Final Report

Abstract

This report presents a comprehensive exploration of Convolutional Neural Networks (CNNs) in image classification, addressing real-world application challenges in the field.

In the project's first phase, we trained a ResNet-18 CNN on the Colorectal Cancer dataset, emphasizing computational efficiency. Through thorough hyperparameter tuning and analysis, we gained critical insights into model optimization. Training the Pretrained ResNet-18 model achieved the highest accuracy at 98.39%, and t-SNE reduced CNN encoder output dimensions, visually enhancing class feature distinctiveness and model performance understanding.

In the second phase, custom-trained (on Dataset 1) and ImageNet pre-trained CNN encoders were used for feature extraction on Prostate Cancer and Animal Faces datasets. t-SNE reduced dimensionality for visual clarity. Using the pretrained CNN encoder, we applied Random Forest (RF) (95.33% accuracy) and K-Nearest Neighbors (KNN) (94.00% accuracy) for Prostate Cancer, and on Animal Faces, RF achieved 99.61% accuracy, and KNN achieved 95.56% accuracy.

1. Introduction

The evolution of machine learning, particularly through Convolutional Neural Networks (CNNs), has revolutionized the field of image classification, a critical component in domains such as medical diagnostics and wildlife recognition. This technology's potential for high accuracy and efficiency in image analysis is pivotal. However, real-world data complexities, characterized by intricate patterns and diverse feature sets, present notable challenges. In medical imaging, for example, the task of differentiating between closely resembling pathological tissues demands a model capable of discerning subtle differences. Similarly, animal classification is complicated by variability in species appearances and environmental factors within images.

Compounding these challenges is the need for efficient adaptability of these models across different datasets, a key area where transfer learning becomes essential. Transfer learning involves leveraging a model developed for one task as a starting point for a model on another task. This approach is particularly beneficial in contexts where datasets, though distinct, may share underlying patterns or features. It allows for the utilization of pre-trained models on one dataset (like medical images) to be adapted to another (such as wildlife images and similar medical images), thereby harnessing previously learned features and knowledge.

In the context of this project, particularly with Convolutional Neural Networks (CNNs), the major challenges arise from the complexity and variability of real-world data in applications such as medical diagnostics and wildlife recognition. For instance, medical imaging demands highly discerning models capable of distinguishing between subtly different pathological tissues, while animal classification must contend with variations in species appearances and diverse environmental backgrounds. Additionally, the need for efficient adaptability of these models across different datasets introduces the complexity of transfer learning. This process, involving leveraging a model trained on one type of data (like human-annotated medical images) for use on another (such as wildlife images), is crucial but challenging, as it requires the model to generalize its learned features and knowledge to new, often vastly different, contexts.

In our methodology, we adopted a dual-faceted approach involving Convolutional Neural Networks (CNNs), particularly focusing on the ResNet-18 architecture. Initially, we trained a CNN model on the Colorectal Cancer dataset, emphasizing hyperparameter tuning to optimize accuracy and efficiency. This process involved adjusting learning rates, batch sizes, and exploring various activation functions and regularization methods to enhance model performance. Subsequently, we extended our approach to the Prostate Cancer and Animal Faces datasets, applying both pre-trained and custom-trained CNN encoders for feature extraction. This step was crucial in assessing the models' adaptability and versatility across different types of data, especially considering the distinct nature of the Animal Faces dataset compared to medical images. To complement our analysis, we employed t-SNE for dimensionality reduction, enabling a visual interpretation of how the models process and categorize different classes within these datasets. Our comprehensive methodology aimed not just at training and fine-tuning the CNN models, but also at evaluating their effectiveness and adaptability across varied datasets using advanced machine learning techniques like transfer learning.

In conclusion, the application of Convolutional Neural Networks (CNNs), specifically the ResNet-18 architecture, to our selected datasets yielded notable results. For the Colorectal Cancer Classification, the CNN model achieved a remarkable accuracy of 98.39%, highlighting its proficiency in medical image analysis. In extending our approach to the Prostate Cancer and Animal Faces datasets, we observed that both pre-trained and custom-trained CNN encoders performed effectively, with accuracies of 95.33% and 94.00% for Prostate Cancer, and 99.61% and 95.56% for Animal Faces, using Random Forest and K-Nearest

Neighbors methods, respectively. These results demonstrate the successful application of transfer learning, showcasing the model's adaptability across diverse datasets. Additionally, the t-SNE visualizations provided valuable insights into the models' feature differentiation capabilities. Overall, these findings underscore the versatility and robustness of CNNs in handling varied and complex image classification tasks, establishing a solid foundation for future explorations in this domain.

