Adding Hierarchy to a Stacked Vector-Quantized Variation Autoencoder for Semi-Supervised White Blood Cell Classification

Amirali Goodarzvand Chegini
amiralic@student.ubc.ca

Jerome Cho
jeromejj@student.ubc.ca

Domenic Mastromatteo
dmastro@student.ubc.ca

Abstract

Recent advances in unsupervised learning with Vector-Quantized Variational Autoencoders (VQ-VAEs) have demonstrated strong performance in medical image classification tasks by leveraging large amounts of unlabeled data. Building on the Stacked Vector-Quantized Variational Autoencoder (SVQVAE) architecture proposed by Wang (n.d.) for white blood cell (WBC) classification, this study investigates the benefits and trade-offs of integrating a hierarchical VQ-VAE architecture within the existing SVQVAE pipeline. Drawing inspiration from the multi-level discrete latent structure of VQ-VAE-2, we augment the original stacked VAE model with an additional VQ-VAE module to form a hybrid hierarchical-stacked architecture. This modification aims to capture richer latent representations by encoding both high-level and fine-grained features. We retrain the original model for reproducibility and evaluate our extended architecture on the Raabin-WBC dataset using the classification metrics of the original work. Our findings indicate that the hierarchical extension can yield improvements in classification performance, even in low-data contexts, with trade-offs in model complexity and dimensionality. This work contributes a novel method for efficient and accurate image classification in environments with scarce labeled data.

1 Introduction

Automated medical image analysis can significantly improve diagnostic workflows by reducing reliance on manual annotation and enabling scalable analysis of large datasets. In the context of hematology (the study of blood disorders), accurate classification of white blood cells (WBCs) plays a crucial role in diagnosing a wide range of health conditions. However, high-performance classification models are often constrained by the limited availability of cable medical image, and large volumes of unlabeled data remains underutilized.

To address this challenge, Wang (n.d.) proposed a Stacked Vector-Quantized Variational Autoencoder (SVQVAE) framework that leverages self-supervised learning to perform unsupervised pretraining on unlabeled blood cell images. This model compresses high-dimensional inputs into a very compact latent representation, before finetuning the model using a separate dataset for classification. The SVQVAE showed strong performance in the classification of WBC types, even in experiments where labeled data was scarce.

While effective, the SVQVAE's architecture may be limited in its ability to capture hierarchical relationships in complex medical images due to the strength of its compression. In this project, we draw inspiration from the VQ-VAE-2 framework and explore whether augmenting the SVQVAE with

Course project for UBC CPSC 440/550, April 2025.

an additional hierarchical vector-quantized autoencoder can enhance the model's performance with a moderate sacrifice of efficiency. Hierarchical VQ-VAEs similar to the proposed model are known for capturing both high-level and fine-grained information by leveraging multiple levels of latent abstraction, a property we hypothesized may further improve classification accuracy.

Our research provides two contributions. First, we replicate the original SVQVAE pipeline to validate its performance and establish a baseline, detailing specific implementation steps for replication that were absent from Wang's manuscript. Second, we design and implement a novel hybrid architecture that combines the strengths of stacked and hierarchical VQ-VAEs, and evaluate its effectiveness in improving classification on the Raabin-WBC dataset. This work aims to detail trade-offs between latent representation complexity and downstream task performance in semi-supervised image classification.

2 Related Work

Our research was grounded in a previous unpublished manuscript: Stacked Vector-Quantized Variational Autoencoders for Unsupervised Pretraining and Classification of White Blood Cells by Wang (n.d.). This manuscript introduced the SVQVAE architecture which leverages VQ-VAEs for unsupervised representation learning on stained cell images. The model compresses images into a compact latent space, enabling effective downstream classification with minimal labeled data. While this approach demonstrates strong performance, it is limited in its ability to model hierarchical dependencies in image features due to the linear stacking of VQ-VAEs.

Generating Diverse High-Fidelity Images with VQ-VAE-2 (Razavi, Oord, and Vinyals 2019) proposes a hierarchical extension to VQ-VAE, introducing multiple levels of discrete latent variables to capture both global and local image features. This model achieved impressive performance in generative tasks, particularly image synthesis. While their primary focus was on generation rather than classification, the hierarchical design demonstrated the potential of extracting richer latent representations. Our work adapts this idea for classification in our semi-supervised learning setting, extending the SVQVAE to incorporate hierarchical latent spaces to with the goal of improving blood cell classification performance.

A common issue in image classification, particularly in the medical field, is the scarcity of high-quality labeled data. The original manuscript by Wang addressed this issue by borrowing ideas from Self-supervised Learning as a Means to Reduce the Need for Labeled Data in Medical Image Analysis (Benčević et al. 2022), which addresses this issue by pretraining on a subset of the dataset which does not contain labels, then finetuning on the remainder of the dataset. This approach drastically reduced the need for labeled data, and is adapted for use in our white blood cell context by pretraining on the pRCC and CAM16 datasets, and performing finetuning and classification on the Raabin-WBC dataset.

