberculosis is the second dominant disease after COVID-19. Around one-fourth of the comprehensive population is ascertained to have tuberculosis. Timely detection and prevention of tuberculosis is a must to overcome its harmful effects. The method most often used in ascertaining whether a patient has tuberculosis, is examining his or her sputum sample. In the process, the isolation of the bacilli is done manually, and hence it is prone to error. Segmentation illustrates and enlightens objects or particles within an image, thus extracting the Region of Interest (ROI). The contemplated study uses TransUNet architecture to segment tuberculosis bacilli from sputum images to increase diagnostic accuracy and performance. The attention mechanism used in the TransUNet model helps to identify the spatial hierarchies present in image. It is an extremely tough task for naive or traditional segmentation algorithms to deal with the inherent complexity of sputum images. Hence, this study introduces an approach to capture the intrinsic features and dependencies needed to segment mycobacterium or TB bacilli by leveraging the TransUNet model. The model achieved an average Dice Score of 92.795%, a mean Intersection over Union (IoU) of 88.845%, and a segmentation accuracy of 99.19% on the Mosaic and Ziehl-Neelsen datasets. These results surpassed several existing state-of-art methods like UNet, clustering and thresholding, depicting the superior capability of TransUNet in segmenting the TB bacilli. It deepens the potentiality of transformer-based CNN models, especially TransUNet, for improving the diagnosis of tuberculosis and supporting disease management.

Index Terms: Semantic Segmentation, Tuberculosis Bacilli, Sputum Smear Images, Attention Mechanism, Transunet

1. Introduction

Tuberculosis (TB) is a cosmopolitan health issue that is caused by Mycobacterium which is an aerobic and immobile bacillus [1]. Around 1.5 million people die because of TB every year due to lack of proper treatment and bad disease management [2]. TB spread's via air and mainly affects the lungs and probably other parts of the body and may lead to chronic diseases. The symptoms of TB involve deep-rooted cough, mucus with blood, fever, and many more.

The continuous cough produces sputum, a mucus like substance mixed with blood and is present in the trachea. Considerably, there are three different arenas in TB which include primary infection, Latent TB infection, and Active TB infection [3]. The form of TB that damages the other organs of the body, not just the lungs, is referred to as extrapulmonary TB. It may also cause Meningitis, Sterile pyuria, Pott's disease and hepatitis [4, 5]. The conventional technique used to diagnose TB is Sputum Smear Microscopy (SSM), a technique that is quite cheaper than chest X-rays. SSM is used to prove the existence of TB causing bacilli, which appears in red and rod-shaped structures. Usually, the sputum sample is collected from the patient, and a fast acid stain like Ziehl-Neelsen (ZN) stain or kinyoun stain is applied to produce the sputum smear sample. A typical sputum smear sample can be used to determine the density and intensity of TB spread in a patient and helps in better disease management. The application of Image segmentation is increasing day to day in the field of medical research and disease diagnosis. Image Segmentation is a process of partitioning the image into different portions that depict objects in an image. In segmentation, every pixel of an image belongs to a different class and hence, it is a process of classifying every image pixel into its respective class label [6]. There are three types of segmentations: semantic, instance, and panoptic. Semantic segmentation is a process in which we partition the image into foreground and background. So, here, we have only two class labels, i.e., foreground and background. The objective of semantic segmentation is to isolate or separate the foreground objects from the background objects. In analyzing the sputum smear sample shown in Fig. 1, the bacilli portions are considered in the foreground, and the rest of the image is the background which consists of mucus and other debris [7].

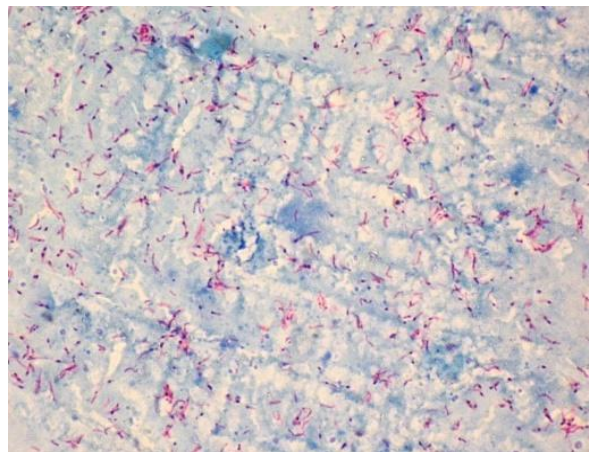


Fig. 1. Sample sputum smear image from ZN dataset.

Transformers are typically used for analyzing sequential data and involve patching, linear projection, patch embedding and attention computation in order to find out the relevance between the data patches. They are well-versed in capturing long-range credence in data using self-attention mechanism which enables them to differentiate with the other data [8]. Although transformers can capture long-range dependencies, they lack localization capability due to scarce low-level details. On the other hand, CNN that is applied to the image and video data, contains multiple convolutional layers and pooling layers that helps them to captures the low-level details or features from the images and isolates the relationship between the different objects in an image [9]. However, this also has its disadvantage which involves failure in capturing the long-range credence. Putting together the Transformers, which are good at collecting the long-range dependencies and CNN models, that are good at capturing the local dependencies, can lead to a state-of-the-art performance in semantic segmentation when compared to the strategies that are solely dependent on CNN models for segmentation. TransUNet is the first constructed state-of-the-art model in the combination of transformers and CNNs for semantic segmentation [10]. That combination ties together the strengths of UNet and Transformers by making semantic segmentation more accurate for medical image analysis. Therefore, this proposed work uses TransUNet to isolate TB bacilli from sputum smear images. As part of this, the work utilizes ZN stained sputum smear images, Mosaic sputum images and involves creating a bacilli segmentation dataset where every image in the dataset consists of a corresponding semantic mask image for analysis. The work addresses the need for TB bacilli segmentation and the research gaps between the proposed work and the existing works. Finally, we collate the performance of various existing models with our work in terms of segmentation accuracy, Dice score, Mean IoU, Recall and Precision, respectively.

2. Literature Review

There has been a growing interest in TB analysis, driven by the need for accurate treatment methods over the past few decades. Various methodologies have been boosted up providing innovative solutions for diagnosis of TB and its timely medication. This is a literature review aimed at providing a set of works on analyzing the sputum smear image to isolate the TB bacilli using image processing techniques and AI models. This review also addresses the advantages and

the potential drawbacks present in the methods that may affect the accurate isolation of the TB bacilli. Owais et al. made semantic segmentation using the modified U-Net model on the BraTs, Heart MRI dataset, and ZN dataset in the year 2022 [11]. Model parameters were also reduced from 30 million to 0.49 million. The evaluation of model's performance in segmentation was done based on the Dice Score (DSC). The DSC obtained was 0.94 using the BraTs and the Heart MRI datasets, while it was 0.74 with the ZN dataset. Wulandari et al. presented an intelligent intuitionistic fuzzy logic-based automatic Acid-Fast Bacillus (AFB) count detection system with self-thresholding techniques in the year 2022 [12]. The paper mainly focused on Y and V-shaped colonies of AFB present in the sputum sample. This technique of segmentation consisted of patch preparation, mask generation and finally, training the U-Net architecture for automating the segmentation task. The maximum accuracy obtained by the propounded scheme in the segmentation of AFB was 94.78%.

Mithra and Emmanuel developed an automated system for diagnosing TB using the Gaussian Mixture Model (GMM), Fuzzy logic, and Neural networks in the year 2021 [13]. The authors addressed limitations, such as low sensitivity and time complexity, in laboratory procedures for diagnosing TB. The authors combined the GMM with fuzzy logic and neural networks to detect TB in a sputum sample. Color Space (CS) transformation was used along with thresholding and the performance of method was judged based on the accuracy of the segmentation. The model achieved a segmentation accuracy of 0.91379 for the ZN TB dataset. Fernández et al. made a work that used the U-Net model for the task of semantic segmentation of less free microscopic sputum images in the year 2020 [14]. The dataset consists of images of 3940 x 2700 pixels in size. These images are divided into patches of 256 x 256 pixels in size. The patches, along with their respective masks, were used for training the U-Net model. Once the prediction process was done, the total image is reconstructed. Mean Intersection over Union (IoU) and segmentation accuracy are used to check the model's performance. The performance metrics obtained were 88% of mean IoU and 92.01% of segmentation accuracy.

Panicker et al. introduced a method that involved the detection of TB using image binarization and classification of TB bacilli using CNN in the year 2018 [15]. The work initially included performing segmentation using image binarization, an unsupervised method. The foreground consisted of simple, touch bacilli and other artifacts. The segmented output was given to the trained CNN architecture to classify the bacilli type, i.e., observed. The performance was analyzed using precision, recall, and F1-score. The model achieved a precision of 78.4%, 97.13% of recall, and 86.76% of F1-score. Mithra and Emmanuel proposed a work for diagnosing TB via segmentation and classification in 2017 [16]. C-Means clustering algorithm supported by fuzzy logic on images of ZN TB was applied for TB bacilli segmentation. After this segmentation, features that represent geometry and intensity were extracted in order to carry out the classification. Ability of the segmentation was tested in terms of precision and the segmentation accuracy obtained for the propounded strategy was 95.1%. Shwetha et al. performed a work that detected TB Bacilli in the ZN TB image dataset in the year 2017 [17]. The work involved utilizing a CNN-based model called squeeze net, and its performance was collated with VGG16 and ResNet50. The author observed that the model performed well compared to the above two models. The propounded work utilized K-Means clustering to perform segmentation for isolating the bacilli with a segmentation accuracy of 97%.

Mithra and Emmanuel introduced a model that performs semantic segmentation on sputum smear images to identify bacilli present in them in 2018 [18]. The data pre-processing step involves noise reduction and intensity normalization to enhance the quality of the input images, ensuring that irrelevant variations are minimized and the significant features of the data are preserved. The propounded model uses a channel area thresholding algorithm for performing segmentation. The author utilized the images of ZN Sputum smears for the research work and achieved a segmentation accuracy of 97.58%. Ayas and Ekinici from Karadeniz Technical University, Turkey, proposed a color information-based segmentation technique for samples stained with ZN using pixel values and their neighboring pixels in the year 2014 [19]. The performance of multidimensional Gaussian probability density functions and support vector machines (SVMs) for segmenting TB bacteria was compared. The system achieved a segmentation accuracy of 99.55% and a recall (sensitivity) of 87.02%. Xavier et al. in the year 2012, introduced a segmentation method for detecting TB bacillus in SSM by selecting scalar features from color spaces like RGB, HIS and YCbCr [20]. The features are extracted through various components via component subtraction from the color spaces, which are then given to the feedforward neural network to classify the pixels as bacilli and background pixels respectively. Geometrical features, notably eccentricity and a new color ratio, were utilized as filters against noise. The model achieved a segmentation accuracy of 91.45%, with a recall of 93.41%.

