M.Sc. Business Administration and Data Science

Data Mining, Machine Learning, and Deep Learning

FINAL EXAMINATION PAPER

Classification of Mars Images

Analyzing the impact of imbalance handling methods on the performance of Convolutional Neural Networks using GoogLeNet and AlexNet

Students

Leonard Brenk (158287)

Finn Feddersen (158294)

Felix Wltschek (158368)

Supervisors

Somnath Mazumdar

Raghava Rao Mukkamala

Hand-In Date

19.05.2023

Pages: 13

Characters: 33.964

Abstract

Convolutional Neural Networks offer a wide range of applications and can understand complex tasks. One of the most common limitations is not the network itself but the amount and quality of data it needs to be trained. This paper analyzes the impact imbalance handling methods for training data – namely ADASYN and Random Oversampling – can have on CNNs, in particular GoogLeNet and AlexNet. The dataset consists of Mars landmark images that are augmented in six different ways to increase the size of the dataset and train more robust classifiers. To evaluate the impact of the methods used to address data imbalances, this paper employs a manual grid search approach to identify optimal hyperparameters for each combination of CNNs and datasets. Subsequently, six tuned models are executed and evaluated, resulting in the observation that both imbalance handling techniques lead to nearly identical results. This study fails to detect any significant differences when training neural networks with balanced data as opposed to imbalanced data. Nevertheless, it reveals that both CNNs outperform an SVM classifier in this image classification task, with the GoogLeNet architecture achieving better performance compared to AlexNet.

Keywords: *Convolutional Neural Network, Imbalance Handling, ADASYN, Random Oversampling, Support Vector Machine, Hyperparameters, GoogLeNet, AlexNet*

Contents

1 Introduction

Computer Vision is currently experiencing rapid development due to recent enhancements in the area of Artificial intelligence. Convolutional Neural Networks (CNNs) are very powerful and effective for image processing and recognition tasks. GoogLeNet (Szegedy et al., 2015) and AlexNet (Krizhevsky et al., 2012) are two of the leading CNNs available today and yield promising results for large-scale image recognition. Amongst other requirements for a well-functioning CNN, a sufficient amount and quality of training data is one of the most important. Especially in applications where images carry complex structures and depth, a high number of training images is necessary for the model to detect and recognize such structures. The limited availability of training data from various sources often constrains the application of CNNs. Furthermore, the distribution of classes in datasets is often very unbalanced resulting in a bias in the model. A possible solution for these two challenges is presented in this paper: Adaptive Synthetic Sampling (ADASYN) and Random Oversampling (RO) - two approaches to increase the number of high-quality data by creating (synthetic) instances based on the original data. This paper evaluates the impact of ADASYN and RO on the performances of GoogLeNet and AlexNet. A benchmark SVM model is developed as an alternative to CNNs for a more comprehensive comparison. In Section 2, we discuss relevant work, followed by the conceptual framework (Section 3) and the methodology in Section 4. The following section highlights and discuss the most important results (Section 5), before drawing a conclusion (Section 6) and elaborating on limitations and future work in Section 7.

2 Related Work

The exploration of Mars has become much more advanced in recent years and has produced new techniques and capabilities, paving the way for comprehensive analyses of the Martian surface. One of those is presented by Di et al. (2014) with developing a way to detect changes on Mars' surface using image comparison techniques. This paper builds on this analysis with a more specific classification of surface characteristics by identifying specific geological features within the images. CNNs are a widely applied technique for image processing stressing the importance of a balanced and sufficiently large dataset. As in reality, many data sources fall short in producing such data quality, RO and ADASYN can increase the number of instances in a set and, at the same time, improve the balance between classes as proposed by Rahman et al. (2023). In their paper, they analyze the two techniques by employing five different Machine Learning Models and evaluating the impact of the adapted datasets. RO is found to improve the overall accuracy from 88.88% to 96.29% and ADASYN even higher to 97.22%, proving that artificially created data based on original instances can have a positive impact on the overall performance of a model. In contrast to Rahman et al. (2023), this paper will use images as data instances with the ultimate goal of optimizing the input for a CNN. In particular, two of the most popular CNN architectures are applied, GoogLeNet (Szegedy et al., 2015) and AlexNet (Krizhevsky et al., 2012). As shown in (Swarup et al., 2023), both models perform exceptionally well, ranging from 98.95% to 99.45% – classifying images almost perfectly. In contrast to Swarup et al. (2023), the focus of this paper lies not only on the model but also on the input data quality. Building on their findings, this work expands the scope to encompass the entire pipeline – from data preparation to model deployment – offering a more holistic approach and finding the most promising combination of imbalance handling methods and CNNs.

3 Conceptual Framework

The main objective of this paper is to examine the classification of images in an imbalanced dataset. There are three hypotheses that are analyzed:

1. Neural Networks are a better approach to classifying images than SVM
2. A balanced dataset yields better results than an imbalanced set
3. ADASYN and RO are two suitable methods to be used for image processing with state-

of-the-art CNNs to create balanced training data

There are three ways, the imbalanced dataset can be incorporated into the training processes. The first approach is applying ADASYN, thereby creating synthetic images similar to the original ones but still unique. The second option is to implement RO and, as such, to create duplicates of existing images to balance the dataset. As a baseline, a third approach is used not employing an oversampling strategy and leaving the dataset imbalanced. This paper tests these three options and compares their results. In the attempt to find the most effective model, the paper applies a CNN and Support Vector Machine (SVM), quickly finding that the non-neural approach results in low accuracy. Thus, further analyses are performed using CNNs, in light of their unmatched proficiency demonstrated in several papers like (Swarup et al., 2023). We choose AlexNet and GoogLeNet as the two neural networks differ in depth and complexity, offering a more diverse analysis. The most effective CNN parameters must be found for comparing the three aforementioned approaches. Therefore, the performance of 4 different hyperparameter combinations for each approach and model is tested, resulting in 24 different models, each being trained for 10 epochs. For every model and every approach, the best model is picked and implemented using an extended training of 30 epochs in order to get a more reliable score. A comparison of the results is to indicate whether a balanced dataset yields better accuracy and whether ADASYN and RO should be considered for imbalanced image datasets. If so, it also suggests which model works best with which data processing approach.