The integration of CNNs in image classification tasks has been a subject of extensive research, especially in fields requiring high precision like medical imaging and natural image classification. Pioneering studies by [4] and [2] have demonstrated the effectiveness of CNNs in complex feature recognition, forming the basis for advanced architectures like ResNet. The ResNet architecture, with its innovative approach to addressing the vanishing gradient problem, has been particularly influential in deep learning networks.

Transfer learning, a key focus of our study, has gained significant traction in the machine learning community. It allows models trained on large, diverse datasets to adapt and perform effectively on new, related tasks. This approach reduces the need for extensive dataset-specific training, optimizing resource utilization. Studies exploring transfer learning in CNNs have shown its potential in enhancing model performance across various applications, from medical diagnostics to wildlife conservation.

Our project builds upon these foundational insights, applying CNN methodologies to our chosen datasets while delving into the practical applications of transfer learning. By comparing the performance of models trained from scratch against those adapted through transfer learning, we contribute to the broader conversation on the adaptability and efficiency of pre-trained models in diverse classification tasks. The use of t-SNE for feature visualization in our study adds a unique analytical dimension, providing insights into the feature extraction capabilities of these models in a visually interpretable manner.

2. Proposed Methodologies

Dataset 1, the Colorectal Cancer Classification dataset, includes image patches categorizing three types of tissues: smooth muscle (MUS), normal colon mucosa (NORM), and cancer-associated stroma (STR), derived from Germany's NCT Biobank and the UMM pathology archive. Dataset 2, aimed at Prostate Cancer Classification, contains image patches identifying three prostate tissue types: Prostate Cancer Tumor Tissue, Benign Glandular Prostate Tissue, and Benign Non-Glandular Prostate Tissue, and is linked to a research paper [3]. Both datasets were meticulously compiled and annotated by pathologists and medical researchers who selected and categorized representative tissue samples, reflecting the complexity inherent in distinguishing between classes with closely resembling patterns, thus presenting significant challenges for accurate classification.

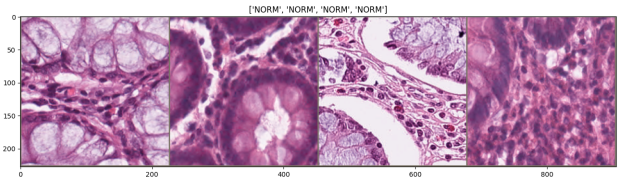

Figure 1. Colorectal Cancer (Dataset 1)

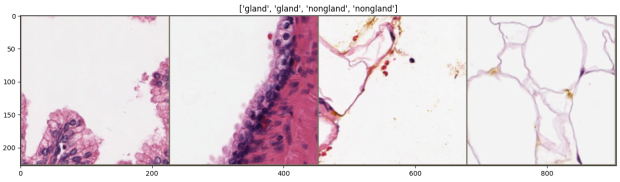

Figure 2. Prostate Cancer Cancer (Dataset 2)

Contrastingly, Dataset 3, the Animal Faces Classification dataset, consists of images of Cats, Dogs, and Wildlife, obtained from AFHQ (Animal FacesHQ). This dataset was collected from natural settings and involved extensive manual annotation, with the primary challenges being the background noise and the diversity of facial patterns within the same animal classes and among different breeds, further complicating the classification process.

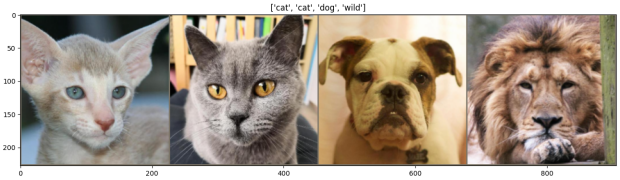

Figure 3. Wild Animal Faces (Dataset 3)

	D1	D2	D3
# Images	6000	6000	6000
Dim	224x224 pixels (RGB)	300x300 pixels (RGB)	512x512 pixels (RGB)
Format	TIFF	JPG	JPG
# classes	3	3	3
# images original D.	100000	6 dataset each containing 120000	16130
# classes original D.	9	3	3
Source	[3]	[5]	[1]

Table 1. Datasets description

For data preprocessing, we utilized the "transforms" module from PyTorch to prepare images for deep learning.