Vanitha et al., from Sri Ramachandra Institute of Higher Education & Research, Porur, Chennai, India, compared supervised and unsupervised segmentation methods for automated TB bacilli segmentation in sputum smear images by overcoming image variability caused by distinct staining procedures and types of microscopes through applying image pre-processing for the standardization of RGB components in the year 2021 [21]. Among the above techniques tested, the U-Net segmentation method has been found to surpass others. The segmentation accuracy attained was also very high at 0.99. Earlier research primarily utilized unsupervised methodologies for segmentation, which frequently exhibit deficiencies in rigorous ground truth analysis and evaluation, consequently constraining their efficacy. Furthermore, the majority of models underwent training and evaluation on identical datasets, resulting in performance that is contingent upon specific datasets and prompting apprehensions regarding their applicability to novel datasets. Traditional machine learning approaches applied to the segmentation process of these images are not good enough since such methods do

not appropriately capture any hierarchical connections or interdependencies for objects like clusters of bacilli, mucus, cell debris, and other artifacts. Moreover, most CNN-based architectures suffer from semantic gap: the low-level attributes of images, namely shape, intensity, color, and texture, do not map satisfactorily onto the high-level semantic concepts. Such a mismatch leads to poor segmentation and generalization capabilities. This limitation points toward the necessity for more advanced models that are capable of adequately accounting for intrinsic complexities and variability inherent in medical image data.

3. Methodology

3.1. Proposed System Architecture

The segmentation of TB bacilli from microscopic sputum smear images includes a set of several stages as shown in Fig. 2. The process begins with data acquisition, where mosaic sputum images from a sputum images database are resized to a uniform resolution of $A \times A$ pixels, where A is 400 pixels. The sputum images database consists of images and their corresponding masks from ZN and Mosaic datasets. Images are first exposed to a noising removal technique in the form of Gaussian filter, and further data augmentation techniques like left, right and rotation flips are applied in order to increase the diversity of the dataset. After that, it further splits the images into train, test, and valid images. The splitting is done in such a manner that 60% constitutes the train set, 20% is classified as the test set, and 20% for the valid set of the image. The methodology uses the TransUNet architecture, a hybrid UNet model that consists of ViT block in its encoder, for segmentation of the TB bacilli or mycobacterium in the sputum sample. The TransUNet model i.e. trained with the Mosaic data is further tested for its generalizability using ZN dataset. The images of ZN dataset are at a resolution of $W \times Y$ pixels, where W is 2112 pixels and Y is 2816 pixels, respectively. The true masks for these images were prepared using MATLAB by manually annotating the regions of interests.

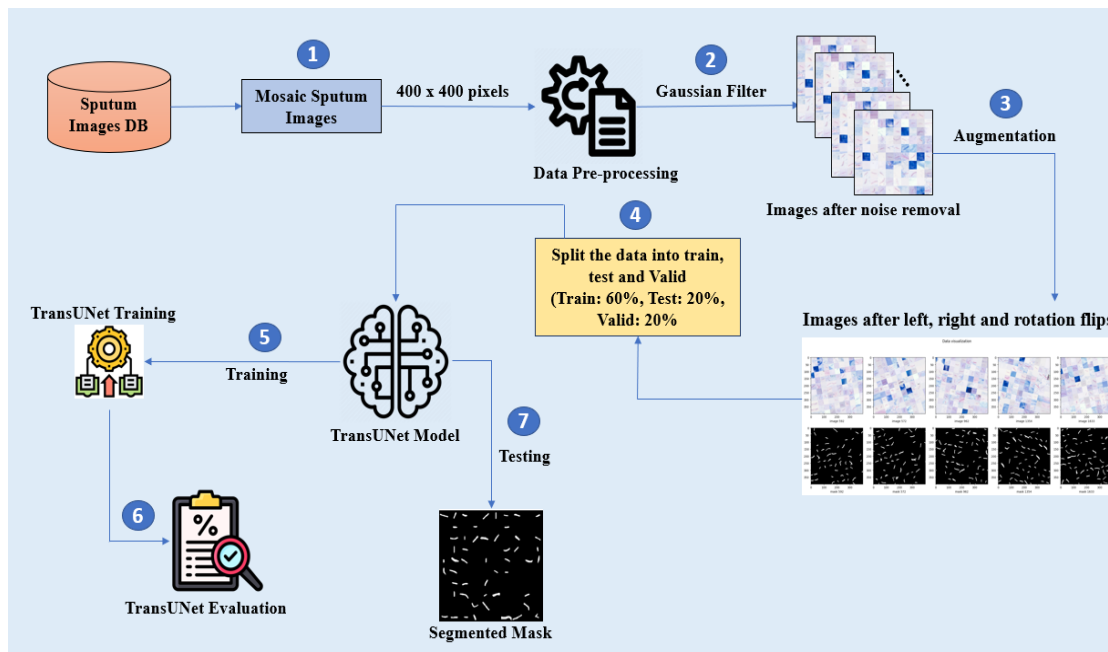


Fig. 2. System architecture for TB bacilli segmentation in sputum image

3.2. Data Preprocessing

Data preprocessing techniques help enhance the semantic segmentation model's performance. The proposed study applies Gaussian filter to remove noise that can significantly affect the model's performance. Algorithm 1 illustrates the procedure for applying the Gaussian filter to the image [23]. The Gaussian filter smoothens the image, preserves the edges, and allows the model to concentrate on the important features by removing the high-frequency components that are subjective to noise. The computational costs associated with gaussian filter in denoising the images are minimum and it is easy to compute when compared to other filtering methods like median and bilateral filters [24, 25]. Moreover, the sputum images contain speckle noise due to various other artifacts present in the image. Hence a gaussian filter is more suitable for reducing the high frequency noise when compared to the other filtering techniques, as it smoothens the noise while retaining the structure of bacilli and leads to better segmentation results.

Algorithm 1. Illustrates the procedure followed when applying Gaussian filter.
Algorithm 1. Gaussian Filter

Input: An image I of size $A \times A$ pixels and Gaussian standard deviation of σ .

Output: A noise-free filtered image I' .

1. Choose the kernel 'k' of size 5.
2. Calculate the value of 'n' as $n = \frac{(k-1)}{2}$ to get the range of kernel.
3. For each $(i, j) \in [-n, n]$ calculate the gaussian value as

$$G(i, j) = \frac{1}{2\pi\sigma^2} e^{-\frac{i^2+j^2}{2\sigma^2}}.$$

4. For each pixel $(x, y) \in I$, place the gaussian kernel over $I(x, y)$ and calculate

$$I'(x, y) = \sum_{i=-n}^n \sum_{j=-n}^n G(i, j) \cdot I(x + i, y + j)$$

5. Apply zero padding for handling the boundaries.
6. Return the filtered image.

Algorithm 1 generates the noise free, filtered image by taking the image I of $P \times P/3$ pixels resolution. The gaussian value is computed across the kernel range using $G(i, j)$ ensuring smooth and noise free output. Here, i, j represents the indices of the image in the kernel range where the gaussian filter is applied. The gaussian value determines the weight i.e. given to each pixel during the process of convolution. The gaussian filter is placed on the input image to perform the convolution, with padding applied to handle the edge cases. The noise free and smoothened image is represented by $I'(x, y)$, where x, y represents the coordinates of the image. This is followed by a data augmentation process to improve the model's generalizability. These filtered images are then applied with augmentation techniques, which include left flip, right flip, and rotation over angle theta. The left flip flips the image along the vertical axis, and the right flip along the horizontal axis. Such kinds of augmentations fetch the features of ROI from different angles. Thus, the model is more generalizable. Algorithm 2 illustrates data augmentation applied to the filtered images [26]. The images and their corresponding masks of $W \times Y$ pixels in the testing data are resized to $A \times A$ pixels to ensure uniformity in input dimensions for the segmentation model.

Algorithm 2. Illustrates the procedure followed when applying data augmentation.
Algorithm 2. Data Augmentation

Input: An image I' of size $N \times M$ pixels

Output: Augmented Images.

1. Set the rotation angle θ .
2. Apply left flip by swapping the x-coordinates of the image as shown below-

$$x' = N - 1 - x, y' = y$$
3. Apply right flip by swapping the y-coordinates of the image as shown below-

$$x'' = x, y'' = M - 1 - y$$
4. Apply rotation by an angle θ at the center as shown below-
 - 4.1. Shift the origin to the center (x_c, y_c) .

$$x_c = \frac{N}{2}, y_c = \frac{M}{2}$$
 - 4.2. Apply rotation by ' θ '

$$x_n = (x - x_c) \cos(\theta) - (y - y_c) \sin(\theta) + x_c$$

$$y_n = (x - x_c) \sin(\theta) - (y - y_c) \cos(\theta) + y_c$$
5. Apply transformations to the image.
6. Return the augmented images.

Algorithm 2 applies the data augmentation techniques like left, right and rotation flips to the filtered images. The input to the algorithm is the $N \times M$ pixels image. The x-coordinates are flipped in the left flip operation and the new coordinates are x' and y' respectively. Similarly, the right flip is applied by flipping the y-coordinated of the image and are represented as (x'', y'') . In rotation flip we utilize an angle of rotation θ and rotate it about the modified origin (x_c, y_c) and the new coordinates of the image become (x_n, y_n) .

3.3. Image Segmentation

3.3.1 TransUNet

TransUNet is a hybrid deep learning approach that combines the strength and power of transformers with traditional U-Net architecture particularly for semantic segmentation tasks. Based on the ability of transformers, it captures global dependencies and long-range context, enhancing the feature extraction process. It incorporates transformers into traditional U-Net architecture, which improves the segmentation, especially of complex and intricate images. This model is so effective for medical image segmentation, where the boundary detection is sharply done and contextual understanding is necessary, as with TB bacilli segmentation in microscopic images. Mosaic images with a $P \times P/3$ pixels resolution are used for training the TransUNet model and for predicting the semantic mask provided the actual sputum image. The traditional methodologies utilizing CNN conceal the images into high-level aspects and then de-conceal them in full spatial resolution. CNNs have effective localization capability but fail to remember the long-range dependencies and thus, suffer in extracting the global context information. The TransUNet model introduces the self-attention mechanism into the encoder part of the model. So, the limited localization of transformers is handled by CNN, and the transformer handles the effective extraction of global context information. The encoder helps in extracting the features by reducing the spatial dimensions of the input image. In TransUNet self attention mechanism is applied to extract the features present in the image and helps in extract meaningful and hierarchical information from data. A view of the encoder architecture of TransUNet is laid out below as shown in Fig. 3. Input images of the form X (B, C, H, W), are passed into the CNN block of the encoder for extracting the local details. Here, B refers to batch size, C denotes the number of channels, H and W represent the number of pixels in terms of height and width of the image. The CNN block of the encoder consists of two convolutional layers of 3×3 filters, each followed by batch normalization and ReLU layers. Moreover, each input of the convolutional layers is associated with the output, which builds a residual connection to ensure that no part of information gets lost. The CNN block of the encoder consists of two convolutional layers of 3×3 filters, each followed by batch normalization and ReLU layers. Moreover, each input of the convolutional layers is associated with the output, which builds a residual connection to ensure that no part of information gets lost.