Data Preparation

Training of the classifier models requires precise data preparation, which is conducted in five steps: data splitting, filtering, normalization, imbalance handling, and augmentation, as shown in Figure 3. The original data, retrieved from NASA's website, consists of seven different categories of landmarks on the surface of Mars and an eighth category summarizing unclassified landmarks as "other". As a first step, we split the dataset into training, test, and validation set. The immediate split ensures that the test and validation set is entirely unknown to our classifiers. Subsequently, all three datasets are normalized. The normalized training set functions as input to the imbalance handling step. Here, we apply ADASYN and RO, two different imbalance handling approaches. Additionally, the initial imbalanced dataset is used as a baseline, resulting in a total of three datasets for analysis. As a next step, these datasets are augmented, increasing their sizes by a factor of seven (using six augmentation techniques) and ensuring sufficient training instances. While these datasets serve as inputs for the CNN classifiers in their original shape, samples of 7000 instances (1000 originals with six augmentations) are used for the baseline model of SVM to speed up the fitting process.

Training Strategy

In this paper, we follow two approaches regarding imbalance handling for data preparation: ADASYN and RO. Additionally, we also create a baseline dataset that will not be balanced in order to compare the effect of the different imbalance handling techniques accurately. These datasets will then be used as input for our two CNN classifiers, AlexNet and GoogLeNet models. Samples of the datasets will also be used for our model baseline, a support vector machine (SVM). Therefore, each model will be trained three times, once for each dataset. To optimize the performance of our models, we performed a manual grid search for the GoogLeNet and AlexNet models for the parameters *batch size* and *learning rate*. Further, grid search cross-validation is performed for training the SVM and identifying optimal parameters. Finally, six combinations of datasets and CNN models are trained, and their results are compared.

4 Methodology

This chapter presents the research methodology used in this paper. Section 4.1 explains the data that was used for the study, while Section 4.2 shows the steps of data pre-processing. Subsequently, Section 4.3 outlines the employed modeling framework and the different techniques implemented to ensure reliable results which are compared using the evaluation metrics described in Section 4.4.

4.1 Data Description

The data used throughout this project was provided by NASA and generated by the MARS Reconnaissance Orbiter. The Reconnaissance Orbiter is a spacecraft used to investigate the atmosphere and terrain of the planet Mars. Throughout its mission, it has provided images of the surface of Mars, which will be the foundation of the input data for the trained models in this paper (Institute of Technology, 2005).

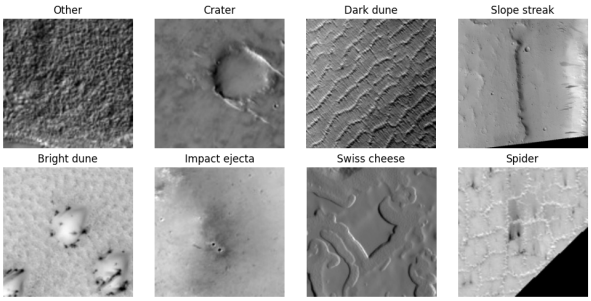

Figure 1: One image per class

The original dataset consists of 10,815 grayscale images of landmarks. The images are distributed over seven different classes of landmarks and one additional class. Examples for each class can be seen in figure 1.

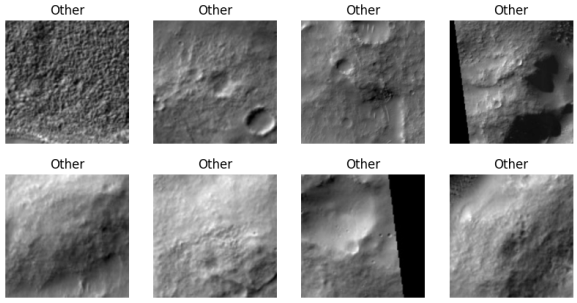

Figure 2: Images from the class "Other"

The class "Other" has to be highlighted. All other seven label categories are used to describe a certain type of landmark on Mars. Other is instead used to classify all included images whose landmarks do not fall into one of the seven categories. Consequently, and in contrast to the other categories, the images labeled as "Other" do not consist of specific characteristics shared by the images of the class (Cf. figure 2).

4.2 Data Preprocessing

Preprocessing is an essential step that must be performed to ensure that the input data provided to the classifiers is suitable for effective training and testing. In the following sections, we delve into a more detailed explanation of the specific steps involved.

Data Splitting

In the beginning, 10,815 original landmark images are split into training and test data. By doing so, we ensure that our final classifiers will be tested against completely unknown test instances which were under no influence of other preprocessing or training steps (e.g., balancing the datasets in section 4.2). For the training-test split, we use 40% as a splitting factor. In light of further CNN training, the test set is split into 30% validation and 70% test set. Through this process, we create test sets to evaluate the accuracy of the CNNs and the baseline method employed. Table 1 gives an overview of the three created sets – training, test, and validation – and their share of the overall dataset.

	Train.	Test	Val.	Total
#	6,489	3,028	1,298	10,815
Share	60%	28%	12%	100%

Table 1: Split of original dataset

Data Filtering

After splitting the dataset into training and test/-validation sets, the training dataset is further filtered. In the original dataset the class "Other" is significantly over-represented. We decreased the number of instances for this class, not in an attempt to apply artificial balancing to the data, but rather

to increase training speed and reduce the risk of overfitting. Considering that it is difficult even for highly specialized experts at NASA, a renowned institution known for its extensive expertise, to accurately distinguish these categories in the published dataset, it is clear that the distinction between these specific classes is inherently complex. Therefore, seeing that the distribution is heavily imbalanced, the reduction in the size of the largest class is intended to support models in identifying specific features. Regardless of this sampling strategy, an imbalance of classes remains, however, less severe than in the initial dataset. As such, this problem continues to affect the performance of our selected models and is accounted for using balancing strategies.

