A Vector Quantized Autoencoder Model for the Interpretable Classification of Hepatic Ultrasound Textures

Fabrizio P. Mello[1], **Alexei M. C. Machado**[1,2]

[1]Programa de Pós-graduação em Informática – Pontifícia Universidade Católica de Minas Gerais (PUC Minas)
R Dom José Gaspar 500 – 30535-610 – Belo Horizonte – MG – Brasil
[2]Departamento de Anatomia e Imagem – Universidade Federal de Minas Gerais
Av Alfredo Balena 190 – 30130-100 – Belo Horizonte – MG – Brasil

fabrizioperagallo@gmail.com, alexeimcmachado@gmail.com

Abstract. *Texture analysis in liver ultrasound is central to steatosis assessment, as pathological alterations are primarily reflected in microtextural patterns. This work proposes a vector quantized variational autoencoder (VQ-VAE) for faithful reconstruction and interpretable latent representation of hepatic textures. Experiments were conducted on 550 image patches of the liver, normalized by the hepatorenal index, where the reconstruction quality and discriminative capacity of the latent space were evaluated. The model based on VQ-VAE architecture achieved the best overall performance among the fully reconstructed autoencoder models with respect to Mean Squared Error, Structural Similarity Index and Peak Signal-to-Noise Ratio. Moreover, a Support Vector Machine (SVM) trained on 32-dimensional latent vectors achieved the same accuracy as a SVM trained on raw pixels while reducing dimensionality from 784 to 32 features. These results demonstrate that VQ-VAE preserves microtextures, organizes the latent space in a structured manner, and produces compact, discriminative representations, highlighting its potential for quantitative and interpretable liver ultrasound analysis.*

1. Introduction

Texture analysis in medical imaging plays a fundamental role in the quantitative characterization of tissues and in the identification of subtle pathological patterns. In modalities widely used in clinical practice, such as ultrasound, the presence of speckle noise, acquisition artifacts, and limited spatial resolution makes it difficult to visualize microstructures relevant to diagnosis. In this context, texture-based approaches become particularly important, as they enable the exploration of granular variations that are often not fully captured by direct visual inspection.

The assessment of hepatic steatosis represents an emblematic example of this scenario. It is one of the most prevalent liver conditions in contemporary populations, associated with obesity and metabolic syndrome, whose manifestation in ultrasound occurs predominantly through alterations in echogenicity and in the textural organization of the parenchyma. Although the experimental domain of this work is steatosis, the challenges involved, such as microtexture preservation, blurring mitigation, and faithful representation of granular patterns, are shared by several medical imaging applications.

Models from the autoencoder family have been widely employed for image reconstruction, dimensionality reduction, and unsupervised learning. However, the literature still lacks systematic investigations that explore, in an integrated manner, three central aspects: texture preservation during reconstruction, controlled generation of plausible variations, and interpretable organization of the latent space. In many studies, the analysis is restricted to visual reconstruction quality or performance in supervised tasks, leaving open the question of how latent representations capture, structure, and render clinically relevant textural patterns interpretable.

Particularly in the context of hepatic ultrasound, few studies investigate generative models from the joint perspective of faithful reconstruction and interpretable latent space analysis. This gap is relevant because latent organization may offer not only efficient data compression, but also structured descriptors capable of supporting classification tasks and quantitative analysis.

In this work, we propose a method based on the Vector Quantized Variational autoencoder (VQ-VAE) for the reconstruction and interpretable analysis of hepatic textures in ultrasound images. The model is evaluated on a dataset composed of 550 regions of interest (ROIs) normalized by the hepatorenal index, enabling the investigation of its ability to preserve granular microstructures, organize textural patterns in the latent space, and concentrate discriminative information in compact representations. For comparative purposes, other autoencoder architectures are considered under the same experimental protocol, serving as methodological baselines.

Texture reconstruction is assessed with quantitative metrics such as the Mean Squared Error (MSE), Mean Absolute Error (MAE), Structural Similarity Index (SSIM) and Peak Signal-to-Noise Ratio (PSNR). A systematic latent space exploration through controlled image generation and the performance of a SVM classifier trained on the learned latent representations are analyzed for confirmatory purposes. The experiments demonstrate that the method reconstructs hepatic textures with high fidelity while organizing the latent space in a structured and discriminative manner.

2. Literature Review

The evaluation of liver diseases using ultrasound remains one of the most widely adopted approaches in clinical practice due to its non-invasive nature, accessibility, and broad availability. In particular, hepatic steatosis is frequently investigated by ultrasonography because of its high prevalence and association with metabolic disorders. However, visual interpretation of the examination is limited by the presence of speckle noise, interobserver variability, and the difficulty of quantifying subtle textural alterations. In this scenario, quantitative methods based on intensity and texture have been proposed as strategies to support diagnosis and severity stratification.

