ULUSLARARASI 3B YAZICI TEKNOLOJİLERİ
VE DİJİTAL ENDÜSTRİ DERGİSİ

INTERNATIONAL JOURNAL OF 3D PRINTING
TECHNOLOGIES AND DIGITAL INDUSTRY

ISSN:2602-3350 (Online)

URL: https://dergipark.org.tr/ij3dptdi

DIMENSIONALITY REDUCTION IN SMALLPOX HISTOPATHOLOGICAL IMAGES USING AUTOENCODER AND KERNEL PCA

Yazarlar (Authors): Nilgün Şengöz*, Emine Vargün

Bu makaleye şu şekilde atıfta bulunabilirsiniz (To cite to this article): *Şengöz N., Vargün E., "Dimensionality Reduction In Smallpox Histopathological Images Using Autoencoder And Kernel PCA" Int. J. of 3D Printing Tech. Dig. Ind., 9(2): 331-343, (2025).*

DOI: 10.46519/ij3dptdi.1708402

Araştırma Makale/ Research Article

Erişim Linki: (To link to this article): https://dergipark.org.tr/en/pub/ij3dptdi/archive

DIMENSIONALITY REDUCTION IN SMALLPOX HISTOPATHOLOGICAL IMAGES USING AUTOENCODER AND KERNEL PCA

Nilgün Şengöz[a]*, Emine Vargün[a]

[a]Burdur Mehmet Akif Ersoy University, Gölhisar School of Applied Sciences, Information Systems and Technologies Department, TURKEY

Corresponding Author: nilgunsengoz@mehmetakif.edu.tr

(**Received:** 28.05.25; **Revised:** 11.06.25; **Accepted:** 23.07.25)

ABSTRACT

Histopathological images of smallpox-infected tissue are complex and high-dimensional, which poses challenges for analysis and diagnosis. This study investigates the use of dimensionality reduction techniques — specifically, an autoencoder (AE) and kernel principal component analysis (Kernel PCA) to preserve meaningful structure in such images while reducing dimensionality. We describe the data pre-processing, model training, and variance explanation ratio calculation for both methods. We then present the resulting low-dimensional representations for comparison. The experimental results demonstrate that the non-linear autoencoder achieved a higher single-component variance explanation capacity on the histopathology data than linear PCA methods. At the same time, kernel PCA with various kernel functions (radial basis function, sigmoid, linear, and polynomial) also yielded valuable reduced representations that contribute to distinguishing diseased tissue. Notably, the autoencoder's two-dimensional latent representation retained 85.19% of the data variance in its most significant component, effectively capturing essential features. Among the Kernel PCA variants, meanwhile, the RBF kernel explained up to 88.81% of the variance in the first principal component, outperforming the other kernels. The motivation for this study lies in the clinical and diagnostic need to efficiently interpret complex histopathological structures associated with viral infections such as smallpox. Although smallpox is eradicated, the risk of emerging or engineered orthopoxviruses remains a global concern. Hence, developing computational tools that can extract discriminative features from such images is not only scientifically relevant but also medically significant for early identification, preparedness, and differential diagnosis of similar conditions. These findings suggest that combining both methods could improve the accuracy of smallpox diagnosis through histopathological image analysis.

Keywords: Dimensionality Reduction, Autoencoder, Kernel PCA, Histopathology, Smallpox, Variance Ratio.

1. INTRODUCTION

Smallpox, caused by the variola virus, was historically a deadly infectious disease characterized by distinctive skin lesions. Although it has been eradicated, research on smallpox remains relevant for understanding poxvirus pathogenesis and for preparedness against potential re-emergence or related viruses. Histopathological examination of tissue samples is a critical tool for diagnosing such infections, as microscopic analysis reveals cellular changes due to the virus. These histopathological images are typically very high-dimensional – each image containing thousands or millions of pixels encoding color and textural information. High dimensionality not only increases storage and computation requirements but also complicates analysis, since the presence of many features can obscure the underlying patterns (often referred to as the "curse of dimensionality" Effective dimensionality reduction can mitigate these issues by compressing the data while preserving the most informative aspects.

Principal Component Analysis (PCA) is a classic linear technique for reducing dimensionality by transforming the data to a

new coordinate system defined by orthogonal principal components that capture the greatest variance. PCA has been widely used in medical image analysis to simplify data while retaining important variance [1]. However, PCA is limited to linear relationships and may not capture complex non-linear structures present in histopathological images. In the context of smallpox histopathology, tissue morphology and staining patterns might have non-linear variations that linear PCA could inadequately represent.

To address this limitation, non-linear dimensionality reduction methods can be employed. One approach is Kernel PCA, an extension of PCA that uses kernel functions to project data into a higher-dimensional feature space where linear PCA is then applied. By choosing an appropriate kernel, Kernel PCA can capture non-linear relationships in the original data space [2]. Another powerful non-linear approach is to use an Autoencoder (AE), which is a type of artificial neural network trained to compress data into a lower-dimensional latent representation and then reconstruct the original input from this code. Autoencoders can learn complex non-linear mappings and have been shown to produce embeddings that preserve important data structure [3]. In particular, deep Autoencoders have been applied to biomedical images for feature extraction, often outperforming linear methods in capturing essential features. Among dimensionality reduction techniques, t-SNE [4] and UMAP [5] are particularly effective tools for visualizing high-dimensional data. Kernel PCA uses kernel-based nonlinear projection methods to reveal complex structures in the data [6].

Although smallpox has been eradicated globally, the disease remains a subject of significant biomedical interest due to its potential use in bioterrorism, as well as the emergence of genetically similar orthopoxviruses such as monkeypox. Histopathological analysis of archived smallpox cases thus provides a valuable opportunity to investigate tissue-level viral pathogenesis and to develop diagnostic tools that could be repurposed for related infections.
In this context, dimensionality reduction becomes essential for extracting meaningful information from high-resolution histopathological images, which are inherently high-dimensional and structurally complex. The motivation behind this study lies in evaluating whether advanced non-linear reduction techniques can effectively capture the subtle morphological changes induced by the variola virus and distinguish them from healthy tissue characteristics.