Classes (Label)	Instances
Other (0)	740
Crater (1)	479
Dark Dune (2)	102
Slope Streak (3)	181
Bright Dune (4)	152
Impact Ejecta (5)	47
Swiss Cheese (6)	181
Spider (7)	107

Table 2: Distribution of instances over classes

Table 2 displays the distribution of the images over the different classes after the filtering. Here a high imbalance in the dataset can be observed. The class "Other" is over-represented with 740 instances and is almost double the size of the second largest class. On the other hand, "Impact Ejecta" consists of only 47 instances and is by far the smallest class, being almost 17 times smaller than "Other". This imbalance will be handled in section 4.2. The final training set consists of 1,989 instances.

Data Normalization

The normalization step is important to ensure consistency within the used dataset. By this, it is avoided that diverging factors get integrated into the model input that could negatively influence the training of models. As the images are already in a 227 x 227 shape, no further size adjustments are required to use the images for model training sufficiently. In addition to the overall size of the images, pixel values are scaled to values between 0 and 1. This is done consistently for the training, test, and validation set.

Imbalance Handling

The training set is highly imbalanced. The dataset includes two highly over-represented classes ("Other" and "Crater") as well as one under-represented ("Impact Ejecta"). Overall, images categorized as "Other" represent 24% of the training set and, therefore, account for two times as much as in a balanced dataset. On the other hand, images of the class "Impact Ejecta" represent only a tenth of what they would in a perfectly balanced dataset (Cf. Table 2).

Using imbalanced training data

One could argue that this is not a problem or can rather be considered a limitation of the training set as it simply represents real-life conditions. As described in 4.1, the class "Other" does not describe a specific characteristic. Instead, it is used for classifying all instances that cannot be assigned to one other seven classes. Even though it is more likely that an image of the surface of Mars taken by the Mars Reconnaissance Orbiter does not contain a landmark out of the seven categories, this over-representation could lead to a heavy bias of the models towards the over-represented classes. This would cause the models to perform really well when trying to identify instances of the respective categories. A common problem could thus be that instances of, e.g., "Crater" that do not clearly demonstrate the characteristics of this landmark are easily mistaken for the class "Other". This is especially the case as this class does not describe a specific pattern but rather a very large "space of possibilities" containing everything that does not look like one of the remaining categories. Due to the high share of "Other" instances in the dataset, the model would still yield relatively high accuracy while performing poorly for the other classes (Kotsiantis et al., 2003). Song et al. (2021) identified that balancing techniques can support eliminating a discriminating

bias against under-represented classes in the models. Reza and Ma (2019) also support this by stating that imbalances in classes could negatively influence the training of CNN models. As one objective of this paper is to evaluate the impact of balancing, the initial imbalanced training set is used for comparison.

General approaches to handling imbalance

In the attempt to mitigate bias caused by imbalanced data, augmentation, and undersampling have shown promising results in the past (Song et al., 2021) (Reza and Ma, 2019). Undersampling describes reducing the number of instances for each class to the number of instances of the smallest class. As described earlier, the class "Impact Ejecta" is the smallest one in the training set. This would lead to a training set size of not even 400 instances. However, this size would be relatively small and likely insufficient for training models for the application discussed in this paper. With augmentation, modifications of the original images would be used to balance the dataset by extending it instead of making it smaller. In the given case, 15 augmentation techniques would be required to increase the category size of "Impact Ejecta" to the required balancing level. A challenge for the use of augmentation would be that augmentation could not be used again to scale the then balanced dataset. However, augmentation will be required to further increase the training data (Cf. Section 4.2). Therefore it should not be used to balance the data as well.

Random Oversampling & ADASYN

Both undersampling and augmentation fail to create a sufficiently large dataset while mitigating imbalance. However, ensuring a large enough training set is a fundamental aspect of data preparation. Training a model with too few instances could limit its ability to identify structures and patterns. This would reduce the overall performance or cause overfitting. Potential solutions for overcoming this challenge are to generate synthetic images or simply use duplicates. Reza and Ma (2019) suggest the use of oversampling for handling imbalances. When using oversampling, especially RO, duplicates of under-represented instances are added to the dataset to balance it. This approach can be advantageous as a sufficient training set size is maintained, and only original images are used. In the given case, 3,931 duplicates would be created. This means that a model trained on the balanced training set would see each image on average 490 times. The high amount of duplicates increases the chance for the model to overfit as it would assimilate the explicit instances more than develop a more generalized understanding of their characteristics. Nevertheless, we can mitigate this concern by implementing augmentation techniques. By expanding the training set, the duplicates would occupy a smaller proportion of the overall dataset, diminishing their frequency for the model. Consequently, the model would have a more significant opportunity to discern and grasp general features as it encounters individual images less frequently. Consequently, this paper will explore the potential of employing RO as one of the discussed approaches. Instead of using duplicates to balance the dataset across all its classes, we could also create synthetic instances (Ali-Gombe and Elyan, 2019). This can be done by applying ADASYN, which attempts to lessen the imbalance in the dataset by creating new images based on the original ones. For each image within a class, approximately as many images are synthesized as necessary to attain a similar count to that of the majority class. Therefore, ADASYN does not rely on the use of duplicates and could therefore limit the risk of overfitting the model. Creating synthetic images would reduce the possibility for the models to simply learn the exact characteristics of the input data instead of only identifying patterns. The success of this technique fundamentally relies on the quality of the generated images. If they are too similar to the original images, we will receive a similar result as using only duplicates. However, if the synthesized images were too different from the original images from their class, models may struggle to discern the common patterns, thereby undermining their performance. Because of this, it is essential to note that the quality of the synthetic images generated by ADASYN can

depend on the complexity of the dataset and the quality of the original minority class samples. We will use RO and ADASYN to balance our dataset. The key difference between the two, using duplicates or synthetically generated images, leads to one of the key research questions of this paper - identifying the impact of synthesized images and duplicates on the model training. Both methods lead to a balanced dataset, even though Random Oversampling creates a perfectly balanced one, and ADASYN still leaves a slight imbalance. Overall, the two approaches created dataset with 5,520 and 5,482 images.