One of the most disseminated semiquantitative measures in the literature is the hepatorenal index (HI), which relates intensity statistics from the liver and renal cortex in order to reduce acquisition variability and enable reproducible estimates of hepatic fat presence. Clinical studies show that HI can be used for steatosis screening and stratification, although factors such as concomitant fibrosis and acquisition conditions may influence its performance [Marshall et al. 2012, Stahlschmidt et al. 2021, Johnson et al. 2021]. Recent works also investigate the automation of HI calculation

through artificial intelligence methods, aiming to reduce dependence on manual ROI selection and increase reproducibility [Zsombor et al. 2023]. These approaches highlight the relevance of quantitative descriptors in hepatic ultrasound, but also reinforce the need for richer representations capable of capturing granular textural patterns beyond simple intensity statistics.

In parallel, a consolidated research line employs texture analysis and machine learning to classify normal and steatotic livers. Studies based on classical features and statistical descriptors extracted from hepatic regions demonstrate that texture can contain relevant discriminative information in ultrasound, provided that preprocessing and feature selection are carefully conducted [Owjimehr et al. 2015]. More recently, deep learning and transfer learning methods have dominated the state of the art, leveraging pre-trained convolutional networks for feature extraction and steatosis classification in B-mode images, often under controlled normalization and clinical evaluation protocols [Byra et al. 2018, Constantinescu et al. 2021]. Clinical reviews also indicate a growing trend toward quantitative ultrasound and the integration of multiple markers to improve diagnostic robustness [Fetzer et al. 2023].

Although effective for classification, many of these strategies prioritize predictive performance and interpretability at the level of features or attention mechanisms, but offer limited capability for image reconstruction and generation. This limitation restricts complementary analyses such as textural variability and systematic inspection of latent factors. In this context, generative models become relevant because they enable learning compact and potentially interpretable representations from unlabeled data, in addition to allowing controlled reconstruction and image synthesis.

Autoencoder-based models have been widely adopted for reconstruction, dimensionality reduction and unsupervised learning in medical imaging. Convolutional autoencoders (CAEs) demonstrate strong capacity to capture local patterns [Chen et al. 2021]. However, due to their deterministic nature and the typical use of mean-error-based losses, they tend to produce smoothed reconstructions, which may compromise the preservation of high-frequency details in noisy modalities. The variational autoencoder (VAE), introduced by [Kingma and Welling 2014], incorporates probabilistic modeling and enables the generation of new samples, although the assumption of simple latent distributions is often associated with edge and microstructure blurring, a phenomenon particularly critical when texture constitutes the primary discriminative signal.

Extensions with Normalizing flows (NF) [Rezende and Mohamed 2015] increase the expressiveness of latent distributions by employing invertible transformations, potentially modeling more complex variations. However, in real biomedical scenarios, their computational cost and sensitivity to hyperparameters may hinder systematic adoption, especially when stable generation and interpretable latent organization are desired.

The Vector-Quantized Variational Autoencoder (VQ-VAE), introduced by [van den Oord et al. 2017], employs a discrete latent space structured by codebooks, reducing the smoothing tendency typical of continuous latents. Extensions such as VQ-VAE-2 [Razavi et al. 2019] demonstrate that hierarchical codebook structures favor more detailed reconstructions, while recent applications in local synthesis and inpainting indicate that vector quantization can recover textural patterns with higher

fidelity [Liang et al. 2024]. Moreover, latent discretization may facilitate interpretable analysis by inducing organization through learned prototypes, which is particularly relevant when reconstruction, generation, and latent structure are intended to be integrated within a single framework.

Finally, recent self-supervised methods such as Masked Autoencoders (MAEs) [He et al. 2022] have achieved strong performance in representation learning through masked region reconstruction, capturing global dependencies. Despite their promise, studies specifically investigating texture preservation in ultrasound and latent space interpretability under integrated quantitative and qualitative criteria remain scarce.

Overall, despite the advancement of deep techniques for steatosis detection and grading in ultrasound [Byra et al. 2018, Constantinescu et al. 2021], a gap persists in the literature regarding the applied use of generative models to integrate texture preservation in reconstruction, plausible generation, and interpretable latent space organization. Recent studies also highlight the relevance of generative and self-supervised models for texture analysis in medical imaging [Gomide and Machado 2025], reinforcing the opportunity to investigate architectures that simultaneously provide textural fidelity and compact representations useful for auxiliary tasks. In this context, the present work focuses on the VQ-VAE as the central model for hepatic ultrasound analysis, aiming to address this gap through an applied evaluation that connects reconstruction, latent structure, and classification utility.