This module resizes the input images to 224x224 pixels, performs a center crop, converts them into PyTorch tensors, and normalizes the pixel values using ImageNet's statistics ([0.485, 0.456, 0.406], [0.229, 0.224, 0.225]). These steps ensure that the input sizes are standardized, the focus is on the relevant regions of the images, the data is compatible with PyTorch tensors, and numerical stability is maintained during model training.

For hyperparameter tuning, the datasets were divided into 56% for training, 14% for validation, and 30% for testing. During the model training phase, the data was further split into 70% for training and 30% for testing. This allocation allows for a focus on model learning with a substantial portion of the dataset, while still reserving a significant portion for evaluating the model's performance on unseen data.

For the image classification task, we opted for the ResNet18 CNN architecture. Our selection was motivated by its demonstrated capability in capturing intricate features through residual learning, as evidenced by its performance in benchmarks such as the ImageNet Large Scale Visual Recognition Challenge. ResNet18's moderate depth offers a compromise between the complexity of feature representation and computational efficiency, rendering it ideal for scenarios with constrained resources.

ResNet18, a derivative of the ResNet architecture, comprises 18 layers organized into several stages. Each stage includes residual blocks with skip connections, which enable the network to learn residual mappings and address the vanishing gradient problem, thereby aiding in training deeper networks. Each block contains two convolutional layers that facilitate the hierarchical capture of features. Downsampling within the network is accomplished using stride convolutions or pooling layers. At the convolutional layers' end, global average pooling is employed to reduce spatial dimensions before transitioning to the fully connected layers. This design is conducive to effective feature extraction and supports transfer learning, making ResNet18 an efficacious model for image classification tasks.

During hyperparameter optimization, we considered variables such as learning rate (lr), batch size (bs), and optimizers. We tested 27 hyperparameter combinations (lr = [0.1, 0.01, 0.001], bs = [16, 32, 64], optimizer = ["SGD", "Adam", "RMSprop"]), training ResNet-18 models from scratch over five epochs. The top five configurations were chosen based on their performance. These were then trained for an additional 15 epochs, with the most effective configuration (lr = 0.01, bs = 32, SGD optimizer) identified via validation accuracy. Additionally, a configuration showing consistent improvement with each epoch (lr = 0.001, bs = 64, SGD optimizer) was selected for further training.

We trained ResNet18 from scratch on Dataset 1, enabling the model to directly learn task-specific features. Additionally, we employed transfer learning using two approaches: first, by fine-tuning the entire network pre-trained with ImageNet weights, and second, by freezing all layers except the final fully connected layer, which was replaced and trained with random weights. These methods allowed us to tailor the model to our specific task while leveraging the knowledge embedded in the pre-trained weights. The optimized hyperparameters identified through the tuning process were applied in all training scenarios.

The pre-trained ResNet-18 model demonstrates a balance between computational complexity and training efficiency, as evidenced in its training and validation phases. During the initial training epoch with a learning rate of 0.01 and a batch size of 32, it achieved a training loss of 0.5856 and an accuracy of 74.86%, with a CPU time of 1 minute and 34 seconds and a relatively swift wall clock time of 21.4 seconds. This efficiency in processing time is significant, considering the model's computational intensity, as indicated by its requirement of 1.82 billion FLOPS (Floating Point Operations Per Second) to process a single image. Along with its substantial computational demand, ResNet-18 is composed of 11.69 million parameters and has demonstrated commendable accuracy on the ImageNet dataset, with a Top-1 accuracy of 70.07% and a Top-5 accuracy of 89.44%. These aspects highlight ResNet-18's capability to manage high-volume computations while maintaining practical execution speeds, making it an effective choice for image classification tasks that necessitate a blend of performance and time efficiency.

For dimensionality reduction, we primarily utilized t-SNE, a non-linear technique, to project the high-dimensional feature spaces of our CNN models onto a two-dimensional plane. This transformation was crucial for visualizing and understanding the feature distributions and class separations within our datasets. A key strength of t-SNE in our application was its ability to preserve local structures of high-dimensional data, allowing us to observe distinct clusters corresponding to different classes. For example, in the colorectal cancer dataset, t-SNE effectively visualized the separation among the three classes, illustrating the model's capability in learning discriminative features.

In Task 1, we adopted the ResNet18 CNN architecture for colorectal cancer classification, involving training the model from scratch and leveraging transfer learning with ImageNet pre-trained weights for adaptation. Our approach also included hyperparameter optimization for performance fine-tuning. The use of t-SNE for dimensionality reduction provided insights into the model's ability to distinguish between classes, offering a clear view of its feature extraction capabilities in the colorectal cancer dataset.