Data Augmentation

In order to further scale the size of the training set and to make trained classifiers more robust, we decided to apply six augmentation techniques and thereby create six augmented images for every instance in the dataset. For the dataset extended using ADASYN, the additional images are based on original and synthetic images. For the dataset that was balanced through Random Oversampling, the added images by augmentation are completely based on existing images but result in more duplicates. The augmentation methods include rotations by 90, 180, or 270 degrees to capture different perspectives of the landmarks. Furthermore, images were flipped horizontally or vertically and zoomed. An exemplary image with its respective augmentations can be seen in figure 4. For the baseline SVM model, a smaller dataset of 1000 images per balancing method is used. Each of these images undergoes one of the aforementioned augmentations. This step completes the data preprocessing.

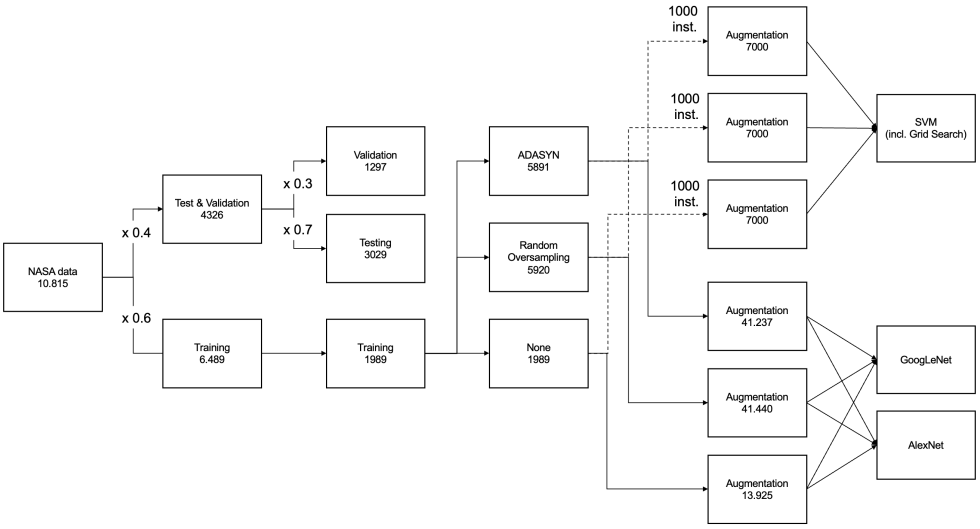

Figure 3: Data Preprocessing

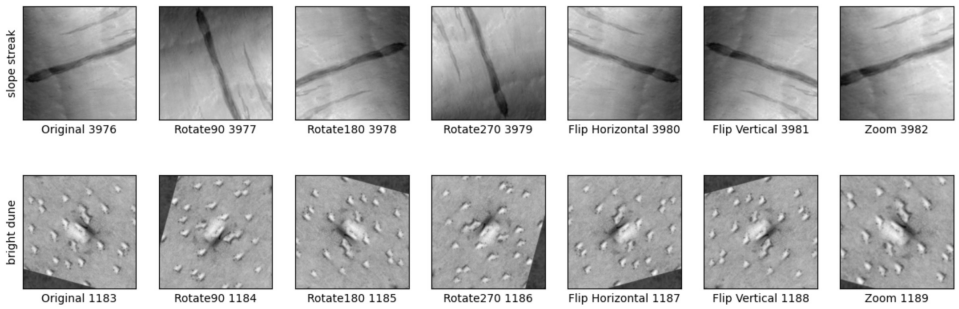

Figure 4: Types of augmentations

4.3 Modeling Framework

In order to make a comparison between the two oversampling methodologies of ADASYN and RO, this project makes use of benchmarks with an SVM and two Convolutional Neural Networks (CNNs): AlexNet and GoogLeNet. Both CNN implementations are conducted without pre-training and are able to show the impact of the oversampling strategy on the training process. While the main structures of the two models were copied, slight adjustments to hyperparameters were made.

4.3.1 Convolutional Neural Network

We choose Convolutional Neural Networks for Mars image processing because they are the leading method in modern image processing. The increasing popularity and continuous improvement of models such as AlexNet, GoogleNet, and ResNet indicate significant developments in this field. Their consistent further development and the constantly growing application potential demonstrate that CNNs can deliver ever better results and are thus the right choice for our research. Compared to other Neural Networks, CNNs have the advantage of being able to capture complex structures by using two-dimensional filters covering not only the current neuron but also its neighbors. Thereby CNNs are able to understand larger complex structures in images and are thus highly beneficial for our use-case (Géron, 2019).

4.3.2 AlexNet

AlexNet was first implemented in 2012 and won the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) (Krizhevsky et al., 2012). With a deep architecture, the neural network outperformed traditional models in the general image classification task. The network encompasses 5 convolutional layers and 3 fully connected layers, amounting to a total of approximately 24.7 million parameters. The input layer of the network is adjusted to take greyscale images of the size 227x227 with one channel (greyscale). The first layers employ large strides of 4 to decrease the shape of the input images. Additionally, max-pooling layers are used to "subsample the input image in order to reduce the computational load, the memory usage, and the number of parameters" and thereby reduce the risk of overfitting (Géron, 2019). The layers use a ReLU activation function to introduce non-linearity into the network. This specific activation function has been shown to converge faster due to its non-saturating nature (Krizhevsky et al., 2012). Generally, the number of filters the network uses increases throughout the CNN – ranging from 96 up to 384. Thereby, the model captures high-level before focusing on low-level features. In the latter part, the network uses 3 dense layers and additionally incorporates dropout layers to reduce overfitting. In consequence, 50% of the layer is temporarily disabled and ignored during the training step. This results in a more robust model that is less sensitive to minor variations because it prevents neurons to create complex co-adaptations with neighboring neurons (Krizhevsky et al., 2012). Finally, the output layer is adjusted to represent the 8 categories of Mars structures in the dataset and employs a softmax function to generate respective probabilities for the classes.