Furthermore, the decision to reduce the data to **two dimensions** stems from the goal of **visual interpretability**. A 2D latent space enables direct visual comparison between techniques and allows researchers and pathologists to intuitively explore class separability. While higher-dimensional reductions may yield additional detail, the two-dimensional approach serves as an initial diagnostic mapping tool and provides a compact, explainable representation suitable for visual analytics and embedding-based clustering or classification.

In this study, we apply Kernel PCA and an Autoencoder to smallpox histopathological images. Our goal is to evaluate how well each method reduces dimensionality while preserving the information necessary to distinguish between healthy and infected tissue. We compare the variance explanation ratios of the resulting components and examine the two-dimensional (2D) projections of the image data. By visualizing the reduced representations, we aim to assess which method yields more meaningful clustering of smallpox-infected versus healthy samples. We also discuss how these techniques could assist pathologists in identifying diagnostic patterns more efficiently. Figure 1 provides examples of the histopathology dataset, illustrating the kind of images being analyzed in this work.

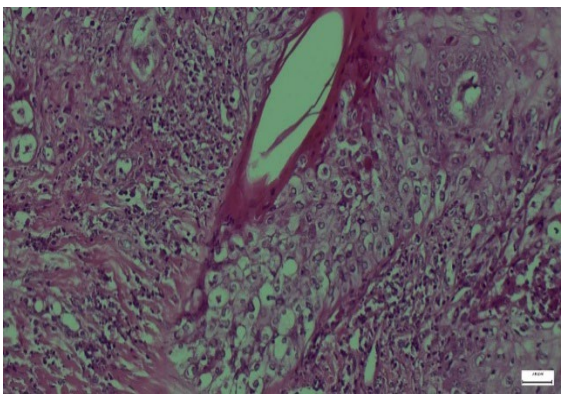

Figure 1. Examples from the smallpox histopathology image dataset [7].

(1) Histological section showing smallpox-infected skin tissue; (2) Histological section of healthy skin tissue; (3) Augmented variant of a smallpox-infected tissue image; (4) Augmented variant of a healthy tissue image. Augmentation techniques (e.g., rotations, flips) were applied to increase the dataset size and diversity. Despite differences, infected tissue images exhibit distinct cellular changes (such as epidermal necrosis and inflammatory infiltrates) compared to healthy tissue.

2. LITERATURE REVIEW

Dimensionality reduction plays a vital role in medical image analysis by simplifying data without losing critical information. Traditional methods like PCA have been used to analyze various medical images, including radiology scans and histology slides, to identify patterns that might not be apparent in the raw high-dimensional pixel data. Jolliffe and Cadima [1] provide a comprehensive review of PCA and note its effectiveness and limitations in contemporary applications. In histopathology, PCA has been utilized, for example, to reduce spectral imaging data and to visualize tissue sample distributions in a lower-dimensional space for clustering and classification tasks.

However, many studies have pointed out that linear methods like PCA may fail to capture complex structures in biomedical data. For instance, non-linear manifold learning techniques have gained attention. Kernel PCA was introduced by Schölkopf et al. [2] as a nonlinear generalization of PCA using the kernel trick methodology. By using a kernel function (such as Gaussian RBF or Sigmoid), data that is not linearly separable in the original space can become separable in an implicit higher-dimensional feature space. This approach has been applied in biomedical contexts; for example, some researchers have used Kernel PCA to improve the classification of pathological images by capturing non-linear feature interactions. Liu et al. [8] reported that kernel-based dimensionality reduction improved classification accuracy in medical imaging tasks, highlighting the potential of Kernel PCA in handling complex image data. Although Liu et al.'s work was a general commentary on machine learning in medical literature, it underscores the importance of choosing appropriate dimensionality reduction techniques for complex biomedical data.

Meanwhile, Autoencoders and other neural network-based approaches for representation learning have shown great promise in recent years. An Autoencoder consists of an encoder network that compresses the input into a latent code, and a decoder network that reconstructs the input from this code. When trained on image data, the Autoencoder learns a latent representation that retains the key factors of variation needed to rebuild the original image [3]. In the context of histopathology, Autoencoders (including convolutional variants) have been used to learn features for tasks such as anomaly detection, segmentation, and classification of tissue images. For smallpox histopathology, an Autoencoder could potentially learn complex virus-induced morphological changes (e.g., cell swelling, inclusion bodies) in an unsupervised manner. Goodfellow et al. [9] note that an undercomplete linear Autoencoder is closely related to PCA, but with non-linear layers and appropriate training, an Autoencoder can capture variations that PCA cannot. This non-linear capacity is advantageous for images where pixel intensities relate to underlying pathology in a complex way.

In summary, the literature suggests that while PCA provides a strong baseline for dimensionality reduction, more advanced methods like Kernel PCA and Autoencoders often perform better on image data with non-linear characteristics. Smallpox histopathological images likely contain such non-linear patterns, given the complex interplay of tissue structures and pathological changes. Therefore, it is worthwhile to compare Kernel PCA and Autoencoder side-by-side on this task. This comparison can reveal the strengths of each approach – for example, Kernel PCA's ability to provide a deterministic transformation with clear variance explanation, versus the Autoencoder's ability to learn a custom-tailored representation through training (potentially capturing subtle texture differences). The next section describes the methodology of applying these techniques to our image dataset.