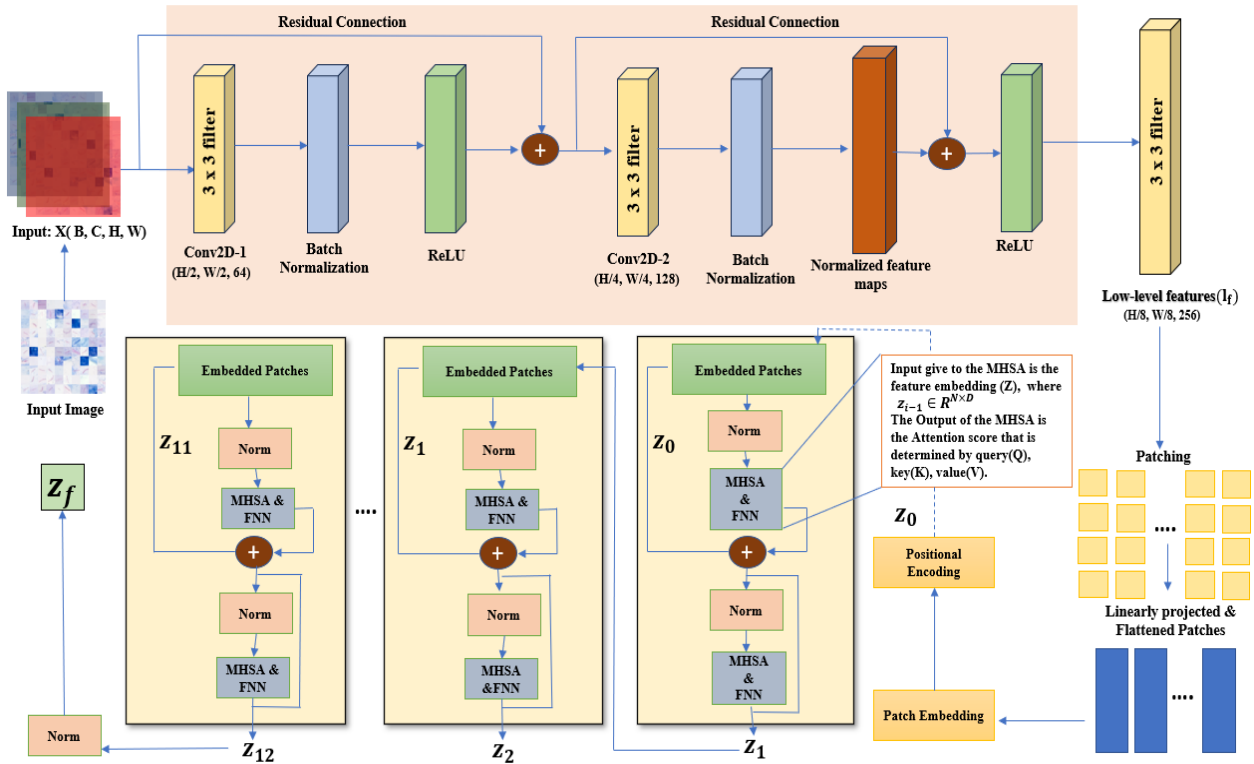


Fig. 3. Encoder of TransUNet model

$$I^c(i, j) = \sum_{m, n} I(i + m, j + n) \cdot k(m, n) \quad (1)$$

Equation (1) represents a typical convolutional operation performed on the input image. I represent the input image; k represents the kernel or the filter applied on the image; (m, n) represents the filter indices; and (i, j) represents the position of the resultant feature map. Batch normalization is performed on the feature map to ensure a harmonious distribution and faster confluence. ReLU helps the model to focus on the positive activations by neglecting the non-informative negative activations, ensuring better feature extraction. Residual connections are set to prevent the loss of important information during training. These connections are used to enable the flow of gradients in the model during

backpropagation and mitigates the vanishing gradient problem. The convolutional block produces the low-level features l_f which are then sent for patching in order to produce a sequence of patches. The low-level features of dimensions $\frac{H}{8} \times \frac{W}{8} \times 256$ are divided into N patches of size 16×16 where N is $\frac{H \times W}{P^2}$ and P is the patch size. Each patch is extracted by passing the $P \times P$ window across the l_f and are then flattened in order to produce the sequence of flattened patches, typically 1-D vectors represented by $X_p \in R^{N \times (P^2 C)}$. The flattened patches are embedded to transform them so the model can process them. As mentioned in Equation (2), the flattened patches are mapped to feature space of dimension 'D'.

$$z = X_p \cdot W_e + b_e \quad (2)$$

Here, X_p is the image patch, $W_e \in R^{(P^2 C) \times D}$ and $b_e \in R^D$. In order to impart more information about the data like patch order and position, positional encoding is applied, which is illustrated in Equation (3). PE is positional encoding pertained to X_p and it is computed using sine and cosine functions as given in the Equations (4) and (5). These periodic functions will introduce unique patterns for every position, and the model will be able to recognize whether patches are spatially adjacent or distant. In this encoding, the model will understand the spatial structure of input, which is very significant in vision tasks where hierarchies of spatial regions play a very important role.

$$z_p = z + PE \quad (3)$$

$$PE_{i,2j} = \sin\left(\frac{i}{10000^{2j/D}}\right) \quad (4)$$

$$PE_{i,2j+1} = \cos\left(\frac{i}{10000^{2j/D}}\right) \quad (5)$$

i is the position of the patch, and j is the dimension index. After performing patch embedding and positional encoding, we get the final embedded patches which are sent into the transformer block that consists of self-attention mechanisms. The transformer has used a self-attention mechanism, which extracts the relationship between different parts of the image. Because of that, it can capture long-range dependencies. Self-attention mechanism comes in computing query (Q), key (K), and value (V) projections. This way, the model learns the importance of a particular patch in the data. The query personifies the current patch to be processed. The key represents all the data patches and helps compute the relevance of the query. The value is the information that a patch contains and, is crucial for determining its contribution to the overall output. The projections of Q , K , and V are computed as following: Q is $z_p \times w_Q$; K is $z_p \times w_K$ and V is $z_p \times w_V$, respectively. These projections work together in the attention mechanism to enable the model to focus on the most relevant regions of the input. This process ensures that both local and global information are effectively utilized for the task. Here, $w_Q, w_K, w_V \in R^{D \times D_{head}}$ and D_{head} is the dimensionality of keys. The attention mechanism computes the score between Q and K . This is then passed on to the SoftMax function to create the probability distribution. This distribution contains weight given to each patch present in the data. Once the attention score is calculated, it is combined with the value to produce the final attention score, as depicted in Equation (6).

$$Attention(Q, K, V) = Softmax\left(\left(\frac{QK^T}{\sqrt{D_{head}}}\right)V\right) \quad (6)$$

In this way, we compute the attention scores for all the patches, concatenate them, and linearly project them to sustain their original dimensional space. This gives the output of Multi Head Self-Attention (MHSA) layer as shown in the Equation (7). The output of MHSA mainly focuses on the areas that have relevance, regardless of their spatial proximities and lets the model to learn long-range dependencies present in the data. This allows the model to distinguish between the data and generalize well, when exposed to new unseen data.

$$MHSA(Q, K, V) = concat(SA_1, SA_2, \dots, SA_h).w_o \quad (7)$$

Here, SA_i is $Attention(Q_i, K_i, V_i)$ and W_o is the learnable weight matrix responsible for linear projection. The output is then given to the feed-forward neural network to provide nonlinearity and to impart local and global context information to the model. Residual connections are also provided to prevent the information loss. Normalization is applied to the output of residual connections to stabilize, improve the learning process, and enhance the layer interactions. This leads to better performance for the overall model. The final output embedding of the encoder Z_f is reshaped to $512 \times \frac{H}{16} \times \frac{W}{16}$. This is sent as input to the decoder. The encoder passes the high-level feature representation that contains semantic information. This high-level information is passed to the decoder for up-sampling which comprises several convolutional layers in series and is followed by batch normalization and ReLU layers. The

architecture of the decoder present in TransUNet is presented in Fig. 4 [7]. The first convolutional layer in the decoder is called the Up-convolution, which is responsible for restoring the original image size. Skip connections combine the feature maps produced at every layer of the encoder and decoder; this ultimately preserves the semantic information at different stages during up-sampling. The output from the new tensor is passed through ReLU and batch normalization layers and after several up-convolution, followed by batch normalization and ReLU layers, the decoder delivers a semantic mask of size $(H \times W)$. This is the pixel-level prediction of the semantic mask. The encoder and decoder interplay are adequately facilitated by skip connections and convolutional operations, which provide comprehensive detail in segmentation maps while capturing the underlying structures within an image. The architecture improves the understanding of spatial hierarchies and significantly improves the accuracy of the segmentation task at hand.

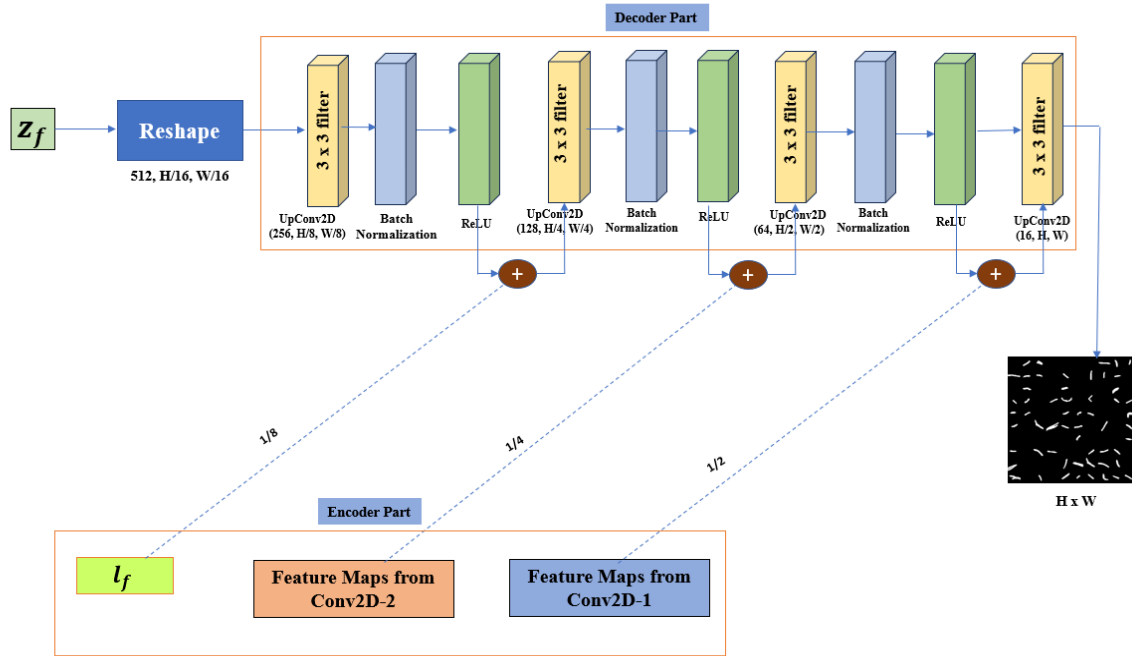


Fig. 4. Decoder of TransUNet Model

4. Experimental Setup

The TransUNet model was tested on mosaic sputum image datasets that were resized to a resolution of $A \times A$ pixels for uniformity, in accordance with the specifications provided in Table 1. Further, to remove noise and ensure that the images are clearer for better optimization of the segmentation performance, a Gaussian filter was applied on the images as a form of preprocessing. To combat the scarcity of data and to improve the generalization capability of the model, horizontal flipping, vertical flipping, and rotation data augmentation techniques were employed, making the dataset more diverse in nature. The dataset was partitioned into training, validation, and testing sets, in the ratio 60:20:20 to enable training of the model and hyperparameter tuning from a portion that would serve for unbiased evaluation. The model is trained up to 20 epochs using Adam and RMSprop optimizers with a set of varying learning rates that includes 0.1, 0.5, 0.01, 0.05, 0.001, 0.005, 0.0001, 0.00001 and 0.00005. The model uses a patch size of 16×16 pixels, hidden size of 768, 12 transformer layers, 12 attention heads, and an MLP dimension of 3072, coupled with a decoder that has 256 channels for the first layer, 128 channels for the second layer, 64 channels for the 3rd layer and 16 channels for the 4th layer, respectively. Its performance is tested with traditional evaluation metrics like DSC, mean IoU, and accuracy. Finally, evaluation is done on the test set where the model can produce masks that could efficiently outline the regions of the bacilli. This large-scale experiment successfully validated the reliability and efficacy of the proposed methodological approach for TransUNet-based semantic segmentation.