4.3.3 GoogLeNet

As a second model to test the oversampling strategies, GoogLeNet is introduced. Also known as Inception-v1, the network was introduced by Google in 2014 and is similar to AlexNet, a deep convolutional neural network. Also similar to the AlexNet, the input layer is adjusted for our needs to be able to capture greyscale images of the size 227x227. While there are certain parallels between the two networks, such as combinations of max pooling layers and convolutional layers with large strides in the first layers, distinct features of the GoogLeNet are its 9 inception modules. The modules employ parallel convolutions using different filter sizes, allowing the network to capture features at multiple scales. By stacking these modules on top of each other, the network can learn increasingly complex features and capture low-level and high-level features simultaneously (Szegedy et al., 2015). Just like AlexNet, the network uses dropout layers as regulation, preventing the model from overfitting and ReLU activation functions in all convolutional layers. Finally, the last layer of the network is ad-

justed to our needs and encompasses a softmax activation function with 8 output neurons.

4.3.4 Hyperparameter Adjustments

Both models are slightly simplified to fit our use case. As such, batch normalization layers in the AlexNet and auxiliary classifiers in the GoogLeNet were omitted. Hyperparameters are chosen according to the recommendations from the AlexNet and GoogLeNet papers. The following fixed hyperparameters are used: Stochastic Gradient Descent optimizer with a momentum of 0.9 and Nesterov, a categorical cross-entropy loss function, and weight initialization according to Glorot uniform. Specifically for the AlexNet, we follow the original structure and introduce weight decay of 0.0005, which is "not merely a regularizer" but "reduces the model's training error" (Krizhevsky et al., 2012). These hyper-parameters are not tested for other values due to our constraints of resources and the complexity of the models. However, to obtain some information about the tuning of the models, we deploy GridSearch for the hyperparameter optimization of learning rate and batch size. Both models were fitted on the dataset generated by the ADASYN and RO algorithms using 4 combinations with batch sizes of 32 and 128 and constant learning rates of 0.01 and 0.005[1]. The fitted results can serve as an indicator of the performance of the models using the respective oversampling algorithm. In total, 24 combinations derived from three datasets, two models, and 2 hyperparameters (with 2 options each) are run as depicted in Figure 5. This initial manual grid search for only 10 epochs served as the foundation for a more elaborate analysis of the best combination of learning rate and batch size. The four combinations of the two models and two datasets are then analyzed more elaborately by fitting for 30 epochs.

4.3.5 Benchmark SVM

To establish a benchmark for evaluating the performance of our modified and trained neural networks, we employ a Support Vector Machine classifier. In its most simple form, an SVM classifier only supports binary classification based on the principle of defining a hyperplane. This hyperplane separates the data into two classes by maximizing the margin or decision boundary between all data points. However, it can also be used for multi-classification by breaking the problem down into multiple binary ("one-vs-one") classifications. Hyperparameters for the approach were chosen according to grid search cross-validation over a subsample of 7000 observations (1000 original instances) from the training data. The resulting parameters are: C=10, gamma="scale", and "kernel"="rbf"

4.4 Evaluation Metrics

The performances of the employed models were evaluated using accuracy, precision, recall, and F1 scores. Especially, for the neural networks that are compared in a more detailed fashion, using precision and recall is of high importance to receive a more elaborate perspective on their performance. The importance is underlined by the imbalance of the test dataset. Because the first category "other" accounts for approximately 82% of the test set on which the models are evaluated, a model that solely predicts this category can reach high accuracy. However, by taking into account false positives and negatives, precision and recall can provide a more comprehensive analysis. For the neural networks, these measures are displayed using a classification report and illustrated using a confusion matrix[2]. More formally, precision is defined as the share of true positives over true and false positives and as such measures what fraction of classified positive instances are actually positive. On the flip side, recall is defined as the share of true positives over true positives and false negatives. It measures a model's ability to identify positive instances cor-

[1] This differs from the original neural networks that employed learning rate schedules to decrease it after certain epochs, or when validation accuracy stagnates.

[2] At this point, we have to point out that the classification reports displayed in the attached Jupyter Notebooks were created with variables in reverse order, resulting in swapped Precision and Recall measures.

rectly. The relationship of the two measures can be described as a trade-off. A model that solely maximizes one, decreases the score of the other. In other words, a model that solely predicts the dominant category can achieve a high recall for this category at the cost of precision. The trade-off between the two measures is captured in their harmonic mean, the F1 score, providing a balanced perspective of maximizing both measures. This per-class measure is then aggregated in a macro- and micro-average. As the test dataset is highly imbalanced, the main focus is laid on the macro-average that equally considers each class by taking the average of all F1 scores.