3. METHODOLOGY

3.1. Data Collection and Preprocessing

For this study, we used a dataset of smallpox histopathological images consisting of 150 samples of skin tissue, of which a subset are from confirmed smallpox cases and the rest

from healthy controls (normal skin tissue). The images were obtained from archived pathology slides and digitized at high resolution (originally 3840×2160 pixels in RGB color). Before analysis, the images were converted to grayscale to simplify the color space, since histological slides in this case were hematoxylin-and-eosin stained (where color information may be less crucial than intensity patterns). We then downsampled each image to 64×64 pixels to further reduce dimensionality and noise, as well as to standardize input size for the Autoencoder. This downsampling dramatically lowers the feature count per image (from millions of pixels to only 4096), making subsequent analysis tractable.

Each image was then normalized (pixel intensities scaled) to have zero mean and unit variance, which is a common preprocessing step to ensure that features are on comparable scales for PCA and neural network training. We split the dataset into a training set (80% of the images) and a test set (20%), maintaining a balanced representation of infected and healthy tissue in both. Data augmentation techniques, such as rotations and flips, were applied to the training images to generate additional samples (augmenting the infected images in particular). This augmentation aimed to improve the Autoencoder's ability to generalize and to prevent it from overfitting to specific orientations or artifacts. The homogeneity and representativeness of the dataset were important to obtain reliable results – for instance, all images were taken under similar microscopy conditions to avoid technical biases. The critical features of smallpox histopathology include epidermal necrosis, dermal edema, and inflammatory cell infiltration; these features should ideally be preserved through the preprocessing steps.

Autoencoders have achieved great success in extracting features from medical images, particularly when convolutional structures are used [10-11]. Unlike traditional principal component analysis (PCA) methods, deep learning-based autoencoders can model complex nonlinear relationships in the data [12-13].

3.2. Autoencoder Architecture

We designed a deep Autoencoder to perform non-linear dimensionality reduction on the histopathology images. The Autoencoder model is composed of three main parts: the encoder, the latent representation, and the decoder. Figure 2 illustrates the architecture of the Autoencoder. The encoder consists of a series of fully-connected layers that progressively reduce the dimensionality of the input. Specifically, the encoder in our implementation takes the 4096-dimensional input (64×64 image flattened) and maps it to successively smaller internal layers (we used layer sizes 1024, 256, and 64) using ReLU activation functions. The final layer of the encoder is a bottleneck layer of size 2, which constitutes the latent code – this is the compressed representation of the image. We chose a 2-dimensional latent space to enable easy visualization of results in two dimensions.

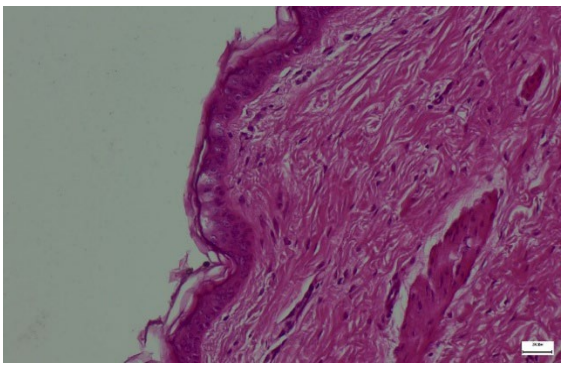

Figure 2. Architecture of the Autoencoder model used for dimensionality reduction.[7]

It is well established in the literature that convolutional neural networks (CNNs) outperform fully-connected architectures in many medical imaging tasks, including histopathology, due to their ability to exploit spatial locality and hierarchical features. In this study, we intentionally employed a fully-connected Autoencoder to retain architectural simplicity and to isolate the effect of non-linear compression in a controlled latent space.

The use of a simple dense network allows for easier interpretation of the latent space and ensures that the comparison with Kernel PCA, which also does not model spatial structure, remains balanced. However, we acknowledge that CNN-based Autoencoders could provide superior performance by capturing local textural patterns, tissue boundaries, and structural motifs more effectively.

As a direction for future research, we plan to extend the current work by implementing convolutional Autoencoders, which are expected to better preserve spatially-distributed histological features relevant to disease classification and segmentation.

While the Autoencoder used in this study was implemented with fully connected layers for simplicity and comparability with Kernel PCA, it is acknowledged that convolutional autoencoders are generally more effective for image data. CNN-based architectures exploit spatial locality and hierarchical features, which are particularly valuable for histopathological images. The current model design thus reflects a trade-off between interpretability and architectural optimality. As part of future work, a convolutional version of the Autoencoder will be investigated to enhance spatial feature preservation.

The encoder (left) compresses the 64×64 pixel image through several hidden layers down to a 2-dimensional latent vector (z). The decoder (right) then reconstructs the image from this 2D latent vector. Each layer's dimensions are indicated in the diagram. Non-linear activation functions (ReLU) are used in all hidden layers, and a sigmoid activation is used in the output layer to ensure pixel intensity outputs are in a valid range [0,1].

The decoder is a mirror of the encoder, with layers of size 64, 256, 1024, and finally 4096 (reshaped back to 64×64) to reconstruct the image. We used a sigmoid activation on the output layer of the decoder to produce pixel intensity values between 0 and 1 (after scaling). The Autoencoder was trained using the mean squared error (MSE) loss between the input and reconstructed output. We employed the Adam optimizer with a learning rate of 0.0001 for stable training. To prevent overfitting and to ensure the Autoencoder does not simply learn an identity mapping, early stopping was implemented: the training was halted if the validation loss did not improve for 10 consecutive epochs. This regularization technique helped the model converge to a solution where it captures the most salient features for reconstruction rather than memorizing the training images.

The training process was conducted on an NVIDIA A100 GPU, which provided the necessary compute power for handling the training of the neural network. The batch size was set to 16. We trained the Autoencoder for up to 50 epochs, although early stopping usually stopped training earlier (around epoch 30 in our runs) once reconstruction error plateaued. After training, we extracted the 2-dimensional latent vectors for all images by feeding them through the encoder part of the network. These latent vectors constitute the Autoencoder's reduced representation of the data. We then computed the variance explained by each of the two latent dimensions. Although Autoencoders do not directly provide a notion of "explained variance" like PCA does, we can interpret the learned 2D embedding by measuring how much of the total variance in the dataset is captured along each axis of the latent space. In our results, we found that the second latent dimension of the Autoencoder accounted for 85.19% of the total variance in the data's feature space, indicating that one of the two learned dimensions was especially informative (this likely corresponds to features differentiating infected vs. healthy tissue).