Table 1. Infrastructure specifications

Components	Specifications
Processor	AMD Ryzen™ 5 5500U @ 2.10 GHz
Memory	32 GB DDR4 RAM
Storage	SSD 1 TB
Graphics card	Radeon Graphics 448SP
Operating System	Windows-11
Deep Learning Framework	Pytorch 2. x

4.1 Data Collection

The data was collected from a conventional image database of Digital Dataset for Sputum Smear Slides 3 (DDS3), which includes 8000 mosaic images whose resolution is 400×400 pixels [25]. Unlike the previous methods which focused on using only one dataset, this work utilizes images that are formed with bacilli of different types, collected from various slides of sputum smears. Each patch accounted for 40×40 pixels and for every image, that is 400×400 pixels in size, there is a corresponding mask. In the training set, the negative patches account for 100%, 90%, and 50% mosaic images. We considered 50% negative image samples of 3000 mosaic images and their corresponding masks such that each image contains a balanced portion of positive and negative patches. Another dataset comprising 35 images from the ZN Sputum images database has been considered to test the model's generalizability [7]. Each image in the dataset is of 2112×2816/3 pixels resolution and was manually annotated using MATLAB; thus, the data was prepared to test the model's generalizability. This means the use of these datasets guarantees a variety of testing environments for the model, such as high-resolution images and smaller segments that reflect real-world situations in the diagnosis of tuberculosis. The mosaic dataset, with pixel-level annotations, is particularly advantageous in the training of deep learning models because it makes the accurate segmentation of bacilli and related artifacts easier. Conversely, the ZN Sputum dataset contains rich, high-resolution imagery, and its test set should pose a difficult test, capable of ascertaining how the model will generalize to data that it has not previously encountered. Further to this evaluation, the training dataset was built carefully, such that the model learned both negative and positive samples with an equal amount in each to ascertain effectively both the presence and absence of bacilli.

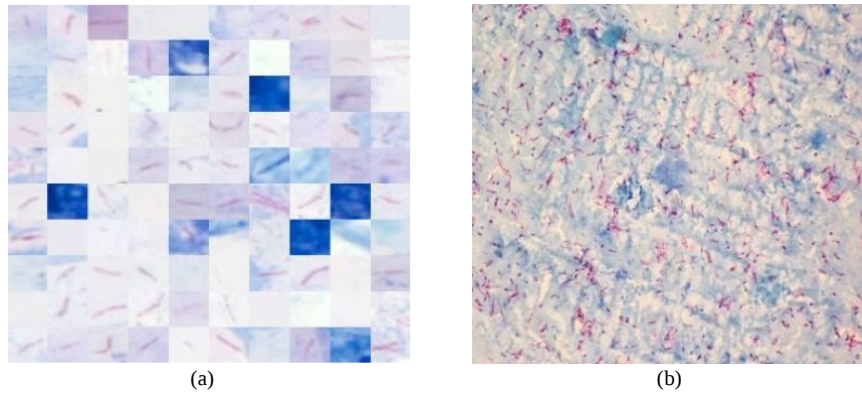


Fig. 5. Sputum Images: (a) Sample sputum image from Mosaic dataset; (b) Sample sputum image from ZN dataset.

The DDS3 Mosaic dataset was selected for training since it offers a high number of annotated images with pixel-level masks, which are critical for proper bacilli segmentation. Its patch-based format enables the model to learn fine-grained information from bacilli of different morphologies, and the presence of positive and negative examples ensures balanced training. The ZN Sputum Images dataset was chosen to evaluate the generalizability of the model since it features high-resolution, real-world clinical images. It presents a perfect candidate for assessing how well the model fares on unseen, real-world scenarios.

5. Results and Discussion

5.1 Evaluation Metrics

Evaluation metrics are important measures used to assess the model's performance. The propounded scheme uses DSC, mean-IoU, segmentation accuracy, precision, and recall as performance metrics. The segmentation accuracy, precision, and recall are calculated based on the True positives (TP), True negatives (TN), False positives (FP), and False negatives (FN). TP is the number of pixels classified correctly according to the target class label. TN is the no. of pixels correctly classified according to the non-target class. FP is the number of pixels incorrectly classified as the target class label, and FN is the number of pixels classified incorrectly as a non-target class label. The measure using a DSC evaluates the overlap between the predicted and the ground-truth segmentation mask as given in Equation (8).

$$DSC = 2 \left(\frac{|x \cap y|}{|x| + |y|} \right) \quad (8)$$

“x” represents the predicted image, and the y represents the ground truth image. As mentioned in Equation (9), mean IoU is given by the average ratio of the number of pixels common in predicted and ground truth masks to the total number of pixels in the ground truth mask. Segmentation accuracy is the measure that calculates the count of correctly classified pixels to the total no. of pixels. Equation (10) depicts the computation of segmentation accuracy. As mentioned in Equation (11), precision is a metric that represents the certainty or veracity of the model's positive prognosis.

$$\text{Mean IoU} = \frac{1}{c} \sum_{i=1}^c \frac{|x \cap y|}{|x| + |y|} \quad (9)$$

$$\text{Segmentation Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (10)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (11)$$

Recall, also referred to as sensitivity, is a measure for how well the model accurately identifies the presence of a particular instance in the dataset. It shows the percentage of actually occurring positive cases that are predicted well by the model, implying its success rate in detecting the presence of the target class.

$$\text{Recall} = \frac{TP}{TP+FN} \quad (12)$$

A higher recall suggests that the model is more sensitive and less likely to miss positive instances, making it an important measure of the model's legitimacy and reliability. The equation for recall is computed as the ratio of TP to the sum of true positives and FN, as shown in Equation (12).

5.2 Data Preprocessing

In the Methodology section, under Data Preprocessing, the first thing to do is apply a Gaussian filter to the images. Gaussian filtering is one of the common image processing techniques that reduce noise and smoothen an image by averaging the pixel values in the neighborhood of each pixel according to a Gaussian distribution. High-frequency variations associated with noise are thus reduced. This quality enhancement not only reduces noise but also enhances the overall image interpretation. More importantly, it retains the core spatial dependencies in the images so that the structural integrity is preserved intact. The images are thus ready for further analysis and processing stages using Gaussian filter. Fig. 6 illustrates the image after applying the Gaussian filter, the output shows smoothness and a decrease in noise, which looks more qualitative. Additional data augmentation techniques involving other transformations have been performed to add much diversity to the datasets and improve the model's robustness. For this analysis, images are augmented, and the outcome images are shown in Fig. 7.

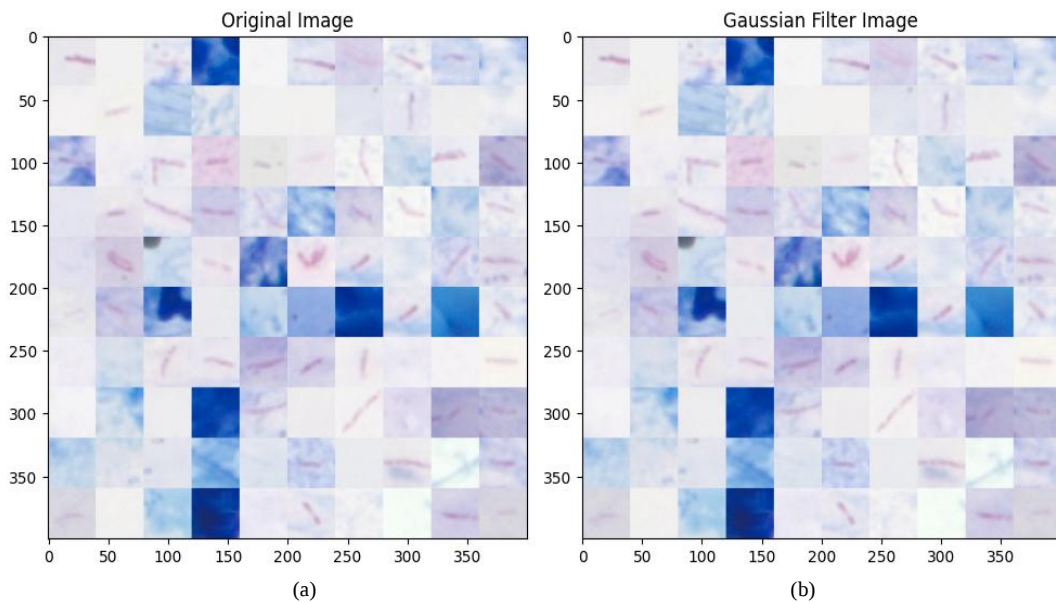


Fig. 6. Gaussian filter: (a) Original sputum sample; (b) Noise-free images obtained by applying gaussian filter