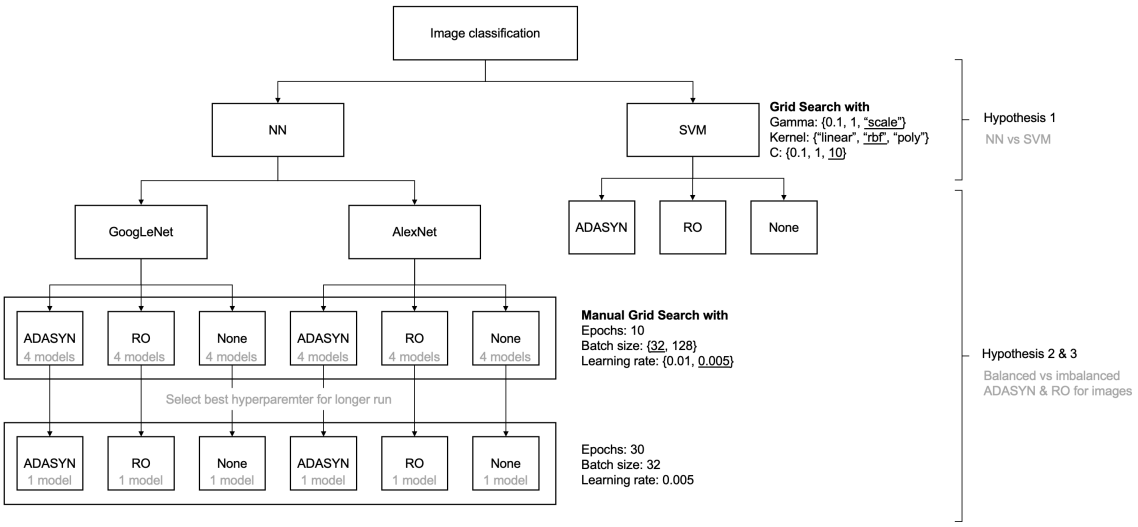

Figure 5: Conceptional framework and used networks

Model	Method	Batch	L. Rate	Accuracy	Precision	Recall	F1
AlexNet	RO	128	0.01	0.80	0.52	0.77	0.60
AlexNet	ADASYN	32	0.005	0.82	0.53	0.80	0.62
AlexNet	None	32	0.005	0.88	0.65	0.74	0.67
GoogLeNet	RO	32	0.005	0.86	0.65	0.81	0.70
GoogLeNet	ADASYN	32	0.005	0.85	0.62	0.84	0.69
GoogLeNet	None	32	0.005	0.89	0.66	0.79	0.70

Table 3: Macro Average of Precision, Recall, and F1 of selected models after 30 epochs of training

5 Results

The evaluation of the relative effectiveness of different machine learning methods shows that Convolutional Neural Networks (CNNs) outperform Support Vector Machines (SVMs) for image classification. In particular, the accuracy achieved by SVMs is less satisfactory, supporting the assumption that CNNs provide more desirable results for this particular dataset. The SVM that is applied in this paper yields an accuracy score of approximately 18% for the balanced datasets and 60% for the imbalanced dataset. Comparing this to the average accuracy achieved by both CNNs of approximately 80% shows that the neural network approach clearly outperforms the SVM, support-

ing hypothesis 1. For this highly complex dataset with difficult-to-detect characteristics in each class, we conclude that a more sophisticated structure is needed to classify images correctly. By employing a deep, multi-layered structure with two-dimensional filters, neural networks are able to distinguish and learn not only high-level features but also minor differences between the eight classes.

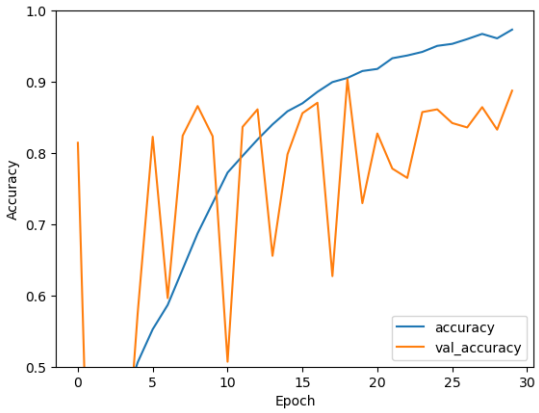

Figure 6: GoogLeNet learning process over 30 epochs for the imbalanced training set

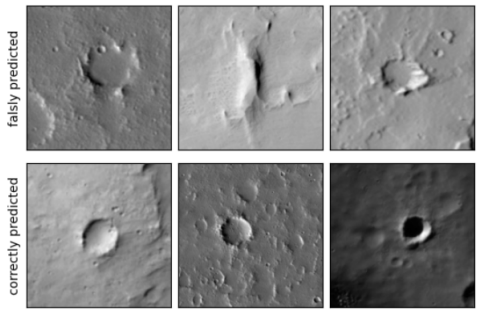

Figure 7: "Other" images classified as "crater"

In regard to the second hypothesis about balanced and imbalanced data, the former is found to deliver a more robust development throughout the first 30 epochs. As shown in Figure 6 the validation accuracy of the imbalanced dataset is very volatile. We assume the model fixates on varying classes when assigning labels. This implies that the model struggles to accurately predict or assign labels to the minority classes in the imbalanced dataset, possibly favoring the majority class instead. In contrast, the balanced set shows a more controlled and steady increase as shown in Figure 10 while achieving an almost identical validation accuracy. We further assume that for the imbalanced data, a longer training process would result in a more stable validation curve similar to the case of balanced data. With regard to the evaluation metrics, only slight differences can be identified. While the models trained on the imbalanced dataset show slightly higher precision at around 0.65 in comparison to the models trained on balanced data with 0.53-0.62, they perform on average 0.04 worse in terms of recall. However, the classification reports show a similar pattern for both training algorithms. While quite accurately classifying all seven landmark classes without many false negatives (high recall), the category "Other" shows a relatively low recall, nevertheless resulting in an overall high macro-average recall. As images that belong to the general category of "Other" can contain parts of landmarks that resemble other classes, they are misclassified as such. This is especially apparent in the classes "Crater" and "Slope Streak". An example of such misclassifications is shown in Figure 7 taken from the GoogLeNet model trained with the dataset created using RO. Clear features that match those of a crater can be identified in the "Other" images, illustrating only slight differences that would be challenging even for humans to distinguish. Due to these false classifications in the two categories, "Crater" and "Slope Streak" show below-average precision. Even though the general recall score is high throughout the classes, "Impact Ejecta" shows a lower recall value. As the category has only a few instances, false negatives have a higher impact on the recall value compared to other classes with more instances. In conclusion, all six models show very similar results and do not show clear differences in any of the measures. As such, we cannot conclude that training with either balanced or imbalanced is preferred for neural networks. The only clear differences are illustrated in the validation curve which could be caused by a decreased training set size and requiring a longer training time. Furthermore, no clear differences can be identified between the oversampling strategies of ADASYN and Random Oversampling. The aforementioned benefit of ADASYN

of generating new instances instead of duplicating existing ones and the hypothesized resulting benefit of being less likely to overfit cannot be observed in the results. Both models, AlexNet and GoogLeNet show almost identical results and demonstrate similar training structures using the respective datasets.