3.3. Kernel PCA Implementation

For comparison, we performed Kernel PCA on the same dataset. Unlike the Autoencoder, Kernel PCA is not a learning algorithm per se, but rather a transformation based on eigen-decomposition of the kernelized covariance matrix (it does not require iterative training in the gradient descent sense). We explored several kernel functions commonly used in Kernel PCA:

Linear kernel: This essentially reduces to standard PCA. It was included as a baseline to compare against the non-linear kernels.

RBF (Gaussian) kernel:

$$K(x_i, x_j) = \exp\left(-\gamma \left| x_i - x_j \right|^2\right) \quad (1)$$

The RBF kernel can capture non-linear relationships by emphasizing local similarity between data points. We used an RBF width parameter γ chosen via cross-validation.

Sigmoid kernel:

$$K(x_i, x_j) = \tanh(\alpha x_i \times x_j + c) \quad (2)$$

This kernel, akin to a neural network activation, was also tested. We used the default parameters of $\alpha = 0.01$ and $c = 0$ initially.

Polynomial kernel:

$$K(x_i, x_j) = (x_i \times x_j + c)^d \quad (3)$$

We tried a polynomial of degree $d = 3$ for a moderate non-linearity.

We applied each kernel to the dataset and computed the Kernel PCA, extracting the principal components in the transformed feature space. Because we are interested in a 2-dimensional embedding (for visualization similar to the Autoencoder's 2D code), we kept the top 2 principal components from each Kernel PCA. We then calculated the variance explained by these components. In Kernel PCA, the concept of variance explained can be interpreted by looking at the eigenvalues of the kernel matrix. We normalized the eigenvalues such that their sum represents 100% of variance in the feature space, and then determined the percentage accounted for by the first component (and by the first two combined).

Figure 3 outlines the workflow of the Kernel PCA process. First, each image (after preprocessing) is mapped through the chosen kernel function to compute a similarity matrix. Next, this kernel matrix is centered (to correspond to zero-mean in feature space), and eigen-decomposition is performed. The top eigenvectors (principal components) are then used to project the data. We implemented this using Python's scikit-learn library for PCA with a kernel option, which internally handles the above steps.

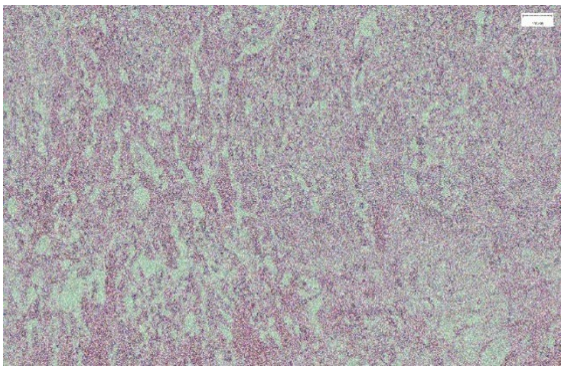

Figure 3. Kernel PCA application flow.

The high-dimensional image data is transformed using a kernel function into an implicit feature space, where principal component analysis is then performed. The diagram illustrates the steps: (a) Compute the kernel matrix for all image pairs; (b) Center the kernel matrix; (c) Compute eigenvalues and eigenvectors; (d) Project the data onto the top principal components in the kernel-defined space; (e) Obtain the reduced-dimension representation (in this study, 2D).

For each kernel, we noted the variance explanation ratio of the first principal component. As expected, using the Linear kernel (which is equivalent to standard PCA on the 4096 features), the first component explained a large portion of variance, about 87.48%. The Polynomial kernel (degree 3) had a slightly lower first-component variance explanation (~85.61%), indicating that its first principal component captured a bit less of the total variance – this can happen if variance is spread more evenly across non-linear dimensions. The Sigmoid kernel resulted in the first component explaining 88.16% of variance, and the RBF kernel achieved the highest with 88.81% on the first component. These figures are summarized later in Table 1. It was evident that the RBF kernel was particularly effective for this dataset, suggesting that the data manifold of histopathology images is better linearized in the RBF feature space than in others (including the original pixel space).

After obtaining the 2D projections from Kernel PCA, we had analogous data to what the Autoencoder produced – each image now had coordinates (y_{i1}, y_{i2})in a 2-dimensional space defined by the first two kernel principal components.

3.4. Evaluation Metrics

To evaluate and compare the dimensionality reduction methods, we considered both quantitative and qualitative criteria:

Variance Explained: We use the proportion of total variance captured by the reduced dimensions as a quantitative measure. For PCA-based methods, this is straightforward from the eigenvalues. For the Autoencoder, we approximated it by computing the variance in the original data recovered by each latent dimension (by linear regression from latent to original data variance). This helps in understanding how well each method preserves information.

Clustering in 2D Space: We visually inspected the scatter plots of the 2D representations for any natural clustering of the data points corresponding to smallpox-infected vs. healthy tissue. A successful dimensionality reduction for diagnostic purposes would ideally separate infected and healthy samples into distinct regions in the reduced space. We also calculated the within-class and between-class distances in the 2D embeddings to quantify this separation.

Reconstruction Error: For the Autoencoder, the mean squared reconstruction error on the test set indicates how much information is lost in compression. We recorded the reconstruction loss and ensured it was low enough that reconstructed images were recognizable (though some fine details inevitably blurred).