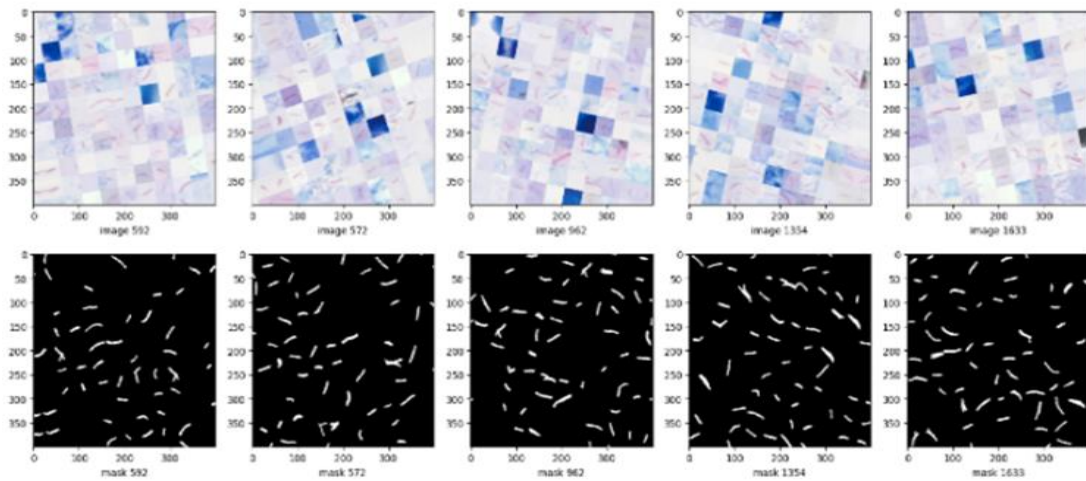


Fig. 7. Data augmentation on sputum images of Mosaic dataset

Applying the data augmentation techniques illustrated in Algorithm 2 to an initial dataset of 3,000 images gives multiple transformed versions of each image. Each image would have undergone a left flip, right flip, and rotation, plus retain the original image, so there would be four versions per image, bringing the total number of images to 12,000. The data is split into three subsets using two different partitioning techniques; [60%, 40%] and [80%, 20%]. The [60%, 40%] split assigned 7,200 images for training subset and 2,400 images each to test and validation subsets. On the contrary, the [80%, 20%] split assigned 9,600 images to the training dataset and divided the remaining 2,400 images between the test and validation datasets, 12,000 images each. Such splits allow different distributions of data that are required for proper assessment of the model and generalization. Table 2 illustrates the details regarding the data partition, highlighting the distribution of images into training, testing, and validation sets using both [60%, 40%] and [80%, 20%] splitting strategies.

Table 2. Training, test and validation data partition details

Total Images	Train set		Test set		Valid set	
	[60%, 40%]	[80%, 20%]	[60%, 40%]	[80%, 20%]	[60%, 40%]	[80%, 20%]
12,000	7,200	9,600	2,400	1200	2,400	1200

5.3 Image Segmentation

After completing data preprocessing techniques on the dataset, the TransUNet model is trained on it by providing the actual sputum image and the ground truth mask. The model achieved good outcomes, achieving an accurately predicted mask when a sample sputum image is provided. The predicted mask is compared with the true mask and both of them are quite similar to each other. This indicates that the model is adeptly learning to differentiate between different classes of pixels, such as bacilli and background. The results of a single image are given in Fig. 8 where Fig. 8(a) shows the actual mosaic sample; Fig. 8(b) shows the ground truth mask and Fig. 8(c) shows the predicted mask generated by the TransUNet model. Fig. 9(a) is one of the images from the ZN Sputum database. It shows a stained sputum smear with visible bacilli. The image contains many structures and artifacts and the bacilli appear as elongated red shapes against a light blue background, indicating the presence of TB bacteria in the sputum sample. Qualitatively, the estimated masks reveal that the TransUNet model is very good at recognizing and segmenting bacilli in challenging and noisy images, even when samples present marked artifacts and structural variations. This proves that the model can be very robust to data from the real world in the ZN Sputum database, where challenges of staining inconsistencies and overlapping entities often prove difficult for classical segmentation techniques.

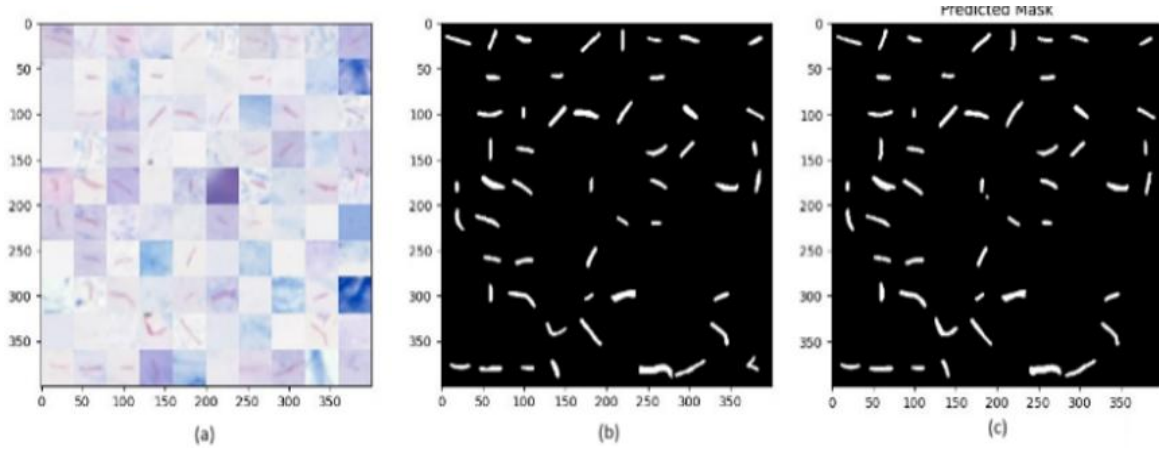


Fig. 8. Segmentation by TransUNet model on Mosaic image: (a) Sample sputum image from the mosaic dataset; (b) Ground truth mask from the mosaic dataset; (c) Predicted mask by TransUNet.

Fig. 9 shows the model prediction for segmentation of bacilli from a ZN sputum smear image. Fig. 9(a) and 9(b) shows the sample sputum image and its corresponding ground truth mask. The white shape against a black background signifies the segmented area of detected bacilli by TransUNet as shown in Fig. 9(c). This helps to differentiate a bacillus from its other region in the imaging and readily identifies it. Fig. 9(d) shows the visual inspection of the segmented image that explains the quality of segmentation. Fig. 10 demonstrates the performance of TransUNet on Mosaic Data. Fig. 10(a) – 10(e) are the ground truth images from the dataset and contain several examples of mosaic sputum images used for the segmentation task. Fig. 10(f) – 10(j) present the true masks where the white regions represent the position of bacilli in ground truth. Fig. 10(k) – 10(o) are the masks predicted by the model, representing the bacilli's detected region. Fig. 10(p) -10(t) illustrates the visual inspection of segmentation which is done by overlapping the predicted mask on the actual images. A comparison of actual, true, and predicted masks will validate this conclusion to show how effectively the model identifies bacilli in mosaic sputum images.

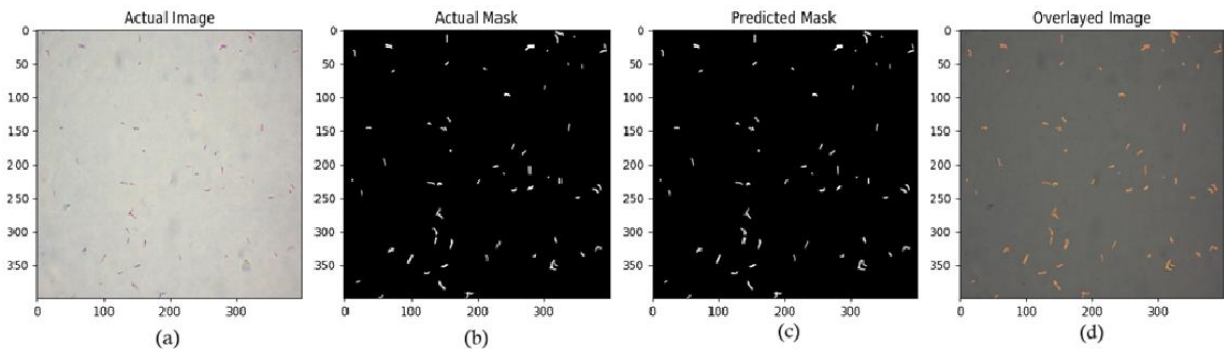


Fig. 9. Segmentation by TransUNet model on ZN sputum image: (a) Sample sputum image from ZN dataset; (b) Actual mask; (c) Predicted mask by TransUNet; (d) Overlaid image

Tables [3-6] summarize the detailed experiments of the performance of the two widely used optimization methods, Adam and RMSprop, using different learning rates. The learning rates include 1.0, 0.5, 0.1, 0.05, 0.01, 0.005, 0.001, 0.0005, and 0.0001 on datasets split with [60%, 40%] and [80%, 20%], respectively. The presented tables illustrate the relative efficacy of each optimizer concerning convergence characteristics, stability, and overall model performance. They offer significant insights into the learning rate selection influence on the optimization efficiency in different contexts of training and validation. This examination contributes to an understanding of how optimization algorithms and learning rates interact in affecting the model's capability to generalize and achieve optimum performance.

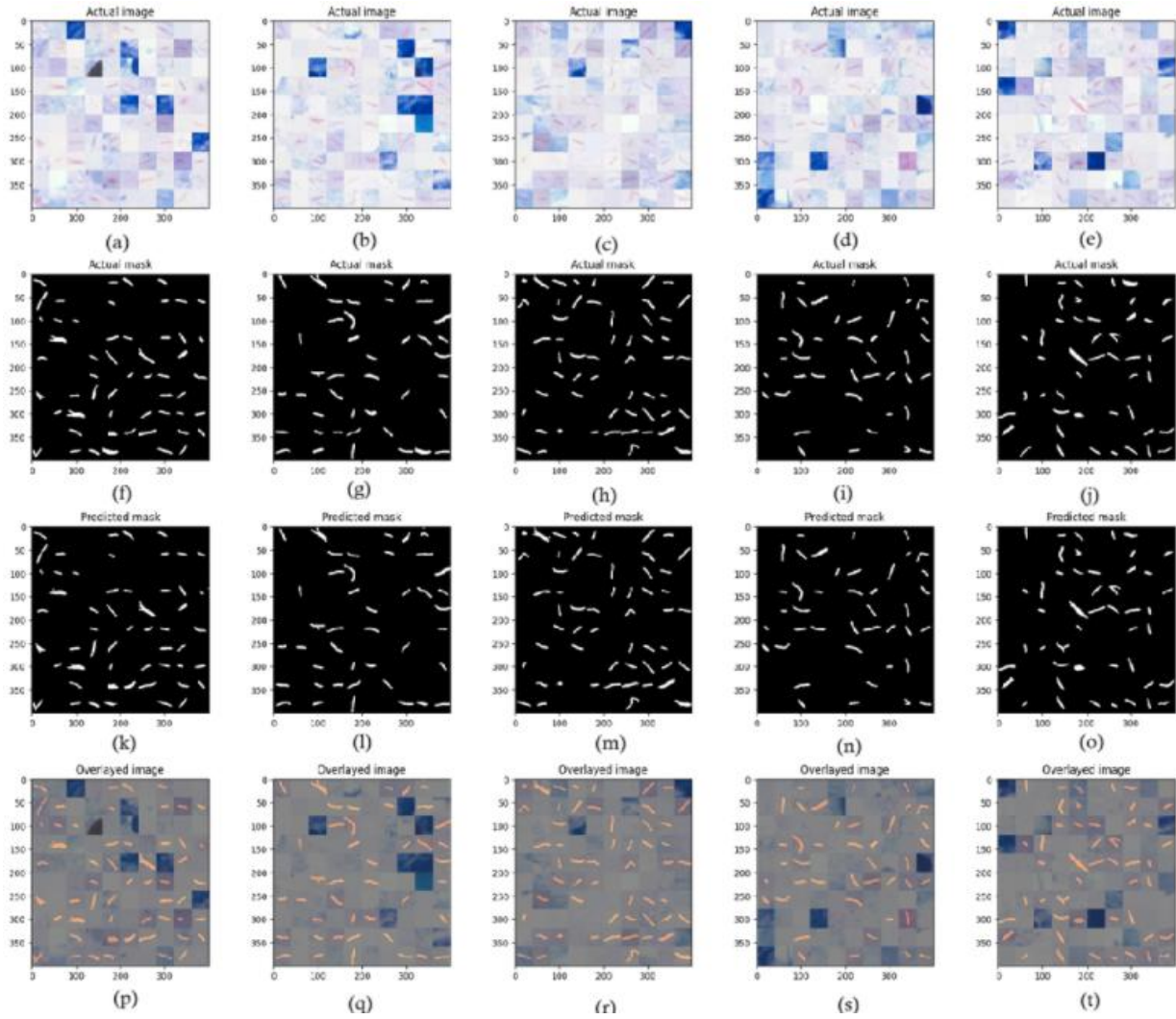


Fig. 10. Segmentation by TransUNet on Mosaic Dataset: (a-e) Sample Sputum images by Mosaic dataset; (f-j) Ground truth masks; (k-o) Predicted masks by TransUNet; (p-t) Overlaid images

Table 3. Performance obtained by TransUNet using Adam optimizer on 60% train and 40% test data.

