Comparison of data balancing techniques for vertebral compression fracture detection

Sejin Yi Yoo
Utrecht University
Utrecht, Netherlands
s.yiyoo@students.uu.nl

Abstract—**Multiple myeloma (MM) is a rare hematological disease that highly compromises the skeletal system, in the worst cases leading to vertebral compression fractures (VCFs). To avoid their progression surgical interventions are carried out although it is key that these VCFs are detected and treated as soon as possible. Automatic detection of these fractures using artificial intelligence (AI) could be very useful and has already been attempted. However, due to the prevalence of VCFs the data is highly imbalanced, which negatively affects deep learning models. In this study the performance of a convolutional neural network (CNN) VCF detection model is evaluated in different data imbalance ratios employing different data balancing techniques. The techniques compared are data augmentation, image generation employing latent diffusion models, a combination of augmented and diffusion generated images and cost sensitive learning. No data balancing technique showed a statistically significant improvement with respect to the baseline. Both data augmentation and diffusion sampled images hindered performance in a statistically significant way specifically in the higher imbalance ratios.**

I. Introduction

Multiple myeloma (MM) is a malignant proliferation of plasma cells in the bone marrow [1]. It is a rare hematological disease comprising 1 percent of all malignancies and 13 percent of hematological malignancies [2]. 80 percent of MM patients develop skeletal complications such as focal lytic lesions, hypercalcemia, and vertebral compression fractures (VCFs).

VCFs are a big healthcare concern in MM patients since they cause severe bone pain, limit mobility and increase spinal instability reducing the independence of the patients. VCFs not only cause a great decrease in quality of life, but are also linked to an increased mortality. Though current treatment lines such as radiotherapy, vertebroplasty or kyphoplasty are effective in delaying fracture progression, it is of utmost importance that they are applied as soon as possible. The reason for this is that surgical treatment reaches peak efficacy when performed before the actual appearance of the VCF. Thus, early detection is important even in diagnosed patients since 61% of diagnosed patients develop new VCFs [3].

There have been several attempts of VCF detection models on CT images employing a CNN [4] or a CNN combined with an recurrent neural network (RNN) [5]. Though they have achieved good performance the scenario of having sufficient training data to achieve these results might not be realistic in every clinical setting.

Due to the rare nature of the disease we are faced with the data imbalance challenge. There are several commonly used approaches to tackle this issue which can be at the algorithm level such as cost sensitive learning or at the data level such as data augmentation.

Recently diffusion models have gained a lot of attention for their widespread applications such as image inpaiting [6], anomaly detection [7] and image generation [8] [9]. In the latter, they have shown a great stability and ability to represent all the data distribution diversely overcoming generative adversarial networks'(GANs) main pitfall mode collapse [10]. These generative models have given rise to the possibility of image generation to create new samples of the minority class. This has been done successfully for other medical classification problems such as skin lesion [11] digitalized microscopic cells images [12] and chest X-ray images [13]. However, image generation has not been compared to other classically used approaches. Since it is a technique dependent on the amount of training data available, there is a possibility that it might not always be the most effective option. This comparison was done in the scope of non medical images using GANs [14] [15] but results were contradictory. Only Suh et al who had larger amounts of training data found GANs more effective than classical approaches. Thus, there is no comparison of diffusion models with respect to other data balancing techniques in medical images.

Therefore the aim of this study is to perform a comparison between employing diffusion models and other classically used data balancing techniques in the VCF detection problem. Plus, since previous studies did not come to consensus possibly due to differences in the amount of training data, imbalance ratio is a new variable introduced in the study. This comparison between data balancing techniques is repeated employing training data of varying imbalance ratios.

II. Methods

The steps taken in this study are the following:

- Train a baseline VCF detection model
- Improve this model using diffusion sampled images
- Separately use other data balancing techniques including data augmentation, a combination of data augmentation and diffusion sampling and cost sensitive learning

- Quantitatively compare the performance of each of this data balancing techniques with respect to the baseline VCF detection model
- Observe if imbalance ratio affects the results in this comparison by training the previously mentioned models with training sets of varying imbalance ratios to quantitatively assess their efficacy

All techniques needed to perform the mentioned steps will be explained below.

A. Detection model

As a baseline detection model a CNN similar to the LeNet architecture [16] is employed. This model employs several convolutional blocks consisting of convolutional layers followed by downsampling layers. The last layer is flattened and several fully connected layers are employed. A schematic of this simplified architecture can be observed in Figure 1.

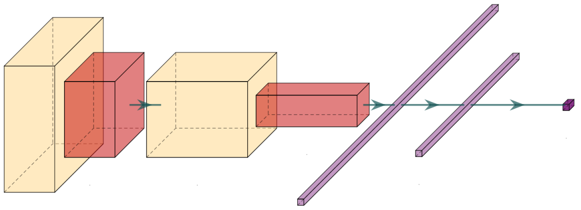


Fig. 1: Schematic of LeNet architecture. Convolutional layers in yellow, downsampling layers in red and fully connected layers in purple.