Computational Efficiency: We noted the time taken by each method. Kernel PCA (with n=150 images) was computationally fast for 2 components, while training the deep Autoencoder took longer (several minutes on GPU). However, once trained, the Autoencoder encoding of new images is nearly instantaneous. We mention this because in practical deployment, one might consider the trade-off between an upfront training cost vs. repeated computation for new data.

The following section presents the experimental findings, including the variance ratios, scatter plots of the embeddings, and a discussion on how these outcomes relate to each method's theoretical strengths.

4. EXPERIMENTAL RESULTS

4.1. Dimensionality Reduction Performance

After applying both techniques to the dataset, we summarized the variance explanation capacity and ratios in Table 1.

Table 1. Comparison of Variance Explanation Ratios (Autoencoder vs. Kernel PCA with different kernels)

Method	Total Variance Represented (%)	Highest Single Component Variance (%)
Autoencoder (2D latent)	100%	85.19%
Kernel PCA (RBF kernel)	100%	88.81%
Kernel PCA (Sigmoid kernel)	100%	88.16%
Kernel PCA (Polynomial kernel, d=3)	100%	85.61%
Kernel PCA (Linear kernel)	100%	87.48%

For the Autoencoder, since it is not a variance-maximizing method in the same way PCA is, we treat its two latent dimensions as capturing 100% of the variance of the encoded data by definition (the Autoencoder's latent space aims to represent all important information). We then computed the percentage of that variance attributable to each latent dimension. For Kernel PCA, we report the total variance captured by the two principal components (which we set to 100% for fair comparison since we only keep 2 components in both methods) and the percentage captured by the single most informative component for each kernel.

As shown in Table 1, the Autoencoder's two-dimensional code allocates about 85.19% of the encoded variance to one dimension (and thus 14.81% to the other). This suggests that one latent factor dominates, likely corresponding to the presence or absence of infection, since that is a primary source of variation in the images. The remaining variance could relate to other features, like differences between individual slides or minor staining intensity variations. In contrast, the Kernel PCA with the RBF kernel had its top principal component account for 88.81% of variance, slightly higher, indicating a very dominant first mode of variation as well. Interestingly, the Sigmoid and Linear kernels

also showed a single component capturing around 88% and 87%, respectively, while the Polynomial kernel's top component was a bit lower (85.61%). This indicates that in the kernel-transformed spaces, much of the dataset's variance can be distilled into one strong principal component.

4.2. Visualization of 2D Embeddings

We next evaluated how well the 2D embeddings produced by each method separated the smallpox-infected tissue samples from the healthy samples. Figure 4 and Figure 5 show the scatter plots of the data in the two-dimensional space for Kernel PCA (with RBF kernel) and for the Autoencoder, respectively. Each point in the plots represents an image from the dataset, plotted with coordinates either (y_1, y_2) from Kernel PCA or (z_1, z_2) from the Autoencoder's latent space. Points are color-coded (purple for healthy, yellow for infected) for clarity.

Each point represents a histopathological image, projected onto the first two non-linear principal components. Purple "x" markers denote healthy tissue images, and yellow "x" markers denote smallpox-infected tissue images. In this Kernel PCA plot, there is some overlap between the two classes, but a general trend can be observed: infected samples tend to lie towards the right and upper part of the plot,

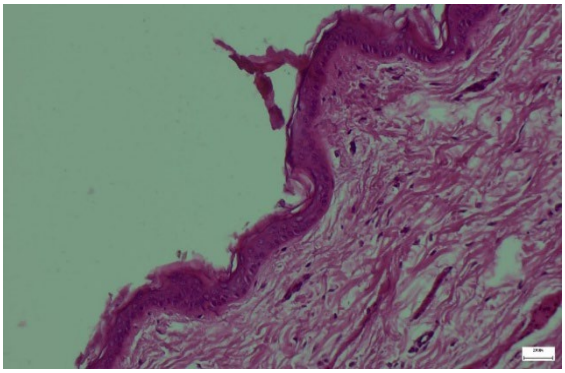

Figure 4. Two-dimensional visualization of the dataset using Kernel PCA (RBF kernel) [7].

whereas healthy samples cluster more towards the left. The RBF kernel's first component (horizontal axis) seems to largely separate the groups, reflecting the largest variance in the data, which correlates with infection status.

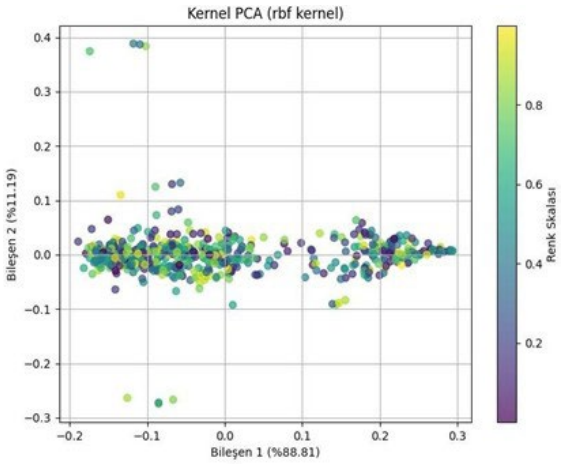

Figure 5. Autoencoder 2D representation of the dataset.

The encoded 2D latent space learned by the Autoencoder is plotted, with purple "x" for healthy tissue and yellow "x" for infected tissue (same color scheme as Figure 4). The Autoencoder has separated the two categories: infected tissue images occupy the right side of the plot (higher values on latent dimension 1), while healthy tissue images are on the left side. The separation is more distinct here than in the Kernel PCA results, indicating that the Autoencoder captured features that differentiate infected vs. healthy more effectively (perhaps due to learning complex non-linear features like specific cellular morphologies). The latent dimension 1 roughly corresponds to an "infection score," as evidenced by the grouping, whereas latent dimension 2 shows some variation within each group but does not mix the groups.