Initial Learning Rate	Performance Metrics									
	DSC (%)		Mean-IoU (%)		Accuracy (%)		Precision (%)		Recall (%)	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
0.1	89.78	92.19	80.39	84.34	99.20	99.40	89.84	94.06	89.78	90.39
0.5	89.09	91.54	79.07	82.70	99.15	99.34	89.17	91.84	89.10	91.26
0.01	90.04	92.77	80.75	85.16	99.22	99.41	90.26	92.26	89.73	92.73
0.05	89.98	92.49	80.64	84.40	99.22	99.41	90.26	92.26	89.73	92.73
0.001	93.06	94.16	92.12	93.49	99.51	99.74	98.10	98.03	96.60	95.36
0.005	90.40	93.02	81.35	85.34	99.25	99.46	90.70	93.53	90.10	92.52
0.0001	90.43	93.04	81.39	85.46	99.25	99.46	90.73	93.50	90.13	92.59
0.00001	90.45	93.04	81.42	85.46	99.26	99.46	90.75	93.50	90.15	92.60
0.00005	90.44	93.05	81.43	85.47	99.26	99.46	90.74	93.50	90.15	92.55

Table 3 shows the evaluation of the TransUNet model on a [60%,40%] train-test split using the Adam optimizer with different learning rates. The results depict that the learning rate is an important factor to decide multiple performance indicators including DSC, mean-IoU, Accuracy, Precision, and Recall for train and test data. Interestingly, it achieved the best mean-IoU and DSC on the test dataset at a learning rate of 0.001, with values of 92.12 and 94.16, respectively. Similarly, this learning rate resulted in an improved accuracy of 99.74%, precision of 98.03%, and a recall of 95.36%, respectively.

Table 4. Performance obtained by TransUNet using RMSProp optimizer on 60% train and 40% test data.

Learning Rate	Performance Metrics									
	DSC(%)		Mean-IoU (%)		Accuracy (%)		Precision (%)		Recall (%)	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
0.1	88.59	91.07	78.64	81.47	97.32	98.43	88.91	92.87	88.27	89.33
0.5	88.22	90.76	77.21	80.65	98.32	99.13	88.56	90.76	87.89	90.76
0.01	89.22	91.46	79.86	83.53	99.02	99.54	89.78	91.43	88.67	91.49
0.05	89.19	91.66	79.45	82.78	98.17	99.01	89.34	91.77	89.04	91.56
0.001	96.34	93.36	91.87	92.76	98.23	98.96	97.51	97.98	95.19	94.80
0.005	88.75	91.34	80.71	84.58	98.30	99.12	88.89	91.90	88.62	90.78
0.0001	89.72	92.06	79.80	83.52	97.87	98.18	89.49	92.67	89.96	91.45
0.00001	87.65	90.97	80.94	88.56	96.91	98.18	87.08	91.79	88.23	90.16
0.00005	89.52	92.01	80.23	84.60	97.11	98.03	89.63	92.71	89.42	91.32

Table 4 shows the performance of TransUNet model that is assessed for different learning rates, using a [60%, 40%] train and test splits with the RMSProp optimizer. It is evident that choices of learning rate have influenced the performance metrics. Among all the learning rates that were tested, the best performance was obtained at 0.001, for which model DSC was 93.36% on the test set and had mean-IoU to be 92.76%, with very good accuracy of 98.96%, precision of 97.98%, and recall of 94.80%; these results indicate good generalization and segmentation at this training rate.

Table 5. Performance obtained by TransUNet using Adam optimizer on 80% train and 20% test data.

Learning Rate	Performance Metrics									
	DSC(%)		Mean-IoU (%)		Accuracy (%)		Precision (%)		Recall (%)	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
0.1	90.07	92.54	80.79	85.17	99.23	99.45	90.37	92.54	89.80	93.55
0.5	89.25	91.89	79.38	83.37	99.16	99.35	89.44	89.90	89.14	93.97
0.01	90.04	92.60	80.90	84.76	99.23	99.42	90.41	91.96	89.90	93.44
0.05	90.19	92.84	81.00	84.96	99.23	99.44	90.51	92.43	89.88	93.26
0.001	90.48	91.06	81.49	85.52	99.26	99.46	90.77	93.58	90.19	92.36
0.005	90.40	93.05	81.36	85.37	99.25	99.45	90.75	92.75	90.06	93.36
0.0001	90.42	93.05	81.38	85.39	99.25	99.46	90.77	92.88	90.08	93.21
0.00001	90.45	93.04	81.43	85.39	99.25	99.46	90.75	92.91	90.15	93.16
0.00005	90.41	93.05	81.43	85.40	99.26	99.46	90.77	92.97	90.14	93.12

The efficacy of the model TransUNet, validated on an [80%, 20%] training-testing partition with Adam optimizer over different learning rates, is summarized in Table 5. The model had good performance on all of the evaluated metrics, though the best results on the testing dataset were seen at a learning rate of 0.0001. It achieved a DSC of 93.05%, mean-IoU of 85.39%, accuracy of 99.46%, precision of 92.88%, and recall of 93.21%. All these metrics reflect that there is a proper balance between the generalization and segmentation accuracy.

Table 6. Performance obtained by TransUNet using RMSProp optimizer on 80% train and 20% test data.

Learning Rate	Performance Metrics									
	DSC(%)		Mean-IoU (%)		Accuracy (%)		Precision (%)		Recall (%)	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
0.1	88.18	90.89	77.98	82.24	97.10	98.32	88.76	90.26	87.60	91.52
0.5	87.12	89.28	76.45	81.30	97.36	98.18	87.90	87.43	86.35	91.22
0.01	87.90	90.82	78.44	82.19	98.03	99.10	88.12	89.78	87.68	91.89
0.05	88.04	90.65	79.67	82.11	97.89	98.35	89.01	90.56	87.09	90.74
0.001	89.43	91.74	79.46	83.18	98.11	99.03	89.65	92.33	89.21	91.16
0.005	88.72	91.16	79.98	84.08	98.23	97.46	88.90	90.47	88.54	91.87
0.0001	90.95	92.00	79.43	84.00	98.12	98.96	89.23	90.10	92.73	93.98
0.00001	89.77	91.87	80.01	84.64	97.90	98.67	88.16	90.98	91.44	92.78
0.00005	89.44	91.01	80.87	83.61	98.11	98.97	89.23	90.16	89.65	91.87

The performance of the TransUNet model, assessed using [80%, 20%] train-test division and with the usage of RMSProp optimizer operating at different learning rates is provided in Table 6. Notably, a learning rate of 0.0001 gave the best test set performance with a DSC of 92.00%, Mean-IoU of 84.00%, Accuracy of 98.96%, Precision of 90.10%, and Recall of 93.98%. The training and testing performance were balanced for this learning rate, and generalization was robust.

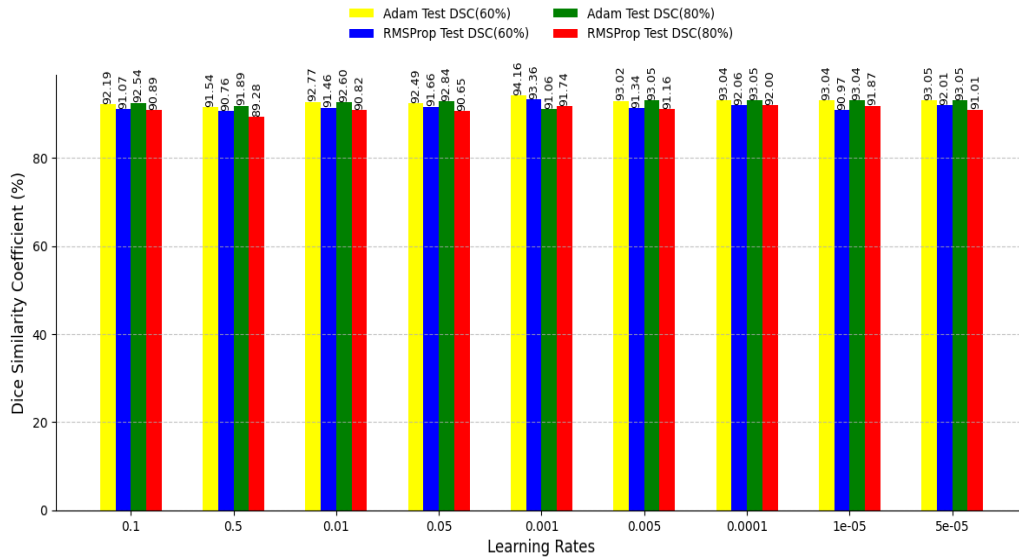


Fig. 11. Comparison between DSC scores of TransUNet model using Adam and RMSprop optimizers across different learning rates.

DSC scores for [60%, 40%] and [80%, 20%] train-test splits across learning rates and optimization algorithms, Adam and RMSProp are compared as shown in the Fig. 11. The bar chart plots learning rates on the x-axis and test DSC on the y-axis. Every bar presents an optimizer and train-test split, where the legends are as follows: yellow for Adam with 60% data split, blue for RMSProp with 60% data split, green for Adam with 80% data split, and red for RMSProp with 80% data split, respectively. The overall results show high DSC scores with slight variations in the learning rates and strategies. The Adam optimizer is stronger than RMSProp, especially at low learning rates like 0.001 and 0.0001, which shows that it is robust for segmentation. The results obtained by the [60%, 40%] train-test split is comparable to those obtained by the [80%, 20%] split, which shows low dependency on the split ratio. These results highlight the stability of optimizers across various hyperparameters.