A number of other papers have covered variational autoencoders and provided background knowledge or inspiration for this research. Autoencoders (Bank, Koenigstein, and Giryes 2021) covered this topic in the greatest breadth, providing a comprehensive overview of autoencoder architectures, including classical, variational, and discrete variants like VQ-VAE. They highlight the use of autoencoders in dimensionality reduction and feature extraction, both key features of the model used in this report. He et al. (2021) presented a masked image modeling strategy in Masked Autoencoders Are Scalable Vision Learners by pretraining vision transformers (ViTs) using randomly masked inputs. While shown to be effective in large-scale settings, the method relies on expensive transformer architectures and high compute resources. Our approach, by contrast, offers a more lightweight solution using convolutional VQ-based encoders, making it far less resource intensive.

3 Model Description

3.1 Datasets

Datasets used were as identical as possible to the original paper, which was necessary for both replication and evaluation of our new method. A variety of datasets were used to train and finetune the model. The pRCC (Gao et al. 2021) and CAM16 (Litjens et al. 2018) datasets were used as unlabeled data used for pretraining, while the Raadbin-WBC (Kouzehkanan et al. 2022) was used

for finetuning and evaluation. The WBC dataset was further broken down into subsets WBC100, WBC50, WBC10, and WBC1, each of which represents 100%, 50%, 10%, and 1% of the original dataset respectively. Only 1418 images were used for pRCC while 1419 were used in the original manuscript, as we found one of the images to be corrupted.

Table 1: WBC_100 Dataset Distribution for Training and Validation Images

Class	**Training Images**	**Validation Images**	**Total**
Basophil	176	36	212
Eosinophils	618	126	744
Lymphocyte	2015	412	2427
Monocyte	466	95	561
Neutrophil	5172	1059	6231
Total	8447	1728	10175

Table 2: CAM16 and pRCC Dataset Distribution

Class	**CAM16**			**pRCC**
	Training Images	**Validation Images**	**Test Images**	**Training Images**
Normal	379	54	108	1418
Tumor	378	54	108	
Total	757	108	216	1418

Image generation for the CAM16 dataset required broad assumptions to be made due to a scarcity of description in Wang's manuscript. PNG images for the CAM16 dataset were creating by cropping 3 tumor TIFF files and 3 normal (non-tumor) TIFF files of the CAM16 dataset. Feature-rich images were then automatically selected from cropped images by removing images with a standard deviation and Laplacian variance below necessary thresholds (10 and 20 respectively). Afterwards, images where the majority of the image consisted of stained cells were selected manually in the quantities described in the Table 2, achieving a 70/20/10 train-test-validation split.

3.2 Model

The Stacked Vector Quantized Variational Autoencoder (SVQVAE) architecture from the original manuscript built upon the VQ-VAE framework by composing multiple quantization stages to progressively compress images into compact latent representations. Each VQ-VAE encodes the input into a lower-dimensional space using vector quantization and reconstructs it via a learned codebook of embeddings. In the SVQVAE setup, three such VQ-VAE modules are stacked, where the output of one module serves as the input to the next. This enables the model to transform a 512×512×3 input image into a compact latent representation of size $2k \times 8 \times 8$ (where k is the dimension of each quantization vector in the codebook), capturing increasingly abstract and compressed features at each stage. These compressed features are then used for downstream classification, with the goal of enabling learning on a much smaller, yet information-rich latent space. The architecture from the original manuscript is displayed in Figure 1.

Building on this architecture, we introduce a dual-stack SVQVAE system designed to further refine and disentangle the information captured at different levels of abstraction. After the original three-level SVQVAE produces its final $2k \times 8 \times 8$ latent encoding, we upscale this encoding using bilinear interpolation to match the spatial resolution of the first level in the original stack. We then concatenate this upscaled representation with the encoding from the first VQ-VAE level, effectively combining coarse global context from the deepest level with fine-grained local details from the initial encoding. This fused representation is then passed as input to a second, independently trained SVQVAE consisting of two stacked VQ-VAEs. The training of the two stacks is performed sequentially: Stack 1 is first fully trained, and only afterward is Stack 2 trained on top of its frozen outputs. This approach encourages Stack 2 to learn the missing details not captured by Stack 1, resulting in a complementary latent space that enhances the overall representation. The second stack compresses this enriched feature set into a latent representation of the same size as the original stack's final output, namely $2k \times 8 \times 8$. This model's architecture is displayed in Figure 2.

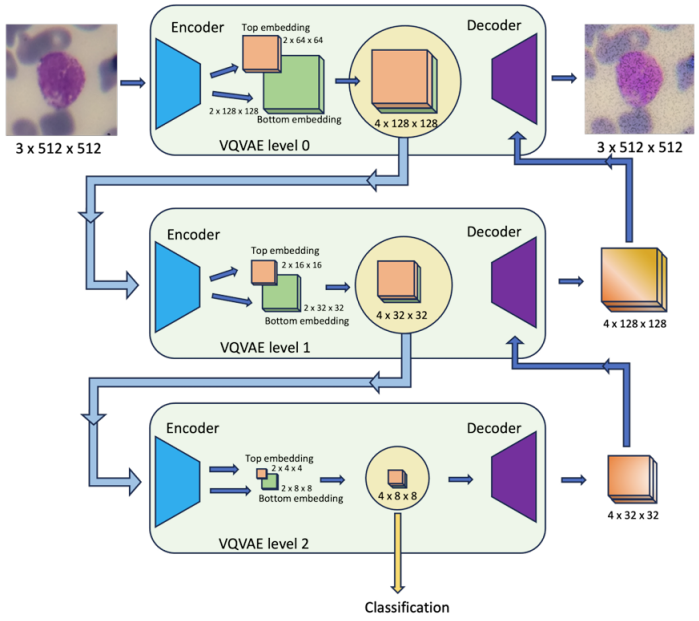

Figure 1: Original model architecture, with input image dimension of 512x512x3 (Wang n.d.)