For Task 2, our focus was on feature extraction using pre-trained encoders for datasets 2 and 3. We employed two approaches: first, utilizing the custom-trained CNN encoder from Task 1, initially used on the colorectal cancer dataset,

to extract features from new datasets, which allowed us to assess how a model trained on specific data performs with different, yet related datasets, and second, applying a standard ImageNet pre-trained encoder to gauge the effectiveness of generic models in specialized contexts. This process involved passing images through the CNN up to the last convolutional layer and extracting the resultant feature maps for classification tasks.

For classifying these extracted features, we implemented two machine learning techniques: K-Nearest Neighbors (KNN) and Random Forest Classifier. KNN helped in analyzing the clusters formed, enabling inferences about the distinctness and separability of features from different classes. The Random Forest Classifier quantitatively measured the effectiveness of the feature extraction process, using metrics like accuracy, precision, and recall. This dual-technique approach provided a comprehensive evaluation of our feature extraction strategy's performance.

3. Results

For Task 1, we employed the ResNet-18 architecture, chosen for its optimal balance between depth and computational efficiency. In this task, we identified three distinct model training approaches. The first involved training a ResNet-18 model from scratch using Dataset 1, enabling the model to learn and adapt to features specific to the task. The second approach utilized transfer learning, where the network, initially pre-trained on the ImageNet dataset, was fine-tuned to suit our specific Dataset 1 better. The third approach also employed transfer learning but differed in its execution; all layers of a pre-trained ResNet-18 model were kept frozen except for the final fully connected layer, which was replaced and trained from scratch to accommodate the specific number of classes in our datasets (three classes).

In hyperparameter optimization for the ResNet-18 model, we explored 27 configurations, combining learning rates (0.1, 0.01, 0.001), batch sizes (16, 32, 64), and optimizers (SGD, Adam, RMSprop), to identify the most effective setup for colorectal cancer classification. This exploration allowed us to find a balance between rapid convergence and the nuanced adjustments necessary for Dataset 1. We tested higher learning rates for expedited learning and lower rates for more precise model tuning. The selected batch sizes helped evaluate the trade-off between computational load and gradient estimation accuracy, crucial for the efficient processing of complex medical images. Each optimizer presented a distinct method, from the simplicity of SGD to the adaptive learning rates of Adam and RMSprop, addressing our dataset's unique challenges. This strategic exploration was aimed at balancing learning speed with model accuracy, ensuring practical and stable training.

During the initial phase of our study, we embarked on training ResNet-18 models from scratch on Dataset 1, rigorously testing each of the 27 different configurations over five epochs. This essential preliminary training phase was instrumental in discerning the most promising configurations, as determined by their performance, especially in terms of validation accuracy. Having identified these metrics, we honed in on the top five configurations that demonstrated superior performance. These were then extended to a more robust training period of 15 epochs, providing us with a more profound comprehension of their learning capabilities and the sustainability of their performance over time.

The best-performing one (lr = 0.01, bs = 32, SGD optimizer), determined by validation accuracy, was identified. Additionally, a configuration (lr = 0.001, bs = 64, SGD optimizer) showing consistent improvement with each epoch was chosen.

We trained models from scratch and pre-trained models using two transfer learning approaches for the selected configurations. Below are the results for the six models.

Table 2. Model Performance Comparison

Training Data Results			
Metric	**Scratch Model**	**Pretrained Approach1**	**Approach2**
lr = 0.01, optimizer = 'SGD' batch size = 32			
Loss	0.002202	0.00032	0.117913
Accuracy (%)	99.976190	100.00	95.54761
Precision	1.00	1.00	0.973333
Recall	1.00	1.00	0.973333
lr = 0.001, optimizer = 'SGD' batch size = 64			
Loss	0.023551	0.007248	0.17127
Accuracy (%)	99.428571	99.952380	94.8095
Precision	0.94	1.00	0.95
Recall	0.94	1.00	0.95
Testing Data Results:			
lr = 0.01, optimizer = 'SGD' batch size = 32			
Loss	0.154925	0.049607	0.1279
Accuracy (%)	95.50	98.38709	95.0555
Precision	0.973333	0.993333	0.95
Recall	0.973333	0.993333	0.95
lr = 0.001, optimizer = 'SGD' batch size = 64			
Loss	0.141247	0.127920	0.1821
Accuracy (%)	95.50	95.055555	94.2222
Precision	0.886667	0.983333	0.94
Recall	0.883333	0.983333	0.95

The results indicate that the pre-trained model using transfer learning approach 1, with a learning rate of 0.01,

batch size of 32, and optimizer SGD, achieves the best performance. The choice of learning rate and batch size had a significant impact on the models performance. Lower learning rates and larger batch sizes tend to have slightly higher losses but maintain a good accuracy.