3. Vector Quantized Variational Autoencoder

The VQ-VAE is a generative model that combines convolutional encoding with a discrete latent space structured by a codebook. Unlike variational models with continuous latent variables, the VQ-VAE replaces Gaussian sampling with a vector quantization process, in which each latent vector produced by the encoder is mapped to the closest element of a finite set of learned embeddings.

Formally, given a continuous latent feature map generated by the encoder, each vector is quantized to its nearest codebook entry according to a distance metric, e.g. the Euclidean Distance, resulting in a discrete latent representation. The decoder reconstructs the original image from this quantized structure. By relying on discrete prototypes rather than smooth Gaussian samples, this mechanism mitigates the blurring effects typically observed in continuous latent models.

In the context of hepatic ultrasound, parenchymal texture is characterized by granular, recurrent patterns that are highly sensitive to excessive smoothing. Continuous latent models often prioritize global intensity statistics at the expense of high-frequency details, which are critical for representing subtle microtextural alterations associated with steatosis. The discrete latent space of the VQ-VAE is particularly suitable for this scenario, as it induces the model to represent images as combinations of learned texture prototypes. This structure favors the preservation of local variations and contrast patterns, leading to sharper and structurally coherent reconstructions.

Beyond reconstruction fidelity, the VQ-VAE also enables interpretable latent analysis. Because quantized vectors correspond to specific codebook entries, controlled variations in latent indices can be directly associated with visual changes in reconstructed

textures. This property allows for a systematic investigation of latent organization and its relationship to patterns observed in healthy and steatotic livers.

In this work, quantized latent maps are aggregated into representative vectors for each ROI, producing compact low-dimensional descriptors. These vectors are subsequently used as input to an SVM classifier to assess whether the learned structure concentrates discriminative information. Thus, the VQ-VAE is employed not only as a reconstruction model, but also as a mechanism for extracting an interpretable latent representation with potential applicability as a quantitative support tool for clinical liver ultrasound analysis.

4. Materials

The dataset used in the experiments is the B-mode liver ultrasound dataset introduced in [Byra et al. 2018]. The original study involved 55 patients with severe obesity (mean age of 40.1 ± 9.1 years; mean body mass index of 45.9 ± 5.6), evaluated during preoperative preparation for bariatric surgery. Images were acquired at the Medical University of Warsaw, Poland, using a GE Vivid E9 system operating at 2.5 MHz.

For each patient, 10 consecutive images were selected along a cardiac cycle, resulting in a total of 550 images with a resolution of 434×636 pixels and 8 bits per pixel. All patients underwent liver biopsy during the surgical procedure, and histopathological classification was performed by a single pathologist. The liver was considered steatotic when more than 5% of hepatocytes presented lipid accumulation, resulting in 17 patients classified as healthy and 38 as steatotic.

In order to reduce acquisition variability and analyze hepatic textures in a standardized manner, a procedure based on the hepatorenal index (HI) was adopted, as commonly used in ultrasound steatosis assessment. For each image, two square ROIs of 28×28 pixels were manually selected: one in the hepatic parenchyma and another in the renal cortex, avoiding vessels and artifacts. The hepatorenal index was computed as $\mathrm{HI} = \mu_{\mathrm{L}}/\mu_{\mathrm{K}}$, where μ_{L} and μ_{K} correspond to the mean intensity values respectively in the liver and kidney. The hepatic ROI was then normalized by multiplying its pixel values by HI and rescaling to the range [0,255], reducing differences related to gain and imaging depth. After normalization, only the hepatic ROI was retained for analysis, resulting in a dataset of 550 ROIs with 28×28 pixels. Fig. 1 presents examples of normalized ROIs from healthy and steatotic patients.

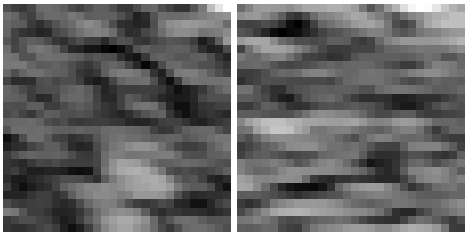

Figure 1. Examples of normalized hepatic ROIs used as input to the models showing a healthy subject (a) and a steatotic patient (b).