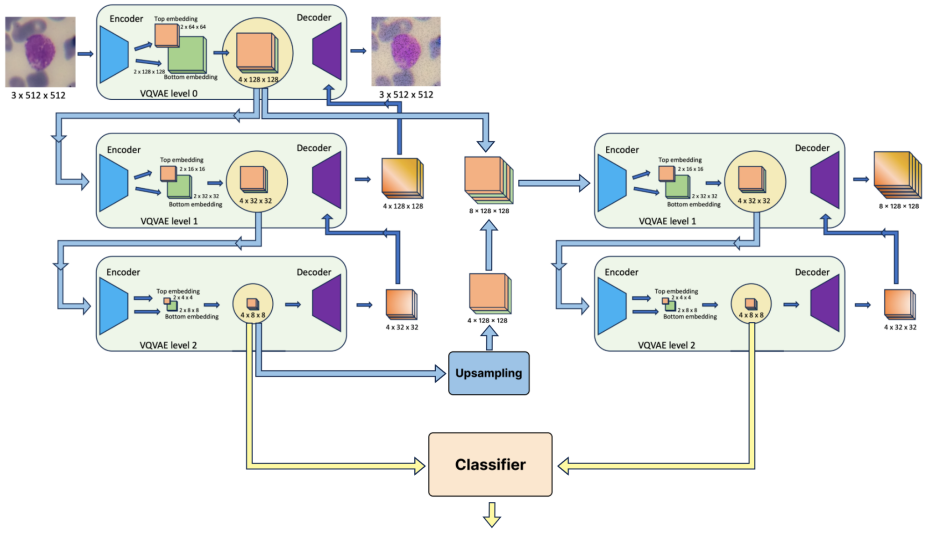

Figure 2: Updated dual-SQVAE architecture, with input image dimension of 512x512x3

For classification, the outputs from both the original and the second SVQVAE stacks are concatenated, resulting in a feature vector of size $4k \times 8 \times 8$. The classifier is correspondingly expanded to accommodate this larger input and trained to perform the final task. This dual-stack design draws inspiration from U-Net architectures, in the sense that high-resolution and low-resolution feature representations are merged to capture both global structure and local detail. The second stack is encouraged to specialize in capturing finer, complementary information not already extracted by the first, effectively promoting a division of labor between the two encoding pathways. This hierarchical, detail-aware compression scheme aims to enhance the expressive power of the latent space for improved classification performance.

4 Experimentation and Results

4.1 Reproduction

To establish a performance baseline, we first reproduced the original SVQVAE model proposed by Wang (n.d.). Following the methodology outlined in the manuscript, the model was pretrained on the unlabeled pRCC and CAM16 datasets using the specified architecture: three stacked VQ-VAE modules, each with 128 encoder/decoder channels, three residual blocks per encoder and decoder with 64 channels each, a codebook with 512 entries of dimension 2, and a codebook exponential decay factor of 0.99. The original manuscript used 2 residual blocks per encoder, with 32 channels instead, which the departure from was a reproduction error on our end. As the goal was merely to provide a baseline for our updated model, and as training the model was costly, we proceeded with our reproduction. The original model contained approximately 3.1 million parameters while our reproduction contained approximately 6.2 million.

Training was conducted using an AdamW optimizer with a learning rate of 1.5×10^{-4}, weight decay of 0.05, and β parameters (0.9, 0.95). A cosine annealing learning rate scheduler was used during pretraining. Contrary to the original protocol of 50 epochs, we pretrained the model for 60 epochs but maintained all other hyperparameters. Finetuning was performed on the labeled WBC datasets for 100 epochs, using the same optimizer and learning rate but without a scheduler.

Overall, reproduction of the original results was largely successful. On the WBC100, WBC50, and WBC10 subsets, our reproduced accuracies closely matched those reported in the original work. However, performance on the WBC1 dataset, which contains only a single training example per class, was noticeably lower. We attribute this discrepancy to variability in the random sampling process when constructing WBC1, which may have strongly influenced our outcome due to the sparsity of some blood cell classes.

Table 3: Original Training and Testing Accuracy Results

Dataset	WBC 100	WBC 50	WBC 10	WBC 1
Training Accuracy (best)	97.7%	97.0%	97.9%	98.7%
Testing Accuracy (best)	97.3%	95.8%	94.4%	88.4%

Despite minor variations, particularly on WBC1, the reproduced results validate the general effectiveness and reproducibility of the SVQVAE pipeline under the same experimental conditions.

Table 4: Reproduction Training and Testing Accuracy Results

Dataset	WBC 100	WBC 50	WBC 10	WBC 1
Training Accuracy (best)	97.0%	96.0%	95.3%	90.0%
Testing Accuracy (best)	97.3%	96.7%	94.4%	75.1%

4.2 Novel results

We next evaluated our proposed extension: a dual-stack hierarchical SVQVAE architecture. After pretraining the original SVQVAE stack, we bilinearly upscaled the final latent encoding and concatenated it with the level-1 encoding. This fused representation was passed through a second, independently trained two-level SVQVAE module. Finally, outputs from both the original and secondary stacks were concatenated before classification.