In light of the third hypothesis, it can be observed that both ADASYN and RO deliver satisfactory results when applied to GoogLeNet and AlexNet. As part of the manual grid search, each combination of imbalance handling methods and CNNs is run four times with different parameters for 10 epochs. The resulting best hyperparameters are used to run the models for a longer training time of 30 epochs which results are displayed in Table 3. It is important to emphasize that five out of six combinations are trained with a Batch size of 32 and a learning rate of 0.005. The results highlight that the GoogLeNet architecture performs better in all dataset variants with an average macro-average F1 score of 0.70 and accuracy of 0.87 in comparison to AlexNet's 0.63 and 0.83. This is likely due to GoogLeNet's more sophisticated structure employing inception layers, being better able to capture structures in classification tasks. These most likely also cause the differences in running time, as shown in Figure 9. It becomes apparent that GoogLeNet, compared to AlexNet, requires about 4 to 5 times more runtime. As such the training process over 30 epochs for GoogLeNet using the balanced datasets takes between 9 to 10 hours to run.

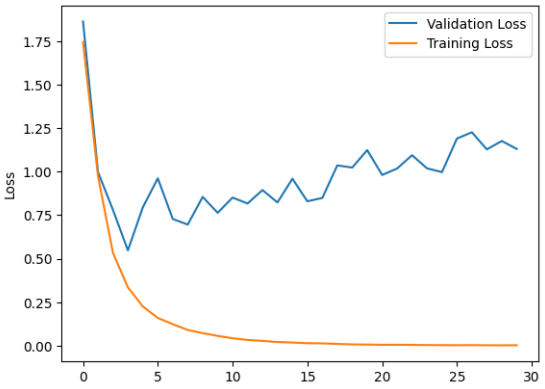

Figure 8: Validation Loss vs Training Loss

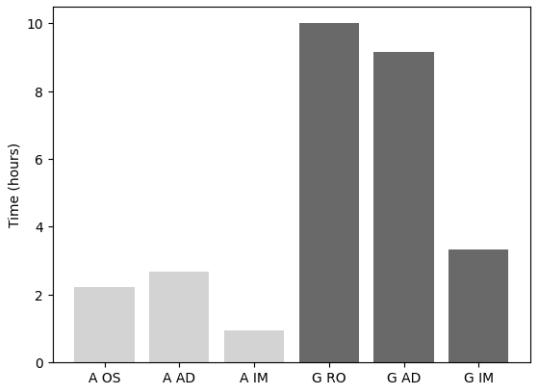

Figure 9: Training time for the six Neural Networks

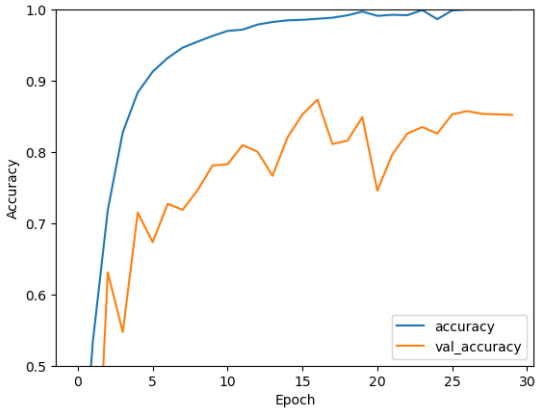

Figure 10: GoogLeNet learning process over 30 epochs for the Rnd OS dataset

Throughout the training and testing of the different models on multiple datasets, indications for overfitting are discovered. An example of this can be seen in Figure 8, where the training loss decreases over time. Therefore, the accuracy of the model improves with regard to the training data (Cf. Figure 10). While the training loss decreases steadily throughout the first five epochs, the validation loss stagnates and starts increasing again (Cf. Figure 8). The development of the validation accuracy can be described vice-versa, as it first increases in parallel to the accuracy, then stagnates and even decreases slightly (Cf. Figure 10). The described observations indicate overfitting. This can be due to the model initially learning characteristic features of the labels but focusing on solely improving its understanding of the training data after a certain number of epochs. This takes the exact shape

shown in the figure where the training accuracy continues to increase while the validation accuracy stagnates or even decreases. A possible explanation for overfitting the training data is the presence of duplicates or very similar images generated by the oversampling techniques.

6 Limitations & Future Work

From an ethical standpoint, further investigation and improved understanding of space and its objects, like Mars, are a general public interest. In the case of a broader application of machine-learning techniques in this field, such as the one described in our paper, it would be important to make the gathered information available. This would not only be true between institutions working in the respective field. Rather should the data be also made available to the public. Therefore, it would be crucial to further ensure the accuracy of used data and provide transparency about the data and machine-learning techniques used in the process.

Especially due to computational constraints, the performed analyses suffer from limitations that need to be addressed. The SVM, which is used for an initial benchmark comparison, can only be trained using a "small" subsample of 1000 instances, possibly introducing a bias due to insufficient training sizes. Further, the two balancing strategies RO and ADASYN are compared using two CNNs that are not tuned optimally. Because of their deep nature, training time is increased extensively, especially without access to GPU acceleration. Consequently, only a limited grid search could be performed that focused on the parameters of learning rate and batch size. This performance of manual grid search for a low number of epochs comes with another limitation. The identified hyperparameters are not necessarily the optimal ones for increased training times. As such, further analyses, should perform more extensive hyperparameter optimization for the models and adjust their structures, in order to avoid overfitting, increase performance, and enhance their ability to capture the subtle differences between Mars structures. This goes hand-in-hand with testing longer training duration. In particular, we hypothesize that the application of the neural networks to the imbalanced dataset requires longer training times as the validation accuracy is still rising after 30 epochs and shows high fluctuations. In this vein, the implementation of early stopping would allow for finding the optimal training duration.