5. Methods

The general architecture of the model used in this study, based on the VQ-VAE, is shown in Fig. 2. The implemented model consists of three main modules: an encoder, a vector quantization layer, and a decoder. The encoder is a CNN that maps an input ROI $x \in [0,1]^{1\times28\times28}$ to a continuous latent representation $z_e(x) \in R^{32\times7\times7}$ through two convolutional layers with stride 2. The vector quantization module discretizes this latent map by replacing each 32-dimensional latent vector with its nearest neighbor from a codebook composed of 64 learned embeddings, producing the quantized latent representation $z_q(x)$. A straight-through estimator is employed to enable gradient propagation during training. The decoder mirrors the encoder using two transposed convolutional layers to reconstruct the image from $z_q(x)$, restoring the original spatial resolution of 28×28. The latent space is therefore structured as a discrete codebook of 64 vectors of dimension 32, providing a compact and organized representation for hepatic texture reconstruction and analysis.

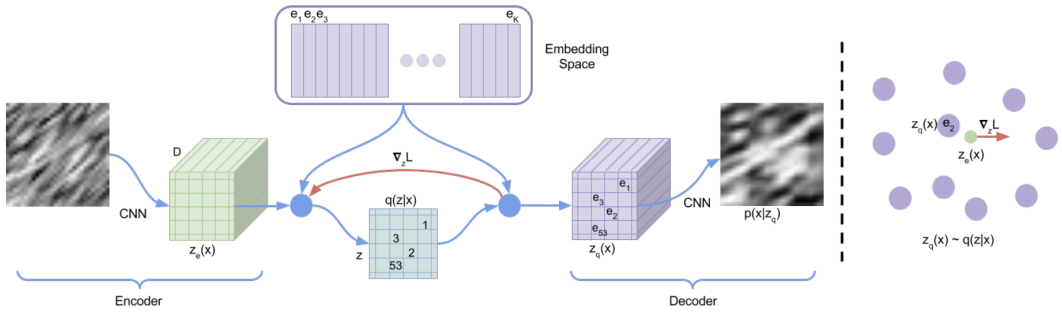

Figure 2. Architecture of the VQ-VAE model, highlighting the vector quantization process through a discrete codebook. Adapted from [van den Oord et al. 2017].

The VQ-VAE loss function consists of three main components: a reconstruction term $\mathcal{L}_{rec}$, a codebook term $\mathcal{L}_{rec}$, and a commitment term $\mathcal{L}_{com}$:

$$\mathcal{L} = \mathcal{L}_{rec} + \mathcal{L}_{cb} + \beta\mathcal{L}_{com},$$

where

$$\mathcal{L}_{rec} = \|x - \hat{x}\|_2^2, \qquad \mathcal{L}_{cb} = \|\mathrm{sg}[z_e(x)] - z_q(x)\|_2^2, \qquad \mathcal{L}_{com} = \|z_e(x) - \mathrm{sg}[z_q(x)]\|_2^2.$$

the reconstruction loss is computed as the mean squared error between the input ROI x and its reconstruction $\hat{x}$; $\mathrm{sg}[\cdot]$ is the stop-gradient operator, $z_e(x)$ is the encoder output, $z_q(x)$ is the quantized latent representation, and $\beta = 0.25$ is the commitment cost hyperparameter.

5.1. Evaluation Protocol

Reconstruction quality was assessed using MSE, MAE, SSIM and PSNR metrics, computed for each ROI. These metrics quantify pixel-wise error, mean absolute difference, structural similarity and perceptual reconstruction quality.

Beyond reconstruction assessment, the utility of the latent space for classification was investigated. The quantized latent maps of the VQ-VAE were aggregated into

representative vectors of 32 features per ROI. A SVM classifier was trained on these vectors and compared with a SVM trained directly on the ROI pixels (784 features). A support vector machine was adopted due to its robustness in small-sample scenarios and its widespread use in texture analysis studies. In particular, SVM-based classifiers have been frequently applied to ultrasound texture characterization, including studies on hepatic steatosis such as [Byra et al. 2018]. Since the objective of this stage was to evaluate the discriminative capacity of the learned latent representation rather than to optimize a deep classifier, the use of SVM provides a stable and well-established baseline while avoiding the additional trainable parameters and potential overfitting associated with multilayer perceptrons.

Accuracy, sensitivity, specificity, and F1-score were computed, along with confusion matrices. This analysis enabled the evaluation of whether the learned latent space concentrated discriminative information relevant for distinguishing healthy and steatotic livers, complementing the visual and quantitative reconstruction assessment. The performance of the proposed model was compared to a convolutional autoencoder (CAE), a variational autoencoder (VAE), a VAE with normalizing flows (VAE+NF) and a masked autoencoder (MAE), all trained under the same experimental protocol using the same normalized ROIs.