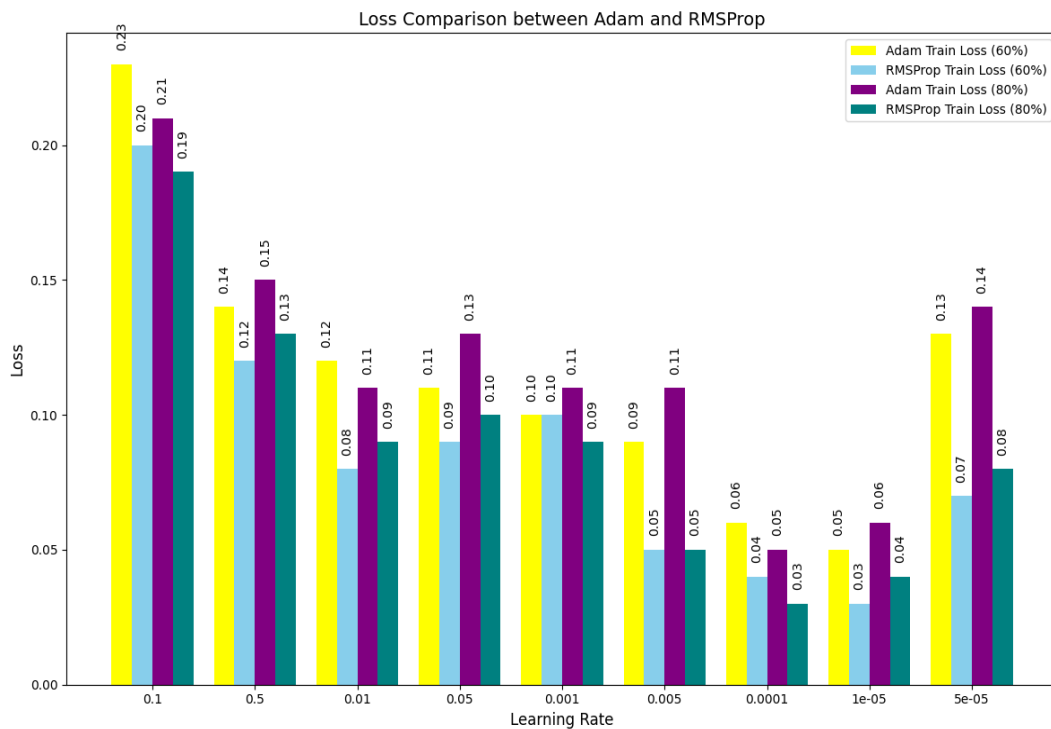


Fig. 12. Comparison between train losses of TransUNet model using Adam and RMSprop optimizers across different learning rates.

An analysis of the training loss values with respect to different learning rates for both Adam and RMSProp optimization algorithms, under two different train-test ratios: [60%, 40%] and [80%, 20%] is shown in the Fig. 12. The x-axis represents the learning rates, and the y-axis represents the corresponding loss values. Each combination of the optimizers is represented by a color: yellow for Adam with 60% data split, blue for RMSProp with 60% data split,

purple for Adam with 80% data split, and green for RMSProp with 80% data split, respectively. It shows that Adam consistently gives lower loss values than RMSProp in most of the learning rates, especially when the rates are smaller, for example, 0.005 and 0.001. The loss values decrease as the learning rate decreases. This shows how learning rate optimization impacts this. For instance, relatively higher loss values are given by RMSProp at bigger learning rates compared to Adam. This analysis focuses on the effectiveness of Adam in minimizing loss over different configurations and emphasizes the critical role of learning rates in the performance of optimization.

As mentioned in the previous sections, the results are evaluated according to the metrics: loss, mean-IoU, accuracy, precision, recall, and DSC. Table 7 shows the average performance metrics that the TransUNet model has achieved on the mosaic dataset using the Adam optimizer and a learning rate of 0.001. The results suggest excellent performance for the model on both training-testing splits, namely, [60%, 40%] and [80%, 20%]. Especially for the [60%, 40%] split, the model performed at a DSC of 94.16%, a mean-IoU of 93.49%, Segmentation Accuracy of 99.74%, Precision of 98.03%, and Recall of 95.36%. On the other hand, on the 80-20% split, the model attained a DSC of 91.06%, a mean-IoU of 85.52%, Segmentation Accuracy of 99.46%, Precision of 93.58%, and Recall of 92.36%. It further proves that the performance of TransUNet with Adam optimizer at an aptly tuned learning rate for the task is indeed very robust. For evaluating the model's performance, it was tested with the ZN dataset, which has given promising results as illustrated in Table 8. The model demonstrates a good segmentation performance, achieving a DSC score of 92.19% and a segmentation accuracy of 98.64%. Precision and recall values stand at 75.64% and 84.96%, respectively. These metrics collectively indicate effective segmentation capabilities and a high degree of reliability.

Table 7. Performance metrics achieved by utilizing TransUNet model on mosaic Dataset using Adam optimizer at a learning rate of 0.001

Performance Metrics	Value (%)	
	[60%, 40 %]	[80%, 20 %]
DSC	94.16	91.06
mean-IoU	93.49	85.52
Segmentation accuracy	99.74	99.46
Precision	98.03	93.58
Recall	95.36	92.36

Table 8. Performance metrics achieved by evaluating the TransUNet model on the unseen ZN test dataset using the Adam optimizer.

Performance Metrics	Value
DSC	92.19%
mean-IoU	85.57%
Segmentation accuracy	98.64%
Precision	75.64%
Recall	84.96%

Table 9. Performance comparison between the existing implementations and proposed work.

S. No.	Existing Work	Datasets	Algorithms/ Models	Performance Metrics	
1	Ali et al. [11]	ZN dataset	UNet	DSC	74%
2	Wulandari et al. [12]	ZN dataset	UNet for Y and V-shaped colonies of AFB	Segmentation accuracy	94.78%
3	Mithra & Emmanuel [13]	ZN dataset	CS transformation along with thresholding	Segmentation accuracy	91.379%
4	Fernández et al. [14]	ZN dataset	UNet	Segmentation accuracy mean IoU	92.01% 88%
5	Panicker et al. [15]	ZN dataset	Image Binarization	DSC	86.76%
6	Mithra & Emmanuel [16]	ZN dataset	C-Means Clustering	Segmentation accuracy	95.01%
7	Shwetha et al. [17]	ZN dataset	K-Means Clustering	Segmentation accuracy	97%
8	Mithra & Emmanuel [18]	ZN dataset	Channel area thresholding	Segmentation accuracy	97.58%
9	Ayas & Ekinci [19]	ZN dataset	Color Information based segmentation	Segmentation accuracy Recall	99.55% 87.02%.
10	Xavier et al.[20]	ZN dataset	Selection of scalar features from color spaces (RGB, HIS etc.,)	Segmentation accuracy	91.45%
12	Vanitha et al. [21]	ZN dataset	UNet	Segmentation accuracy	99%
13	Serrão et al. [22]	Mosaic dataset	UNet	Segmentation accuracy	99.64%
14	TransUNet for Semantic Segmentation of TB Bacilli	Mosaic dataset	TransUNet	DSC	94.16%
				Segmentation accuracy	99.74%
		ZN dataset		DSC	92.19%
				Segmentation accuracy	98.64%

Comparison was carried out against others works, that have already been published by using similar metrics appropriately. Table 9 also gives a clear idea on the performance of the proposed model with the improvement found through continuous assessment and comparison. The proposed work using TransUNet for semantic segmentation of TB bacilli shows a considerable improvement in segmentation accuracy over several existing methods. A variety of conventional segmentation techniques, including UNet, K-Means Clustering, and C-Means Clustering, have been applied to the ZN dataset with remarkable results.

Owais et al. used a modified U-Net model for semantic segmentation on the BraTs, Heart MRI, and ZN datasets, significantly reducing the model parameters from 30 million to 0.49 million [11]. The model obtained a DSC of 0.94 on the BraTs and Heart MRI datasets, while for the ZN dataset, the DSC was 0.74. This method pointed out the superiority of the model in medical image datasets but with a low application on the ZN dataset. Wulandari et al. proposed an intelligent intuitionistic fuzzy logic-based AFB count detection system [12]. The method used self-thresholding techniques in conjunction with U-Net architecture for segmentation. This system focused on the detection of Y and V-shaped colonies in sputum samples and achieved segmentation accuracy of 94.78%, which proved fuzzy logic useful in enhancing tasks of segmentation for AFB in TB diagnosis. Mithra and Emmanuel developed an automatic TB diagnosis system by putting together GMM with fuzzy logic and neural networks [13]. Using CS transformation and thresholding, the model is able to get an accuracy of 0.91379 for the segmentation of the ZN TB dataset. The proposed method resolved the issues of low sensitivity and time complexity associated with the traditional laboratory procedures that are generally followed for TB diagnosis. The proposed TransUNet methodology outperforms these implementations by offering high segmentation accuracy of 99.74% and DSC of 94.16% respectively, showing its effective performance.

Fernández et al. applied U-Net for semantic segmentation of less free microscopic sputum images by splitting them into 256×256 -pixel patches for training [14]. The model achieved a segmentation accuracy of 92.01% and mean IoU of 88%, which shows that U-Net can handle big, high-resolution sputum smear images for TB bacilli detection. Panicker et al. proposed a method that used image binarization and CNN-based classification for the purpose of detecting TB bacilli [15]. The obtained segmentation accuracy using unsupervised binarization techniques and trained a CNN for classifying the type of bacilli. Precisions and F1-score were also used as evaluating performances of the model, reporting a recall of 97.13% and the F1-score of 86.76%. Other TB diagnosis methodologies include segmentation and classification done by Mithra and Emmanuel, which used C-Mean clustering combined with fuzzy logic and obtained successful application in ZN TB image, with the segmentation accuracy of 95.1% [16]. Shwetha et al. made use of a CNN-based model labeled SqueezeNet to predict TB Bacilli in images of ZN TB through comparison with VGG16 as well as ResNet50 [17]. The accuracy of 97% achieved by that model was outperforming the other two, with great potential displayed by CNNs in being able to extract the bacilli from a sputum smear images. Compared to the methods discussed, TransUNet offers an advanced approach for TB bacilli isolation through, particularly in handling complex and high-resolution images. It combines the strength of CNN and the power of attention mechanism used in ViT to effectively learn the intrinsic patterns present in the data and gives an appreciable accuracy of 99.74%.

Mithra and Emmanuel proposed a model for semantic segmentation of sputum smear images using a channel area thresholding algorithm [18]. Their model achieved a segmentation accuracy of 97.58% after applying preprocessing steps like noise reduction and intensity normalization, showing the importance of data enhancement techniques in improving segmentation results. Ayas and Ekinci introduced a color information-based segmentation technique that utilizes pixel values and Gaussian probability density functions in 2014 [19]. Comparing SVMs for segmenting TB bacteria, their method resulted in a segmentation accuracy of 99.55% and a recall of 87.02%, thereby establishing the efficiency of color-based segmentation in the detection of TB. Xavier et al. presented a color space-based segmentation method, utilizing RGB, HIS, and YCbCr spaces along with feedforward neural networks for the identification of TB bacilli in sputum smear images [20]. The model obtained a segmentation accuracy of 91.45%, with a recall of 93.41%, thus underlining the significance of color space features for the enhancement of segmentation accuracy in the detection of TB bacilli.