Training of the dual-stack model followed the same procedure as the original model, with no changes to optimization settings or dataset usage. While the increase in model complexity—approximately doubling the number of parameters from 6.2M to 11.4M—approximately doubled the pretraining time required, the model remained computationally manageable and achieved improved performance across all datasets, indicating the extra complexity was useful in this context.

Compared to the reproduced baseline, the dual-stack model achieved higher test accuracies across all WBC datasets. Most notably, the model demonstrated substantial gains on the extremely low-data

Table 5: Hierarchical Model Training and Testing Accuracy Results

Dataset	WBC 100	WBC 50	WBC 10	WBC 1
Training Accuracy (best)	96.8%	97.3%	96.0%	87.7%
Testing Accuracy (best)	97.7%	97.3%	95.3%	81.1%

WBC1 subset, improving testing accuracy from 75.1% to 81.1%. These results support the hypothesis that introducing hierarchical structure into the latent space enables more effective feature extraction, even under severe constraints to labeled data.

5 Discussion

5.1 Conclusion

The performance improvements observed with the dual-stack SVQVAE indicate that enriching the latent space with multi-scale feature information can meaningfully enhance classification outcomes without the need for additional labeled data. By merging fine-grained and high-level representations before further compression, the second stack likely captures complementary features that the original SVQVAE alone could not fully represent.

However, this improvement comes at a cost. The dual-stack architecture roughly doubles the number of parameters of the original model, and the two-stage training results in a doubling of pretraining time.

Overall, our experimental results suggest that hierarchical stacking of VQ-VAEs offers a promising direction for improving semi-supervised learning in domains characterized by large amounts of unlabeled data and limited labeled examples, such as medical imaging.

5.2 Extensions

The current architecture demonstrates how a secondary SVQVAE stack, informed by both high-level and low-level encodings from the original stack, can be used to enhance feature representation for downstream classification. However, several promising extensions remain to be explored that may offer further insight into the utility of the second stack.

One direction involves simplifying the classification pipeline by retaining the classifier used in the original SVQVAE and training it using only the final encodings produced by the second stack. This would allow for a direct comparison between the representational power of the original three-level stack and the new, U-Net–inspired two-level stack. If the second stack, trained with access to multi-scale features from the first stack, yields comparable or better classification results, this would support the hypothesis that it acts as a superior feature extractor, even without the need for larger classifier capacity.

Another, more targeted, extension involves designing a custom encoder and decoder for the first level of the second stack to better specialize it for extracting residual or fine-grained information. In this setup, instead of performing an explicit upscaling and concatenation of encodings, the encoder would take two inputs: the level 1 encoding and the level 3 encoding from the original stack. While it only encodes the level 1 input, its operations would be conditioned on the level 3 encoding, allowing it to selectively refine low-level features based on the high level context. Correspondingly, a custom decoder would attempt to reconstruct the original level 1 encoding using both the latent encoding from the custom encoder and the level 3 encoding as additional context. This encourages the second stack to focus specifically on capturing complementary details that may have been lost in the coarse compression of the first stack, acting as a detail-aware refinement mechanism.

These extensions offer pathways to evaluate and enhance the modular design of stacked VQ-VAE systems for classification, with the ultimate goal of balancing model complexity with performance.

Acknowledgments

We gratefully acknowledge the support provided by the University of British Columbia's SOCKEYE computing cluster.

References

Dor Bank, Noam Koenigstein, and Raja Giryes (Apr. 2021). *Autoencoders*. arXiv:2003.05991. DOI: 10.48550/arXiv.2003.05991. URL: http://arxiv.org/abs/2003.05991 (visited on 04/25/2025).

Marin Benčević, Marija Habijan, Irena Galić, and Aleksandra Pizurica (June 2022). *Self-Supervised Learning as a Means To Reduce the Need for Labeled Data in Medical Image Analysis*. arXiv:2206.00344. DOI: 10.48550/arXiv.2206.00344. URL: http://arxiv.org/abs/2206.00344 (visited on 04/25/2025).

Zeyu Gao, Bangyang Hong, Xianli Zhang, Yang Li, Chang Jia, Jialun Wu, Chunbao Wang, Deyu Meng, and Chen Li (June 2021). *Instance-based Vision Transformer for Subtyping of Papillary Renal Cell Carcinoma in Histopathological Image*. arXiv:2106.12265. DOI: 10.48550/arXiv.2106.12265. URL: http://arxiv.org/abs/2106.12265 (visited on 04/25/2025).

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick (Dec. 2021). *Masked Autoencoders Are Scalable Vision Learners*. arXiv:2111.06377. DOI: 10.48550/arXiv.2111.06377. URL: http://arxiv.org/abs/2111.06377 (visited on 04/25/2025).

Zahra Mousavi Kouzehkanan, Sepehr Saghari, Sajad Tavakoli, Peyman Rostami, Mohammadjavad Abaszadeh, Farzaneh Mirzadeh, Esmaeil Shahabi Satlsar, Maryam Gheidishahran, Fatemeh Gorgi, Saeed Mohammadi, and Reshad Hosseini (Jan. 2022). "A large dataset of white blood cells containing cell locations and types, along with segmented nuclei and cytoplasm." en. *Scientific Reports* 12.1, p. 1123. ISSN: 2045-2322. DOI: 10.1038/s41598-021-04426-x. URL: https://www.nature.com/articles/s41598-021-04426-x (visited on 04/25/2025).