6. Experimental Results

The dataset consisted of the 550 normalized 28×28 pixel hepatic ROIs described in the Materials section. For CAE, VAE, VAE+NF and VQ-VAE, images were used in the range $[0, 1]$, whereas the MAE employed rescaling to $[-1, 1]$ to explicitly represent masked regions. The ROIs were split into training and testing sets (80%/20%), preserving the proportion of healthy and steatotic patients, and keeping the images of each subject either in the training or in the testing sets in order to avoid data leakage.

All methods were implemented in PyTorch and executed on a Windows 11 workstation equipped with a 14th-generation Intel Core i7 processor, 64 GB of RAM and an NVIDIA GeForce RTX 4070 GPU (12 GB VRAM). The architectures were trained with the Adam optimizer for 250 epochs with batch size of 64. Hyperparameters were determined through preliminary experiments performed exclusively on the training subset, evaluated based on reconstruction stability and consistency across quantitative metrics (MSE, SSIM). No data augmentation techniques were employed, due to the small ROI size and the objective of preserving the original texture distribution, which also reinforces the need for careful interpretation of potential overfitting effects. The final configurations were defined as follows: CAE with learning rate (α) of 2×10^{-4}; VAE with $\beta = 1.0$ and $\alpha = 10^{-4}$; VAE+NF with $\beta = 0.01$, $\alpha = 5 \times 10^{-5}$ and gradient clipping (maximum norm 5.0); VQ-VAE with a 64-entry codebook of 32-dimensional vectors and $\alpha = 2 \times 10^{-4}$; and MAE with 4×4 patch embedding, dimension 256, a six-layer four-head Transformer encoder, $\alpha = 10^{-4}$ and a 30% masking ratio.

6.1. Reconstruction Comparison and Latent Space Analysis

Table 1 reports the mean and standard deviation of the MSE, MAE, SSIM and PSNR metrics computed over the 110 ROIs of the validation subset for each evaluated architecture. Overall, the VQ-VAE achieved lower reconstruction error and higher structural similarity than the convolutional autoencoder baselines, standing out as an effective architecture

for preserving hepatic microtextures under the evaluated conditions. VAE and VAE+NF produced stronger smoothing and lower structural similarity, reflecting limitations in retaining high-frequency details under the adopted probabilistic regularization.

The MAE presentslower reconstruction errors and visually sharper reconstructions. However, this behavior was influenced by the masking strategy used during training, where only 30% of the pixels were predicted while the remaining portion of the image was directly observed. This relatively low masking ratio was adopted due to the small spatial resolution of the ROIs (28×28) and the limited dataset size. Higher masking ratios would reduce the available spatial context, that makes the reconstruction task considerably more challenging for such small images. As a result, a large part of the original structure is preserved during reconstruction, favoring pixel-wise metrics, but limiting the direct comparability with models such as VQ-VAE that reconstruct the entire image from a fully compressed latent representation.

Table 1. Mean±standard deviation of reconstruction metrics (MSE, MAE, SSIM and PSNR) computed over the 110 ROIs of the validation subset for each evaluated model. The p-values of the statistical hypothesis tests comparing the proposed VQ-VAE (shown in bold) and the CAE baseline are also reported.

Model	MSE (mean ± std)	MAE (mean ± std)	SSIM (mean ± std)	PSNR (mean ± std)
CAE	0.0208 ± 0.0047	0.1157 ± 0.0122	0.4076 ± 0.0684	16.90 ± 0.88
VAE	0.0324 ± 0.0048	0.1341 ± 0.0142	0.3822 ± 0.0713	14.82 ± 0.74
VAE+NF	0.0301 ± 0.0042	0.1268 ± 0.0133	0.4015 ± 0.0674	15.42 ± 0.76
MAE	0.0068 ± 0.0021	0.0615 ± 0.0100	0.9514 ± 0.0130	21.86 ± 1.27
Proposed	$\mathbf{0.0131 \pm 0.0025}$	$\mathbf{0.0909 \pm 0.0093}$	$\mathbf{0.6764 \pm 0.0496}$	$\mathbf{18.92 \pm 0.85}$
p-value	< 0.001	< 0.001	< 0.001	< 0.001

For a joint qualitative analysis, Fig. 3 shows reconstructions of four reference ROIs, including two healthy samples and two steatotic samples. Visually, the MAE reconstructions may appear sharper in some regions due to the partial observation of the input during training.