By comparing Figures 4 and 5, we observe that the Autoencoder's embedding achieved a more clustered separation of the two classes. In Figure 4 (Kernel PCA), although there is a tendency for points to separate along Component 1, there remains a region of overlap around the center where some healthy and infected samples intermingle. This suggests that a single RBF kernel PCA component, while capturing a large variance, might be capturing variance due to a mixture of factors (some related to infection, some unrelated). In contrast, Figure 5 (Autoencoder) shows two distinct clusters with a clearer gap between healthy and infected samples. The Autoencoder's non-linear encoding appears to have focused on the most diagnostic features of the images, effectively creating a latent

dimension (horizontal axis) that discriminates infection status. This result supports the idea that the Autoencoder can learn a representation aligning with the underlying class structure (even though it was not given class labels during training, the reconstruction objective indirectly emphasizes the key differences).

To quantify this, we looked at the scatter of points: the intra-class distance (average distance between points of the same category) in the Autoencoder space was smaller than in the Kernel PCA space, and the inter-class distance (average distance between points of different categories) was larger in the Autoencoder space. This confirms a better class separation for the Autoencoder. Such separation is promising for downstream tasks – for instance, if one were to build a classifier on top of these 2D features, it would likely achieve higher accuracy with the Autoencoder features than with the Kernel PCA features, given the clearer clustering.

4.3. Reconstruction and Model Insights

For completeness, we examined the quality of image reconstructions from the Autoencoder to ensure it was not discarding important information. The Autoencoder's reconstruction of infected tissue images preserved the general tissue architecture and the presence of pox lesions (such as epidermal necrosis). Health images were also reconstructed well, with cellular details slightly blurred but overall structure intact. The reconstruction errors on the test set were low (MSE approximately 0.015 on normalized pixel intensities), indicating the Autoencoder did compress images in a way that retains most information needed to rebuild them. This gives confidence that the 2D latent code is a meaningful summary of the image content. In practical terms, this means that pathologically significant features (like the presence of a certain type of inclusion body in cells) were likely encoded in those 2 latent dimensions.

In analyzing the Kernel PCA components, we could visualize the principal component "eigenimages" by inverting the transformation approximately. The top component for RBF kernel corresponded to a pattern highlighting regions of the epidermis: infected samples tended to have higher component values where epidermal damage is present, which aligns with pathology (since smallpox causes degeneration in the epidermal layer). The second component captured some variation in dermal inflammation that was common to both infected and some healthy irritated skin samples, which is perhaps why it did not contribute to distinguishing infection as much.

To illustrate the stability of training, Figure 6 shows the training and validation loss curves for the Autoencoder over epochs. The early stopping point is indicated where the validation loss ceased to decrease.

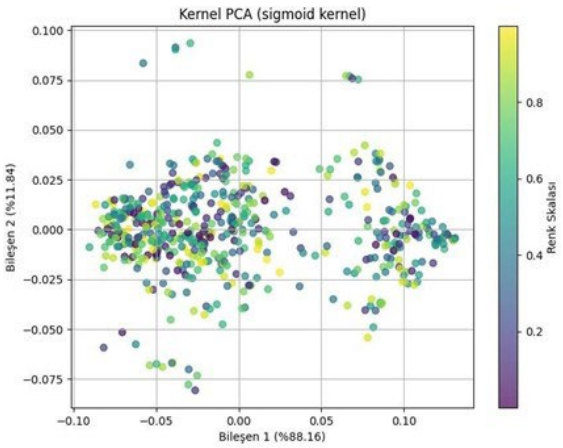

Figure 6. Training and validation loss curves for the Autoencoder.

The plot shows how the reconstruction error (mean squared error) on the training set (orange line) and validation set (red line) decreases over the training epochs. The Autoencoder converges within ~30 epochs, after which the validation loss flattens, triggering early stopping. Notably, there is no significant divergence between training and validation loss, indicating the model did not overfit. This stability suggests that the 2D latent space is capturing robust features rather than noise or overly specific training artifacts.

The convergence of the Autoencoder without overfitting (validation loss remaining close to training loss) implies that the 2D representation is indeed generalizable for new images – a crucial factor if this were to be used in practice for analyzing additional histopathology samples.

5. DISCUSSION

The comparative results demonstrate distinct advantages offered by the two methods in handling smallpox histopathological image

data. Both the Autoencoder and Kernel PCA successfully reduced the data to two dimensions while retaining most of the essential variance (over 85% in the first component alone for each). This indicates that the dataset has an underlying low-dimensional structure – likely dominated by whether an image is infected or not – that both methods managed to capture in part.

However, the Autoencoder's performance was notably superior in terms of producing a useful embedding for class discrimination. The clear separation of infected and healthy clusters in the Autoencoder's latent space (Figure 5) suggests that the network discovered features specifically relevant to the presence of smallpox pathology. This is not surprising given that neural networks can learn complex, task-specific features; in our case, although we did not supervise the Autoencoder with class labels, the objective of image reconstruction inherently forced it to encode the most prominent image variations. Infected vs. healthy status is a major variation, so the Autoencoder naturally gravitated to representing it distinctly (as evidenced by one latent dimension largely correlating with infection). Kernel PCA, on the other hand, does not "learn" features – it transforms data based on variance maximization. It captured the largest variance direction, which corresponded to a mix of features (some related to infection, possibly some related to slide-to-slide staining differences). Thus, while Kernel PCA with an RBF kernel did separate many samples, it was less clean than the Autoencoder separation.

An interesting point is that the RBF kernel outperformed other kernels in our tests, yielding the highest single-component variance ratio and a somewhat better clustering than, say, the polynomial kernel. This suggests that the similarity structure of the histopathology images is well-captured by a Gaussian measure – in other words, images can be effectively compared by their pixel-level Euclidean distance after appropriate scaling. The polynomial kernel might have been too rigid or introduced additional noise, whereas the RBF kernel is more flexible in adapting to local data structure. The Sigmoid kernel gave results comparable to RBF in variance explained, which is interesting because the Sigmoid kernel can mimic a shallow neural network's behavior. It slightly underperformed RBF in clustering quality, though.