Geert Litjens, Peter Bandi, Babak Ehteshami Bejnordi, Oscar Geessink, Maschenka Balkenhol, Peter Bult, Altuna Halilovic, Meyke Hermsen, Rob Van De Loo, Rob Vogels, Quirine F Manson, Nikolas Stathonikos, Alexi Baidoshvili, Paul Van Diest, Carla Wauters, Marcory Van Dijk, and Jeroen Van Der Laak (June 2018). "1399 H&E-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset." en. *GigaScience* 7.6, giy065. ISSN: 2047-217X. DOI: 10.1093/gigascience/giy065. URL: https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giy065/5026175 (visited on 04/25/2025).

Ali Razavi, Aaron van den Oord, and Oriol Vinyals (June 2019). *Generating diverse high-fidelity images with vq-vae-2*. arXiv:1906.00446. DOI: 10.48550/arXiv.1906.00446. URL: http://arxiv.org/abs/1906.00446 (visited on 04/25/2025).

David Wang (n.d.). "Stacked Vector-Quantized Variational Autoencoders for Unsupervised Pretraining and Classification of White Blood Cells." unpublished. URL: https://davidw0311.github.io/assets/files/SVQVAE.pdf.

A Supplementary material

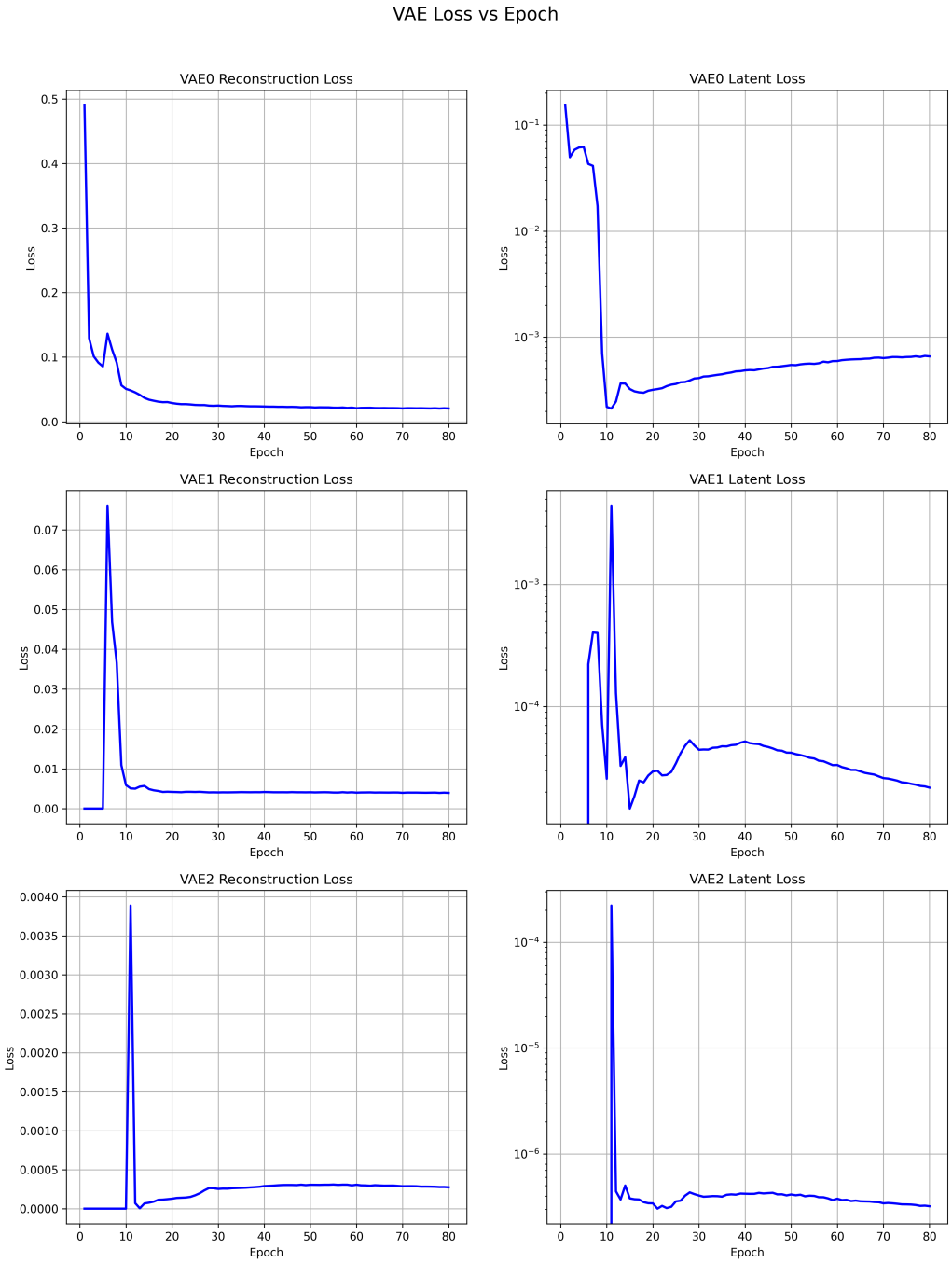

Figure 3: VAE losses for the original model (reproduction)