In contrast, the VQ-VAE reconstructs the entire image from a fully compressed and discretized latent representation, without direct pixel preservation. Although visual differences may appear subtle in individual samples, the quantitative metrics in Table 1 indicate that the proposed method achieves lower reconstruction error and higher structural similarity than the convolutional autoencoder baselines. The differences become more evident when considering the global behavior across all ROIs, where the VQ-VAE consistently outperforms CAE, VAE and VAE+NF across all evaluated metrics.

One of the most important aspects of the proposed method based on VQ-VAE, its potential for interpretability, is highlighted in Fig. 4. The figure shows 5 patches artificially generated from the learned latent space of the codebook during the unsupervised training process. The 5 patches were generated by using the mean values of the 32 latent variables and varying just one of them, from -2σ to $+2\sigma$. It is visually possible to interpret this chosen latent variable as a change in the contrast between the granular patterns. This generative procedure was therefore effective to produce plausible samples, consistent with the hepatic texture observed in the dataset, preserving granularity and local contrast. This behavior suggests that the codebook is able to encode learned recur-

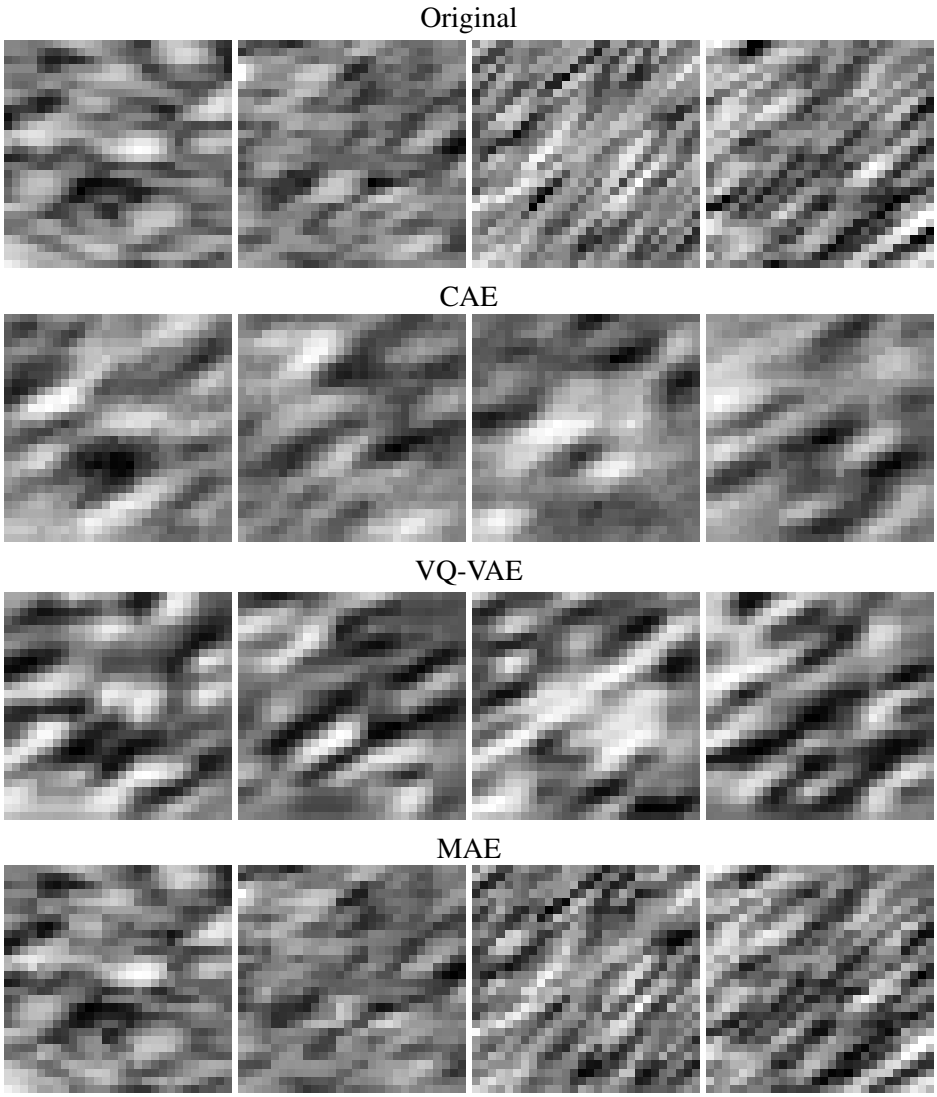

Figure 3. Reconstructions of four reference ROIs: original images and reconstructions produced by CAE, VQ-VAE and MAE. The first two columns correspond to healthy liver ROIs, while the last two columns correspond to steatotic liver ROIs.

rent textural prototypes, enabling latent manipulation while maintaining high-frequency characteristics.