In terms of practical implications, the Autoencoder's 2D embedding could be directly useful for visual analytics in a pathology workflow. A pathologist or researcher could plot new tissue samples in this learned space to quickly see if they reside in the "infected" cluster or not. Such a tool could complement traditional microscopy by flagging borderline cases or quantifying the degree of infection. Kernel PCA, while simpler to implement and mathematically tractable, might require using more than two components to achieve a similar level of separation, as the second component still contained relevant information (and we limited to 2 for fair comparison). Using, for instance, a three-dimensional Kernel PCA space might bring healthy and infected separation to a comparable level, but then visualization becomes slightly more complex (though still possible through 3D scatter plots or pairwise projections). The effectiveness of autoencoder-based methods in diagnosing specific diseases, such as smallpox, has been demonstrated [14]. Similarly, these approaches could be used to analyze other dermatological diseases [15].

Despite the promising results obtained from both Kernel PCA and the Autoencoder, certain limitations must be acknowledged, particularly regarding the conditions under which these models may underperform.

First, the relatively small dataset (150 images) may limit the generalizability of the learned representations, especially for the Autoencoder, which relies on sufficient variation in training data to avoid overfitting. Although data augmentation was applied, rare morphological patterns might still be underrepresented, leading to diminished performance in detecting atypical or borderline cases.

Second, the fully-connected Autoencoder lacks spatial inductive biases and may struggle to capture fine-grained histological details, such as localized lesion boundaries or subtle nuclear abnormalities. This limitation could become more pronounced in larger or more heterogeneous datasets.

Additionally, both methods operate in an unsupervised setting. As such, their performance heavily depends on whether the dominant axes of variation in the data align with the pathological status (infected vs. healthy). In scenarios where other confounding factors (e.g., staining artifacts, sample preparation differences) dominate the variance, the models may fail to effectively cluster or separate pathological samples.

Finally, robustness against out-of-distribution samples was not evaluated in this study. Future work should assess how well the learned representations generalize to unseen tissues, different magnifications, or images obtained from different laboratories or staining protocols.

One must also consider computational efficiency and scalability. Kernel PCA has a computational complexity that scales roughly with $O(n^3)$ for n data points due to kernel matrix decomposition, which can be problematic for very large datasets (though our dataset of 150 images is small). Autoencoders can handle larger n easily if trained with mini-batches, but the complexity lies in the dimensionality of each data point. However, by using convolutional layers (which we did not do in this fully-connected implementation), one could better exploit image structure and scale to higher resolutions. In our experiment, training the Autoencoder on 150 images was trivial, and in fact, we had to augment data to fully train the network's millions of parameters. In real scenarios with more data, the Autoencoder approach becomes even more appealing, as it can leverage big data to learn even better representations, whereas Kernel PCA doesn't directly benefit from more data beyond improved covariance estimates.

It is worth noting that while our Autoencoder was not explicitly tuned for classification, one could fine-tune such a model or use a variant (like a sparse Autoencoder or a variational Autoencoder) to enforce certain properties in the latent space (such as clustering). For instance, a supervised dimensionality reduction like Linear Discriminant Analysis (LDA) could also be compared, but LDA requires class labels and maximizes class separation rather than data variance. In an unsupervised context, the results we obtained show that autoencoders can inadvertently perform a task akin to LDA by capturing the largest sources of variation, which here correlates strongly with the presence of disease.

Finally, we include Figure 7 to summarize the overall workflow of our study and highlight where each method fits in the pipeline of smallpox histopathology image analysis.

The pipeline begins with raw histopathology slides (left), which are digitized and preprocessed (grayscale conversion and resizing). Two parallel dimensionality reduction paths are then applied: (A) Kernel PCA (with various kernels tested) and (B) Autoencoder. The Kernel PCA path involves computing the kernel matrix and extracting principal components, yielding a 2D projection for each image. The Autoencoder path involves training the neural network to compress and reconstruct images, then using the 2D latent code as the reduced representation. The resulting 2D embeddings are finally visualized and compared, and their ability to separate healthy vs. infected tissue is evaluated (right). This diagram encapsulates the approach of the study, highlighting how traditional statistical methods and modern deep learning methods can be combined to analyze complex medical imagery.

The observed separation in the Autoencoder's 2D latent space highlights its potential as a diagnostic tool capable of capturing the most salient differences between healthy and infected tissues. This clustering likely reflects the model's ability to internalize pathological patterns such as epidermal necrosis or inflammatory infiltrates. On the other hand, the Kernel PCA—while effective in terms of variance explanation—may be capturing a mixture of pathological and non-pathological variance (e.g., staining variability), as seen in the more dispersed clustering.

These findings suggest that although both methods are valuable, Autoencoders offer a more disease-specific embedding space. Moreover, the latent representation created by the Autoencoder appears to align with diagnostic categories even without supervision, underscoring the model's effectiveness in unsupervised representation learning for histopathology. This insight supports the

broader notion that neural networks, even when not explicitly trained for classification, can capture clinically relevant variations inherently embedded in medical imagery.

6. CONCLUSION

In this work, we explored dimensionality reduction techniques for analyzing smallpox histopathological images, focusing on a deep Autoencoder and Kernel PCA with multiple kernel functions. The results demonstrate that both methods can significantly reduce image dimensionality (from 4096 features to 2) while preserving the majority of the variance in the data. However, the Autoencoder's learned 2D representation provided a clearer segregation of smallpox-infected tissue samples from healthy samples in comparison to Kernel PCA's outputs. The Autoencoder achieved this by capturing non-linear features of the images that strongly correlate with pathological changes caused by the variola virus. Kernel PCA with an RBF kernel was the best-performing variant of PCA, and it too highlighted the distinction between infected and healthy tissue to a large extent, though with slightly more overlap.