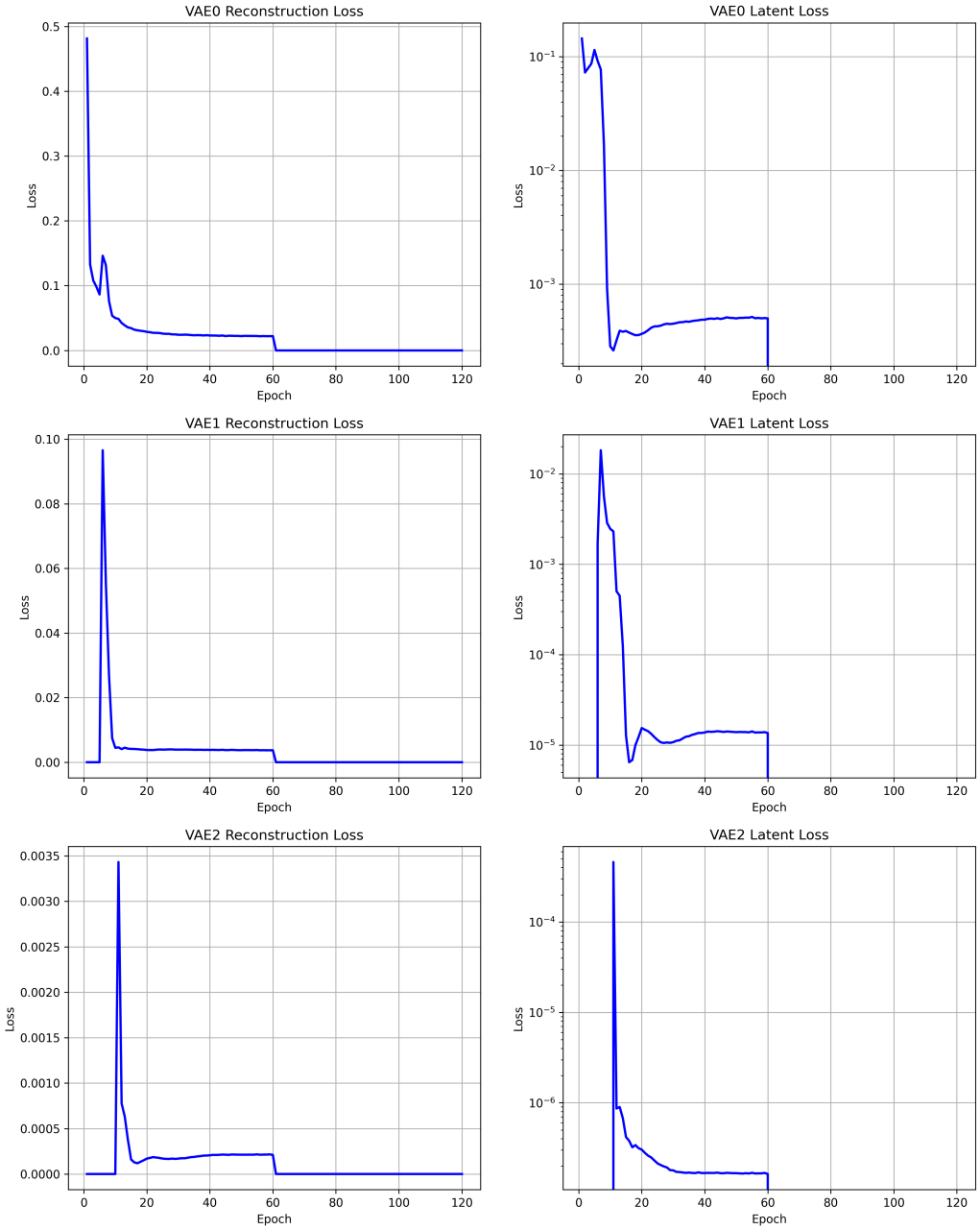

Figure 4: VAE losses for our updated model

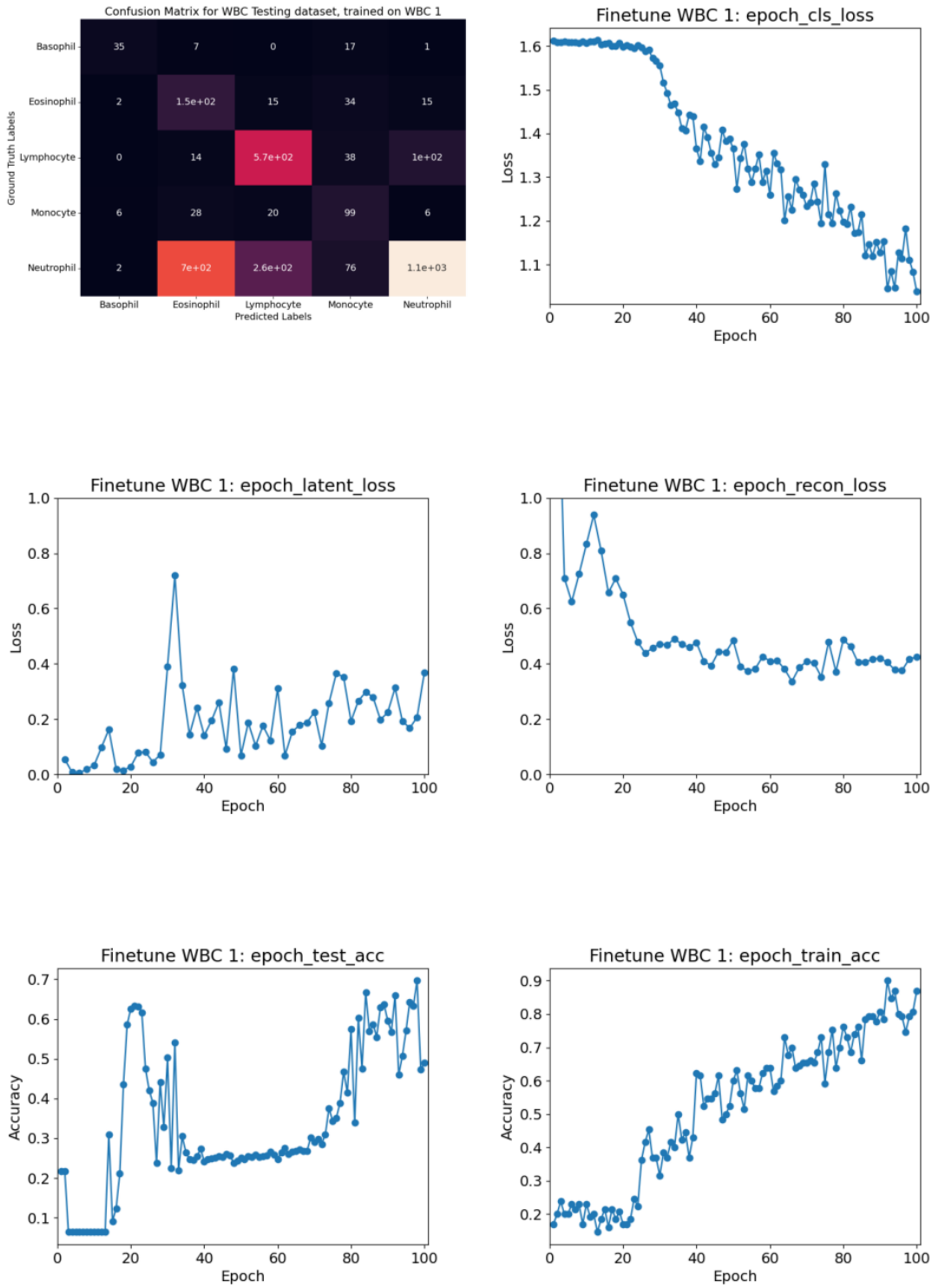

Figure 5: Statistics for finetuning of WBC1 (reproduction)