6.2. Classification based on the VQ-VAE Latent Space

In order to assess the discriminative potential of the learned representations, two SVM classifiers were compared: one trained directly on ROI pixels (784 features) and another trained on the mean VQ-VAE latent vectors (32 features). The classifiers were trained using the 110 ROIs from the validation subset, which were further split into training and testing partitions (80/30 samples). Table 2 presents the resulting confusion matrices, where rows represent true classes and columns represent predictions.

The SVM trained on pixels achieved an accuracy of 0.70, with sensitivity of 0.95 for the steatosis class and specificity of 0.20 for the healthy class (F1-score = 0.81 for steatosis). In contrast, the SVM trained on the VQ-VAE latent space achieved the same

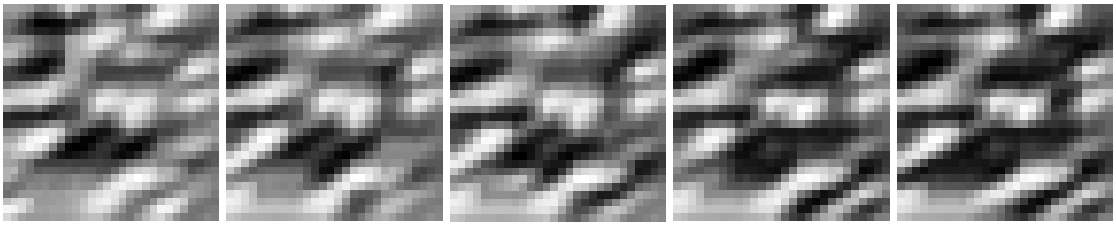

Figure 4. VQ-VAE latent space exploration. Columns represent variations from -2σ to $+2\sigma$ around the latent mean.

Table 2. Confusion matrices of SVM classifiers trained on ROI pixels and on mean VQ-VAE latent vectors (test set with 30 ROIs).

	SVM on pixels		SVM on latents	
	Healthy	Steatosis	Healthy	Steatosis
Healthy	2	8	6	4
Steatosis	1	19	5	15

accuracy of 0.70, with sensitivity of 0.75 and specificity of 0.60 (F1-score = 0.77 for steatosis).

Although the pixel-based classifier obtained higher sensitivity for steatosis, it exhibited a strong bias toward the diseased class, with low specificity for healthy samples. The latent-based classifier, in contrast, produced a more balanced performance between classes while operating on a substantially more compact representation (32 features instead of 784). These results indicate that the VQ-VAE latent space retains relevant discriminative information while providing a structured and interpretable representation.

7. Conclusion and Future Work

This work presented and evaluated a vector quantized variational autoencoder as a central model for reconstruction and interpretable analysis of hepatic textures, combined with a SVM classifier for assessing discriminative capacity. Using a dataset composed of 550 normalized ROIs from a limited number of patients, experiments investigated the model's ability to preserve relevant granular microstructures, organize the latent space in a structured manner, and produce compact descriptors useful for auxiliary classification tasks. To contextualize the results, additional architectures were used as references under the same experimental protocol.

The experiments showed that the VQ-VAE provides a favorable trade-off between reconstruction fidelity and latent organization, producing sharper reconstructions with higher structural similarity. This characteristic is particularly important in ultrasound, where high-frequency textural patterns are sensitive to smoothing. Controlled exploration of the latent space generated plausible samples while preserving granularity and local contrast consistent with the textures observed in the dataset, suggesting that the discrete codebook learned textural prototypes that are meaningful for the domain.

Beyond reconstruction, the VQ-VAE demonstrated quantitative relevance by concentrating discriminative information in a low-dimensional representation. The SVM classifier trained directly on ROIs of 784 pixels achieved an accuracy of 0.70, whereas

the SVM trained on mean VQ-VAE latent vectors of 32 features also achieved an accuracy of 0.70, with a more balanced confusion matrix between healthy and steatotic livers. This result indicates that the learned latent space compresses and organizes information in an interpretable form for separating clinical texture patterns. Overall, these findings reinforce vector quantization as a promising alternative to integrate faithful reconstruction, compact representation and interpretable latent organization, particularly when compared to approaches optimized primarily for reconstruction metrics. The work therefore contributes to reducing a gap in the literature regarding the use of generative models oriented towards texture preservation and interpretability, providing evidence that VQ-VAE can support quantitative analyses that are complementary to the clinical assessment of steatosis.