Future research could consider implementing more regularization methods that were used in the original versions of the AlexNet and GoogLeNet but omitted in this paper for simplicity. These simplifications can be a factor contributing to the apparent overfitting in a number of models.

Because the researched imbalance strategies can work differently for other classifiers, this research can be extended by comparisons to additional methodologies. As such, benchmarks against pretrained CNNs, other CNN structures (e.g. ResNet-50), and other non-neural approaches (e.g. Random Forest) can be considered. Additionally, the impact of other balancing algorithms for example undersampling, SMOTE, and data augmentation is of interest in this domain. Finally, the used dataset of Mars landmarks is very challenging with a category that summarizes many different structures and not demonstrating consistent features.

All of the previously mentioned insights for the use of imbalance handling techniques in combination with CNNs can be useful to develop an enhanced classifier. NASA can integrate these insights into its existing pool of resources to develop a high-performing classifier, enabling it to analyze the surface of Mars with increased accuracy over existing approaches. This can foster space-related research and have a great impact on the knowledge captured about Mars and space itself. Our paper provides insightful knowledge for this as we identify that ADASYN and Random Oversampling in the data preparation for CNNs have a similar effect as using imbalanced data. Extending on this initial research by conducting analyses with access to increased computational resources, can allow to identify ways of enhancing the training processes of neural networks. In this context, taking into account our recommendations for future work and address-

ing the identified limitations serve as a significant starting point.

7 Conclusion

In this paper, the two imbalance handling methods ADASYN and Random Oversampling were applied to determine their ability to artificially increase datasets and thereby improve the performance of AlexNet and GoogLeNetet CNNs. The resulting datasets, after applying ADASYN, Random Oversampling, and the original without any modification, were then applied to these machine learning models. It can be conclusively stated that in the case of Mars surface recognition, artificially balanced datasets can be used to develop machine learning algorithms for landmark classification. Even though the application of either ADASYN or RO did not notably enhance the performance relative to the non-modified dataset in this specific instance, they are still considered a valid approach to balance and simultaneously enlarge a dataset. Furthermore, our findings attest to the capability of both GoogLeNet and AlexNet architectures in executing the classification process to generate valuable and accurate results, with GoogLeNet being especially effective. The evaluation suggests however to always be aware of similarities and overlaps between features of classes and to preprocess the data accordingly to avoid errors and misinterpretations.

References

A Ali-Gombe and Eyad Elyan. Mfc-gan: Class-imbalanced dataset classification using multiple fake class generative adversarial network. *Neurocomputing*, 361:212–221, 10 2019. ISSN 0925-2312. doi: 10.1016/J.NEUCOM.2019.06.043.

Kaichang Di, Yiliang Liu, Wenmin Hu, Zongyu Yue, and Zhaoqin Liu. Mars surface change detection from multi-temporal orbital images. *IOP Conference Series: Earth and Environmental Science*, 17: 012015, 03 2014. doi: 10.1088/1755-1315/17/1/012015.

Aurélien Géron. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: concepts, tools, and techniques to build intelligent systems.* O'Reilly Media, Inc, Beijing [China] ; Sebastopol, CA, second edition edition, 2019. ISBN 978-1-4920-3264-9.

California Institute of Technology. Mars reconnaissance orbiter - mars missions - nasa jet propulsion laboratory, 2005. URL `https://www.jpl.nasa.gov/missions/mars-reconnaissance-orbiter-mro`.

Sotiris Kotsiantis, P E Pintelas, and S Kotsiantis. Mixture of expert agents for handling imbalanced data sets active learning view project intelligent tutoring systems view project mixture of expert agents for handling imbalanced data sets. *COMPUTING & TELEINFORMATICS*, 1:46–55, 2003. URL `https://www.researchgate.net/publication/228084517`.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton. Imagenet classification with deep convolutional neural networks. *Neural Information Processing Systems*, 25, 01 2012. doi: 10.1145/3065386.

Azmi Rahman, Sri Prasetiyowati, and Yuliant Sibaroni. Performance analysis of the imbalanced data method on increasing the classification accuracy of the machine learning hybrid method. *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, 8:115–126, 02 2023. doi: 10.29100/jipi.v8i1.3286.

Md Shamim Reza and Jinwen Ma. Imbalanced histopathological breast cancer image classification with convolutional neural network. volume 2018-August, pages 619–624. Institute of Electrical and Electronics Engineers Inc., 2 2019. ISBN 9781538646724. doi: 10.1109/ICSP.2018.8652304.

Bofan Song, Shaobai Li, Sumsum Sunny, Keerthi Gurushanth, Pramila Mendonca, Nirza Mukhia, Sanjana Patrick, Shubha Gurudath, Subhashini Raghavan, Imchen Tsusennaro, Shirley T. Leivon, Trupti Kolur, Vivek Shetty, Vidya Bushan, Rohan Ramesh, Tyler Peterson, Vijay Pillai, Petra Wilder-Smith, Alben Sigamani, Amritha Suresh, Moni Abraham Kuriakose, Praveen Birur, and Rongguang Liang. Classification of imbalanced oral cancer image data from high-risk population. *Journal of Biomedical Optics*, 26, 10 2021. ISSN 15602281.

Chetan Swarup, Kamred Singh, Ankit Kumar, Saroj Pandey, Neeraj varshney, and Teekam Singh. Brain tumor detection using cnn, alexnet & googlenet ensembling learning approaches. *Electronic Research Archive*, 31:2900–2924, 03 2023. doi: 10.3934/era.2023146.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. pages 1–9, 06 2015. doi: 10.1109/CVPR.2015.7298594.