Vanitha et al. compared supervised and unsupervised segmentation techniques for TB bacilli detection in sputum smear images, using U-Net for segmentation [21]. The method resulted in an outstanding accuracy of 0.99 in segmentation, demonstrating how U-Net performs much better over variation in images due to variability in staining procedures and microscopes. TransUNet showed excellent performance in the isolation of TB bacilli by achieving an accuracy of 99.74% for mosaic images and 98.64% accuracy for ZN images, TransUNet shows its superiority not only through its approach but also in its intrinsic nature by effectively using the attention mechanism that helps in capturing the long range dependencies present in the image. The high accuracy of TransUNet can be attributed to its advanced architecture that effectively utilizes the strengths of both Transformer and U-Net models, making it highly efficient in segmenting TB bacilli in sputum smear images. Serrão et al. proposed a method for segmenting the TB bacilli from mosaic images using UNet architecture [22]. Although the above methods gave good performance in terms of segmentation accuracy, they also possess few limitations in terms of noise, generalizability towards unseen data, suffers from information gap. TransUNet which is trained on mosaic images, overcomes these limitations by including the strengths of CNN and ViT and captures the complex inherent properties and generalizes well with unseen data i.e. ZN sputum images by offering a good DSC and segmentation accuracies of 92.19% and 98.64% respectively.

6. Conclusion

A novel deep learning-based technique for TB bacilli segmentation from sputum smear images has been developed, helping laboratory technicians and professionals in related fields to diagnose tuberculosis. We have used the TransUNet as it is highly recognized for its achievement in high efficiency for both local feature extraction using CNN and global feature extraction using transformers. Mean IoU, DSC, segmentation accuracy, precision, and recall are calculated to evaluate the proposed work, and the results are compared with the existing studies in terms of DSC and segmentation accuracy. Results prove that TransUNet outperforms the existing works and achieves a good accuracy even with unseen data. This indicates that the model performs very well with the sputum data, irrespective of the images' stain colors and different characteristics. A user-friendly interface is developed that simplifies uploading the sputum image and getting the results so that people with less technical knowledge can utilize this model. The model achieved a DSC of 94.16% and 99.74% accuracy on the mosaic dataset. In contrast, on the ZN sputum dataset, it achieved a DSC of 92.19% and 98.64% accuracy, which means it is performing good with ZN data even it is trained on mosaic images. Enhancements to the proposed work includes integrating explainable AI techniques like Grad-CAM, Occlusion sensitivity to the model, ensuring transparency in models decision making process there by gaining the trust of medical practitioners.

References

- [1] Mejbél, F. A. H., Aljanaby, I. A. J., & Aljanaby, A. A. J. (2023). Pulmonary tuberculosis risks and challenges. In E3S Web of Conferences (Vol. 381, p. 01101). EDP Sciences. J. Clerk Maxwell, *A Treatise on Electricity and Magnetism*, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
- [2] Chakaya, J., Khan, M., Ntoumi, F., Aklillu, E., Fatima, R., Mwaba, P., ... & Zumla, A. (2021). Global Tuberculosis Report (2020).—Reflections on the Global TB burden, treatment and prevention efforts. *International Journal of Infectious Diseases*, 113, S7-S12.
- [3] Ismail, Z., Musa, L. H., Muslimah Ithnin, N. A. M., Baharom, N., Ramli, S., & Pathmanathan, S. G. (2022). A Systematic Review of Prevalence and Risk Factors of Latent Tuberculosis Infection Among Medical and Nursing Students. *Malaysian Journal of Medicine and Health Sciences*, 18(6), 270-279.
- [4] Cesur, S. (2023). Clinical forms and diagnosis of tuberculosis. *Journal of Pulmonology and Intensive Care*, 1(1), 13-19.
- [5] Manalan, K., Barrett, J., & Kon, O. M. (2021). Tuberculosis clinical presentation and differential diagnosis. *Essential Tuberculosis*, 79-85.
- [6] Zhang, Y., Zhou, T., Wang, S., Liang, P., Zhang, Y., & Chen, D. Z. (2023, October). Input augmentation with sam: Boosting medical image segmentation with segmentation foundation model. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 129-139). Cham: Springer Nature Switzerland.
- [7] Costa Filho, C. F., Levy, P. C., Xavier, C. M., Costa, M. G., Fujimoto, L. B., & Salem, J. (2012, August). Mycobacterium tuberculosis recognition with conventional microscopy. In: 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp 6263–6268. IEEE.
- [8] Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., ... & Tao, D. (2022). A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 45(1), 87-110.
- [9] Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., ... & Farhan, L. (2021). Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *Journal of big Data*, 8, 1-74.
- [10] Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., ... & Zhou, Y. (2021). Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*.
- [11] Ali, O., Ali, H., Shah, S. A. A., & Shahzad, A. (2022). Implementation of a modified U-Net for medical image segmentation on edge devices. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 69(11), 4593-4597.
- [12] Wulandari, S. A., Mayasari, D. A., Kurniatie, M. D., & Tjahyono, R. (2021, September). Simplification of Mycobacterium tuberculosis segmenting algorithm in sputum images based on auto-thresholding. In: 2021 International Seminar on Application for Technology of Information and Communication (iSemantic), pp 12–16. IEEE.
- [13] Mithra, K. S., & Emmanuel, W. S. (2021). GFNN: Gaussian-Fuzzy-Neural network for diagnosis of tuberculosis using sputum smear microscopic images. *Journal of King Saud University-Computer and Information Sciences*, 33(9), 1084-1095.
- [14] Núñez-Fernández, D., Ballan, L., Jiménez-Avalos, G., Coronel, J., & Zimic, M. (2020). Automatic semantic segmentation for prediction of tuberculosis using lens-free microscopy images. *arXiv preprint arXiv:2007.02482*.
- [15] Panicker, R. O., Kalmady, K. S., Rajan, J., & Sabu, M. K. (2018). Automatic detection of tuberculosis bacilli from microscopic sputum smear images using deep learning methods. *Biocybernetics and Biomedical Engineering*, 38(3), 691-699.
- [16] Mithra, K. S., & Emmanuel, W. S. (2017, April). Segmentation and classification of mycobacterium from Ziehl Neelsen stained sputum images for tuberculosis diagnosis. In 2017 International Conference on Communication and Signal Processing (ICCSPP) (pp. 1672-1676). IEEE.
- [17] Shwetha, V., Prasad, K., Mukhopadhyay, C., Banerjee, B., & Chakrabarti, A. (2021, December). Automatic detection of bacilli bacteria from Ziehl-Neelsen sputum smear images. In 2021 2nd International Conference on Communication, Computing and Industry 4.0 (C2I4) (pp. 1-5). IEEE.
- [18] Mithra, K. S., & Emmanuel, W. S. (2018, December). Segmentation of Mycobacterium tuberculosis bacterium from ZN stained microscopic sputum images. In 2018 International Conference on Smart Systems and Inventive Technology (ICSSIT) (pp. 150-154). IEEE.
- [19] Ayas, S., & Ekinci, M. (2014). Automatic segmentation of Mycobacterium tuberculosis in Ziehl-Neelsen sputum slide images using support vector machines.

- [20] Xavier, C. M. (2012). Segmentação, classificação e quantificação de bacilos de tuberculose em imagens de baciloscopia de campo claro através do emprego de uma nova técnica de classificação de pixels utilizando máquinas de vetores de suporte.
- [21] Vanitha, V., Angayarkanni, S. P., & Sathishkumar, M. (2021, December). Segmentation of tuberculosis bacteria: Evaluation and comparison. In: 2021 International Conference on Computational Science and Computational Intelligence (CSCI), pp 1690–1696. IEEE.
- [22] Serrão, M. K. M., Costa, M. G. F., Fujimoto, L. B. M., & Ogusku, M. M. (2024). Automatic Bright-field Smear Microscopy for Diagnosis of Pulmonary Tuberculosis. Computers in Medicine and Biology, accepted for publication Feb. 15, 2024.
- [23] Misra, S., & Wu, Y. (2019). Machine learning assisted segmentation of scanning electron microscopy images of organic-rich shales with feature extraction and feature ranking. Machine learning for subsurface characterization, 289(3), 4.
- [24] Singhal, P., Verma, A., & Garg, A. (2017, January). A study in finding effectiveness of Gaussian blur filter over bilateral filter in natural scenes for graph based image segmentation. In 2017 4th international conference on advanced computing and communication systems (ICACCS) (pp. 1-6). IEEE.
- [25] MR, D. S. (2020). Comparison of Gaussian and Median Filters to Remove Noise in Dental Images. Institute of Scholars (InSc).
- [26] C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. Journal of big data, 6(1), 1-48.

Authors' Profiles



SCIE (WoS) or Scopus.

Ashutosh Satapathy is working in the Department of Computer Science and Engineering as an assistant professor (Selection Grade) at Siddhartha Academy of Higher Education, Deemed to be University, Vijayawada-520007, Andhra Pradesh, India. He completed his Ph.D. at Vellore Institute of Technology, India, and his Master of Technology (M. Tech) in Information Security and Computer Forensics at SRM University, India. During his Ph.D. program, he completed a research internship at Information and Communications Research Laboratories, ITRI-Taiwan. Before this, he was a lecturer at Odisha University of Technology and Research (formerly College of Engineering and Technology, Bhubaneswar). He has published 24 research articles, including international journal articles, book chapters, and conference proceedings, all indexed in



Praneeth Vallabhaneni is a computer science graduate student from Siddhartha Academy of Higher Education, India. He has been associated with many projects in Machine learning and Deep learning applications pertained to healthcare, exhibiting his diversified skills and technical knowledge. He has also published papers in IEEE conference proceedings, further proving his contributions to the area of research. He has also improved his skills by completing courses in Machine learning, Soft Computing, DevNet, Database Foundations. His dedication to continuous learning and development is perceptible in his involvement in several workshops and internships, where he gained skills in practice and valuable technical knowledge.



Manisha Indugula is a computer science graduate student from Siddhartha Academy of Higher Education, India. She has been involved in multiple Machine learning projects, showcasing her technical capabilities and skillset. Additionally, she has published research papers at Springer conference proceedings, reflecting her contributions to the field. To cap it all off, she has taken Cloud Computing Courses and Pega Systems courses, giving her a more applied knowledge and skills for the industry. She wants to use her knowledge and skills to solve global challenges by applying new developments in AI and Data Science technological advancements. She exhibits such commitment that drives her hopes in this field.

How to cite this paper: Ashutosh Satapathy, Praneeth Vallabhaneni, Manisha Indugula, "Semantic Segmentation of Tuberculosis Bacilli from Microscopic Sputum Smear Images Using TransUNet", International Journal of Image, Graphics and Signal Processing(IJIGSP), Vol.17, No.6, pp. 19-37, 2025. DOI:10.5815/ijigsp.2025.06.02