According to the results the use of the pretrained model has a high accuracy because the pretrained model are trained on large and diverse datasets like ImageNet, have already learned a broad spectrum of features. This rich feature extraction capability makes them highly adaptable to new tasks, providing a strong foundation for feature recognition right from the start, in contrast to models trained from scratch. Additionally, due to a significant portion of the model being pre-trained, it requires fewer epochs to achieve high accuracy, leading to faster convergence and a more efficient training process.

Feature visualizations (t-SNE) for the output features of two CNN encoders from task 1, the best scratch model, and the best pretrained model are provided below

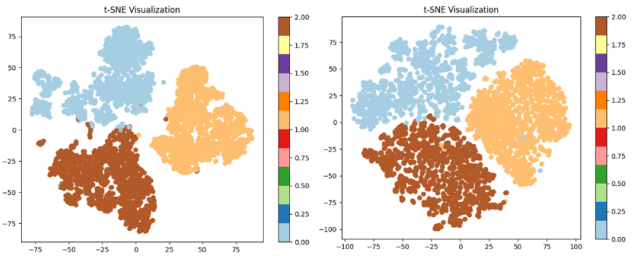

Figure 4. t-SNE for model from scratch vs t-SNE for pre-trained model

In the model trained from scratch (left plot), the clusters, while distinct, show overlap, notably where the orange and brown clusters meet, and the blue cluster is broadly spread, suggesting a greater variance within that class. This indicates a less defined class distinction, which may point to potential challenges in classification accuracy. The gradient scale across these clusters varies, implying differences in class confidence or feature density without a single class appearing dominant.

On the other hand, the pre-trained model (right plot) exhibits tightly packed clusters with well-demarcated boundaries, particularly between the orange and brown clusters, indicative of a strong separability of features. The blue cluster in this plot is noticeably more compact, reflecting a consistent feature representation within that class. Furthermore, the gradient scale across the clusters is more uniform, suggesting that the pre-trained model has potentially achieved a more balanced feature recognition across the classes. This uniformity and the clear separation of clusters in the pre-trained model's visualization underscore its superior capability in distinguishing between classes, likely benefiting from the diverse feature set it was initially exposed to during pre-training.

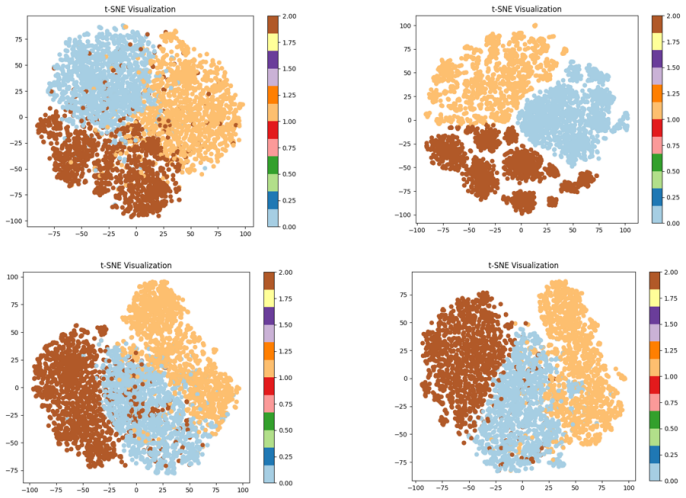

Figure 5. (1,1) t-SNE pre-trained model dataset3 vs (1,2) t-SNE task-1 model dataset3 vs (2,1) t-SNE pre-trained model dataset2 vs (2,2) t-SNE task-1 model dataset2

The t-SNE visualization(1,1) for animal faces dataset using a Task 1 trained CNN encoder , provides insights into the feature clustering effectiveness. The visualization shows moderately defined clusters with considerable overlap, particularly between the brown and orange, brown and blue clusters, indicating some ambiguity in the feature space learned by the CNN encoder. The distribution of the clusters, while somewhat even, features transitional areas where clusters merge, suggesting that certain features may be shared across different animal classes. This lack of distinct inter-cluster boundaries suggests the encoder have not fully captured the complexity needed for discrete classification of animal faces, which could impact real-world classification accuracy. The gradient scale on the visualization hints at a range of feature concentrations or classification confidence, with mixed values in overlapping regions pointing to lower model confidence. Overall, the CNN encoder from Task 1 has identified relevant features for animal faces; however, the clarity of feature separation is suboptimal, implying that further model refinement or additional animal-specific training could enhance performance.