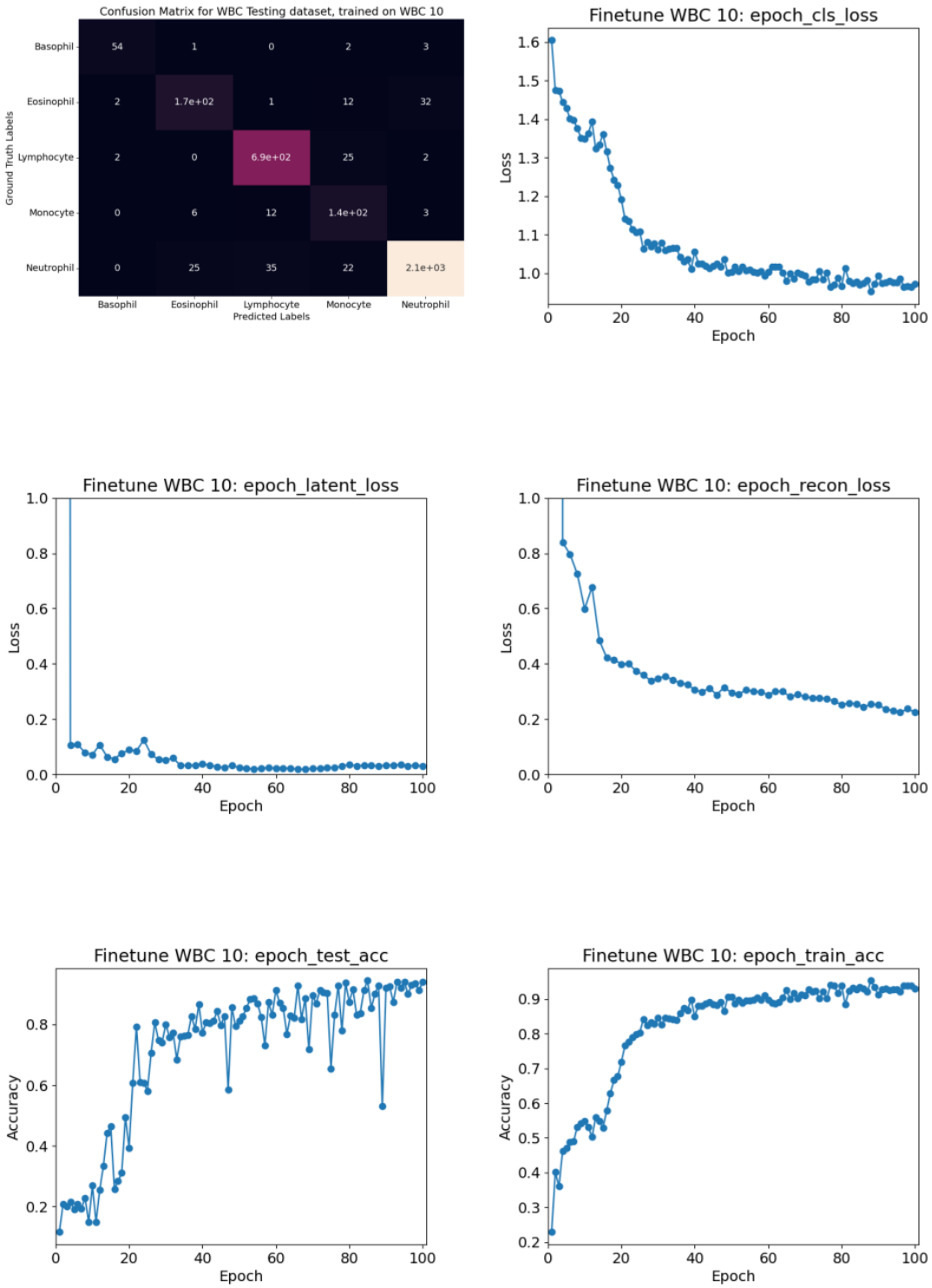

Figure 6: Statistics for finetuning of WBC10 (reproduction)