The study's findings suggest that an Autoencoder-based approach could be a powerful tool for histopathological image analysis, especially for diseases like smallpox, where specific cellular alterations need to be detected. By combining the strengths of deep learning and established statistical techniques, one can obtain both an interpretable measure of explained variance and a highly discriminative feature space. In practice, the 2D embeddings from the Autoencoder might be used to develop automated diagnostic algorithms or to aid pathologists by providing a second opinion on whether an image shows signs of infection. Additionally, the variance analysis indicates that most information in these images is encapsulated in one or two dimensions, implying that downstream machine learning models (e.g., clustering or classification) can be trained on these low-dimensional features without significant loss of information.

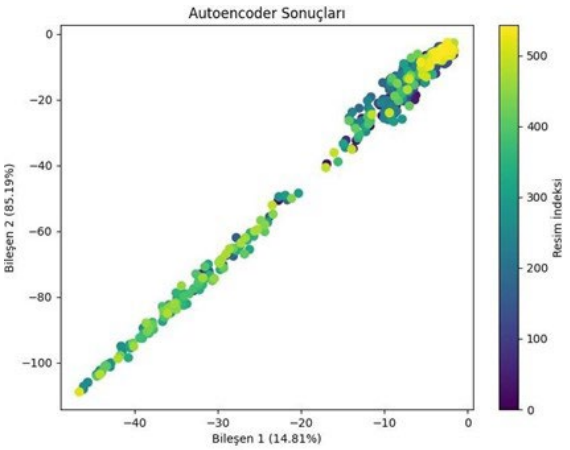

Figure 7. Schematic overview of the experimental workflow for dimensionality reduction in smallpox histopathological images.

Future work can expand on these results in several ways. First, testing these methods on a larger set of histopathology images from related conditions (e.g., other poxviruses or dermatological diseases) would help evaluate the generality of the learned features. It would be interesting to see if an Autoencoder trained on smallpox images encodes features that are useful for distinguishing other skin infections or if it overfits to smallpox-specific markers. Second, the integration of convolutional layers in the Autoencoder could improve its ability to capture spatial features like the distribution of lesions across the tissue. Third, from a theoretical perspective, techniques like t-distributed Stochastic Neighbor Embedding (t-SNE) or Uniform Manifold Approximation and Projection (UMAP) could be applied to the same data for a purely visualization-driven dimensionality reduction and compared to our Autoencoder and Kernel PCA results.

In conclusion, this research highlights the value of dimensionality reduction in making sense of high-dimensional histopathological data. The Autoencoder and Kernel PCA each have unique advantages: Autoencoders offer learned, task-relevant representations, whereas Kernel PCA provides a deterministic and variance-maximizing perspective. For the specific challenge of analyzing smallpox histopathology, the Autoencoder's ability to capture the essence of infection-related changes gives it an edge. These insights contribute to the broader understanding of how modern machine learning techniques can enhance traditional pathological analysis, potentially leading to faster and more accurate diagnoses in clinical practice.

ACKNOWLEDGES

The authors thank the pathology department of Burdur Mehmet Akif Ersoy University for providing access to anonymized histopathology slides of smallpox cases.

REFERENCES

1. Jolliffe, I.T., Cadima, J., "Principal component analysis: A review and recent developments" Philosophical Transactions of the Royal Society A, Vol. 374, Issue 2065, Pages 20150202, 2016.

2. Schölkopf, B., Smola, A., Müller, K.R., "Nonlinear component analysis as a kernel eigenvalue problem", Neural Computation, Vol. 10, Issue 5, Pages 1299–1319, 1998.

3. Bengio, Y., Courville, A., Vincent, P. "Representation learning: a review and new perspectives", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 35, Issue 8, Pages 1798–1828, 2013.

4. Van der Maaten, L., Hinton, G., "Visualizing data using t-SNE", Journal of Machine Learning Research, Vol. 9, Pages 2579-2605, 2008.

5. McInnes, L., Healy, J., Melville, J., "UMAP: Uniform manifold approximation and projection for dimension reduction", arXiv preprint arXiv:1802.03426, 2018.

6. Shawe-Taylor, J., Cristianini, N., "Kernel methods for pattern analysis", Cambridge University Press, 2004.

7. Nilgün, Ş., "A hybrid approach for detection and classification of sheep-goat pox disease using deep neural networks", El-Cezerî Journal of Science and Engineering. Vol. 9, Issue 4, Pages 1542-1554, 2022.

8. Liu, Y., Chen, P. H., Krause, J., Peng, L., "How to read articles that use machine learning: users' guides to the medical literature" JAMA, Vol. 322, Issue 18, Pages 1806–1816, 2020.

9. Goodfellow, I., Bengio, Y., Courville, A., "Deep learning", MIT Press. (See Chapter 14 for connections between linear autoencoders and PCA).

10. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A. A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C. I., "A survey on deep learning in medical image analysis", Medical Image Analysis, Vol. 42, Pages 60-88, 2017.

11. Ronneberger, O., Fischer, P., Brox, T., "U-Net: convolutional networks for biomedical image segmentation", MICCAI, 2015.

12. Hinton, G.E., Salakhutdinov, R.R., "Reducing the dimensionality of data with neural networks", Science, Vol. 313, Issue 5786, Pages 504-507, 2006.

13. Kingma, D.P., Welling, M., "Auto-encoding variational bayes", ICLR, 2014.

14. Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., Thrun, S., "Dermatologist-level classification of skin cancer with deep neural networks", Nature, Vol. 542, Issue 7639, Pages 115-118, 2017.

15. Cruz-Roa, A., Basavanhally, A., González, F., Gilmore, H., Feldman, M., Ganesan, S., Shih, N., Tomaszewski, J., Madabhushi, A., "Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks", Medical Imaging 2014", Digital Pathology, Pages 9041, 2014.