The t-SNE visualization(1,2) of the Animal Faces dataset, utilizing features extracted by the ImageNet pre-trained CNN encoder, reveals a promising capacity for animal faces classification. The distinct and separate clusters corresponding to different classes denote the encoder's ability to discern discriminative features effectively. Variance in cluster density reflect the model's confidence in feature representation, with denser regions signifying common features within a class and sparser areas indicating unique variations. The minimal overlap between clusters underscores

the encoder's proficiency in extracting high-level, transferable features. This analysis suggests that the pre-trained CNN encoder has robustly captured the essence of different animal classes, indicating the potential for accurate classification upon reintroducing a classification head.

The t-SNE visualization(2,1) for prostate cancer classification, using features extracted by an ImageNet pre-trained CNN encoder without classification heads, reveals several insights about the model's feature processing capabilities. Notable cluster overlap indicates that the features for prostate cancer classification are not as distinctly separated as those in the animal faces dataset, suggesting that certain features or cases are challenging for the encoder to differentiate clearly. Varying cluster densities, with some densely populated areas, could represent common features across classes, while sparser regions might indicate outliers or unique characteristics. The blurred cluster boundaries suggest that the encoder struggles with effective discrimination between different classes of the prostate cancer dataset, possibly due to the dataset's complexity or a divergence from the features it was trained to recognize in ImageNet. In essence, the t-SNE visualization indicates that the pre-trained CNN encoder has captured some class-relevant features for prostate cancer classification, but the class distinctions are subdued, which affects classification accuracy and could benefit from additional model tuning or feature enhancement efforts.

The t-SNE visualization for Dataset 2, concerning prostate cancer classification, illustrates the clustering of features using a CNN encoder from Task 1. Significant overlap between clusters, especially the blue and brown, indicates the encoder's challenge in distinctly separating the prostate cancer classes. The clusters are unevenly distributed, with dense regions possibly reflecting common features across many samples, and sparse areas suggesting unique or rare attributes. This blurring of cluster boundaries reflects the inherent complexity of the prostate cancer classification or a divergence from the encoder's original training in Task 1, pointing to the difficulty of distinguishing these features. The adjacent gradient bar likely denotes feature density or probability, shedding light on the prevalence of certain features or the encoder's confidence in class assignments. In sum, the encoder trained in Task 1 demonstrates some capability in feature representation for prostate cancer classification, yet the substantial cluster overlap calls for additional model refinement or prostate cancer-specific training to enhance discriminative performance, a crucial step for medical imaging tasks.

The Table 3. presents results for RF and KNN trained on features extracted using a Pre-Trained CNN Encoder. Both KNN and RF models achieved flawless 100% accuracy during training for Prostate Cancer classification, indicating a perfect fit to the training data. However, when tested on new data, KNN and RF exhibited slightly diminished accuracy, with KNN achieving 94% and RF achieving 95.333%. This suggests potential overfitting during training, where the models may not generalize as effectively to unseen data. Improvement is needed for better generalization in the Prostate Cancer classification task. Conversely, on the Animal Faces dataset, both models displayed consistently high accuracy in both training and testing, with RF achieving a flawless 100% accuracy on both sets. These results underscore the models' exceptional performance on the Animal Faces dataset, highlighting their effective generalization capabilities.

Table 3. Task 2 Performance

Prostate Cancer Dataset		
Metric	**KNN**	**RF**
Training Data Accuracy (%)	100.00	100.00
Testing Data Accuracy (%)	94.00	95.333
Animal Faces Dataset		
Training Data Accuracy (%)	99.833	100.00
Testing Data Accuracy (%)	99.555	99.611

References

[1] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[3] Jakob Nikolas Kather, Niels Halama, and Alexander Marx. 100,000 histological images of human colorectal cancer and healthy tissue, May 2018. 2

[4] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 2

[5] Yuri Tolkach. Datasets digital pathology and artifacts, part 1, June 2021. 2