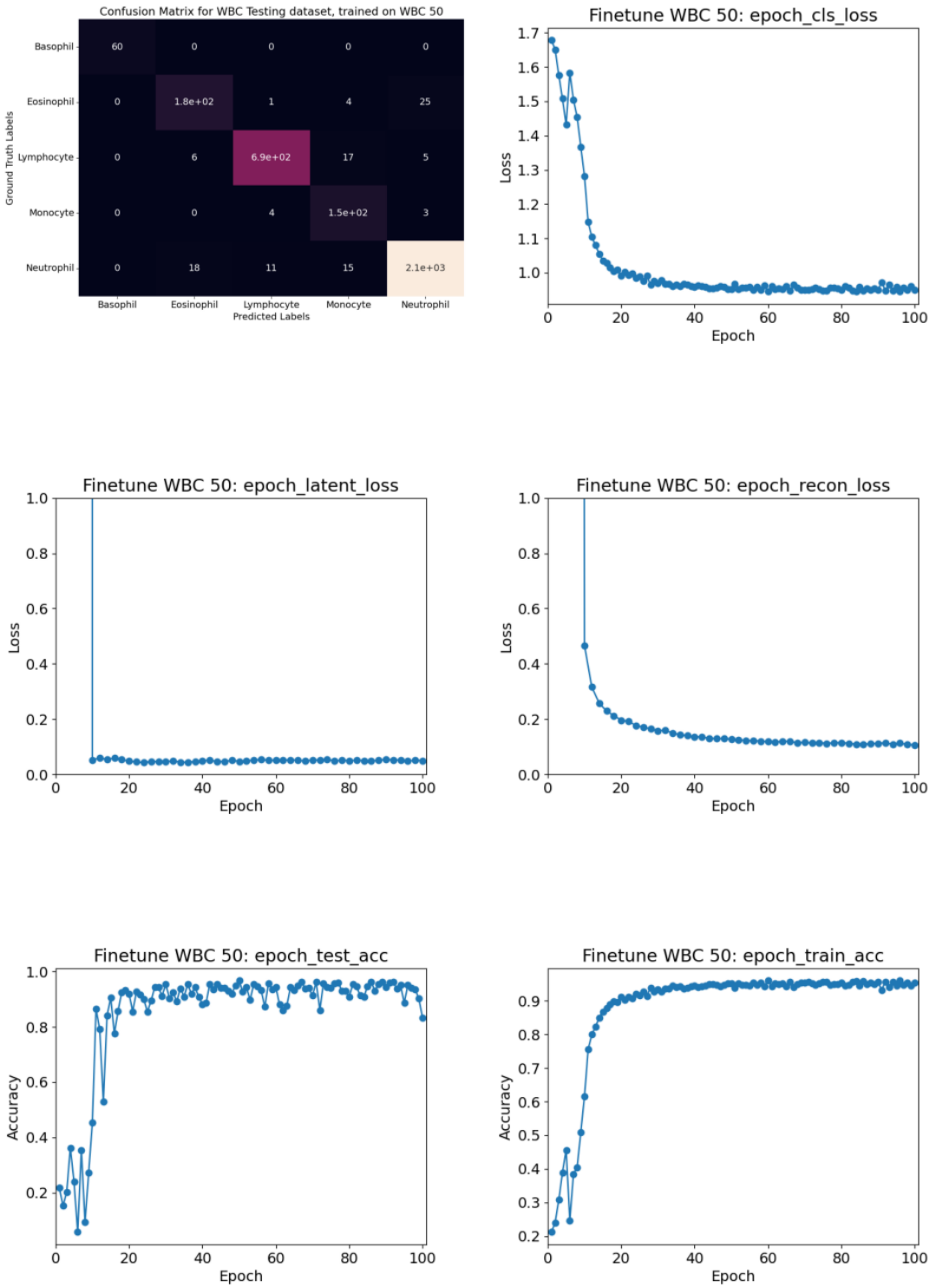

Figure 7: Statistics for finetuning of WBC50 (reproduction)