A natural continuation of this work may further investigate latent space interpretability by associating latent components with clinically meaningful textural attributes, such as granularity, echogenicity and heterogeneity, through sensitivity analyses and controlled code manipulation. Another relevant direction is to expand and diversify the data, incorporating larger ROIs, higher resolution, and datasets acquired with different devices and protocols, as well as multiple degrees of steatosis, in order to evaluate generalization and robustness. Additionally, it is promising to investigate out-of-distribution detection scenarios, exploring reconstruction error and density measures in latent space to identify patterns not observed during training, potentially associated with severe acquisition variations or other pathologies. Finally, integrating latent representations with supervised strategies, including fine-tuning and comparisons with deep classifiers, may further consolidate the utility of the VQ-VAE as a component of quantitative decision-support systems for hepatic ultrasound diagnosis.

Acknowledgments — This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001, and by Fundação de Amparo à Pesquisa de Minas Gerais (FAPEMIG) grants APQ-02753-24 and APQ-06556-24.

References

Byra, M. et al. (2018). Transfer learning with deep convolutional neural network for liver steatosis assessment in ultrasound images. *International Journal of Computer Assisted Radiology and Surgery*, 13:1895–1903.

Chen, M., Shi, X., Zhang, Y., Wu, D., and Guizani, M. (2021). Deep feature learning for medical image analysis with convolutional autoencoder neural network. *IEEE Transactions on Big Data*, 7(4):750–758.

Constantinescu, E. C., Udriștoiu, A.-L., Udriștoiu, C., Iacob, A. V., Gruionu, L. G., Gruionu, G., Săndulescu, L., and Săftoiu, A. (2021). Transfer learning with pre-trained deep convolutional neural networks for the automatic assessment of liver steatosis in ultrasound images. *Medical ultrasonography*, 23(2):135–139.

Fetzer, D. T., Pierce, T. T., Robbin, M. L., Cloutier, G., Mufti, A., Hall, T. J., Chauhan, A., Kubale, R., and Tang, A. (2023). Us quantification of liver fat: past, present, and future. *Radiographics*, 43(7):e220178.

Gomide, L. C. and Machado, A. M. C. (2025). Classificação de texturas em imagens médicas através de modelos generativos e aprendizado autossupervisionado. In *Anais do XXV Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS 2025)*, pages 967–972, Porto Alegre, RS, Brasil. SBC.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. (2022). Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15979–15988.

Johnson, S. I., Fort, D., Shortt, K. J., Therapondos, G., Galliano, G. E., Nguyen, T., and Bluth, E. I. (2021). Ultrasound stratification of hepatic steatosis using hepatorenal index. *Diagnostics*, 11(8):1443.

Kingma, D. P. and Welling, M. (2014). Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*.

Liang, Y. et al. (2024). Hierarchical vector-quantized variational autoencoder and vector credibility mechanism for high-quality image inpainting. *Electronics*, 13(10):1852.

Marshall, R. H., Eissa, M., Bluth, E. I., Gulotta, P. M., and Davis, N. K. (2012). Hepatorenal index as an accurate, simple, and effective tool in screening for steatosis. *American journal of roentgenology*, 199(5):997–1002.

Owjimehr, M., Danyali, H., and Helfroush, M. S. (2015). An improved method for liver diseases detection by ultrasound image analysis. *Journal of Medical Signals & Sensors*, 5(1):21–29.

Razavi, A., van den Oord, A., and Vinyals, O. (2019). Generating diverse high-fidelity images with vq-vae-2. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Rezende, D. and Mohamed, S. (2015). Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pages 1530–1538.

Stahlschmidt, F. L., Tafarel, J. R., Menini-Stahlschmidt, C. M., and Baena, C. P. (2021). Hepatorenal index for grading liver steatosis with concomitant fibrosis. *PLoS One*, 16(2):e0246837.

van den Oord, A., Vinyals, O., and Kavukcuoglu, K. (2017). Neural discrete representation learning. In *Neural Information Processing Systems*, pages 1–10, Long Beach.

Zsombor, Z., Rónaszéki, A. D., Csongrády, B., Stollmayer, R., Budai, B. K., Folhoffer, A., Kalina, I., Győri, G., Bérczi, V., Maurovich-Horvat, P., et al. (2023). Evaluation of artificial intelligence-calculated hepatorenal index for diagnosing mild and moderate hepatic steatosis in non-alcoholic fatty liver disease. *Medicina*, 59(3):469.