B. Generative AI

1) Diffusion models: The diffusion process learns the distribution of a dataset and is able to create new samples. It is separated into the forward and backward processes. The forward process q is a Markov chain where gaussian noise is added sequentially to the original image x_0 in a number of timesteps T to obtain $x_0, x_1, ..., x_T$ images. They are noised according to a fixed variance schedule $\beta_1 ... \beta_T$. This can be adjusted on different ways such as a linear scheduler or an exponential scheduler.

$$q(x_t|x_{t-1}) := N(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I) \tag{1}$$

As a result T images that are progressively noisier are obtained, the higher the timestep t, the lower the signal to noise ratio in x_t. By implementing a change of notation the forward process of any arbitrary timestep t can be performed in a non iterative way. Having $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$

$$q(x_t|x_0) = N(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I) \tag{2}$$

The reverse process p_θ aims to learn how to denoise the image between two timesteps.

$$p_\theta(x_{t-1}|x_t) := N(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \tag{3}$$

The variance Σ_θ is fixed as a hyperparameter like explained previously. Instead of predicting the remaining unknown μ_θ Ho et al [8] founds better results when predicting the noise ϵ instead. Thus the loss function to be optimized is a mean square error between the actual noise ϵ and the predicted noise $\epsilon_\theta(x_t, t)$.

$$L = E_{t,x_0,\epsilon}[||\epsilon - \epsilon_\theta(x_t, t)||^2] \tag{4}$$

Lastly, once the model is trained x_0 can be reconstructed from x_T by sequentially repeating the reverse process T times. Both the forward and backward process are illustrated in Figure 2.

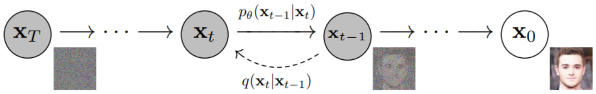

$\mathbf{x}_T$
$\mathbf{x}_t$
$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$
$\mathbf{x}_{t-1}$
$q(\mathbf{x}_t|\mathbf{x}_{t-1})$
$\mathbf{x}_0$

Fig. 2: Schematic visualization of forward (q) and backward (p) processes. The forward process (q) can be seen from right to left as noise is sequentially added to the original image. The backward process (p) can be seen from left to right as the image is sequentially denoised [8]

Once the model is trained it indirectly learns the distribution of the training dataset. Therefore, the trained reverse process can be used to generate new samples of the original distribution. This is done by inputting random noise to the reverse process which will be sequentially denoised till a new sample is reached. For every timestep t = T,..,1:

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)) + \sigma_t z \tag{5}$$

where z $\sim$ N(0,1) and $\sigma_t^2 = \beta_t$ or $\sigma_t^2 = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$ except when t = 1 for which z = 0

It should be noted that it is also possible to parametrize the variance Σ_θ instead of fixing it as a hyperparameter.

2) Latent diffusion models: As the number of timesteps employed can be on the scale of the thousands, the diffusion process can be quite computationally expensive. This along with other factors such as image size can become an issue. To avoid this latent diffusion models [17] can be employed. The key difference with this models is that prior to the diffusion process latent diffusion models employ an autoencoder to reduce the dimensionality of the images. This way the diffusion model can be trained on the smaller encoded images reducing computational expense. Nevertheless, the end results remain the same since the generated synthetic outputs are translated back to the pixel domain by the decoder as can be seen in Figure 3.

C. Data augmentation

Data augmentation is a technique commonly used to increase the amount and diversity of training data. It has been shown to increase performance and also serve as a form of regularization. It should be carried out with caution ensuring variability is introduced in a realistic way. Some of the commonly used augmentation techniques in medical images

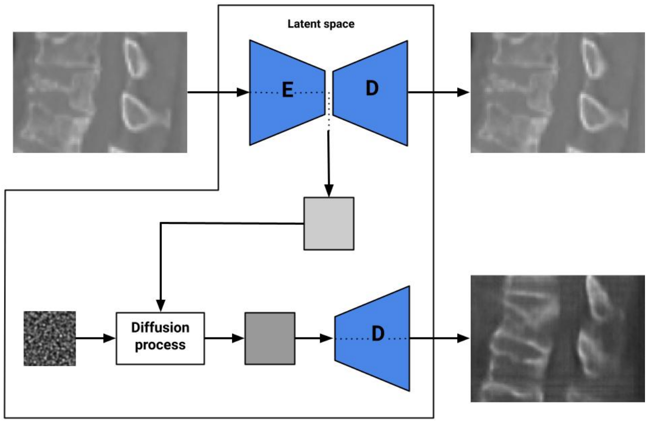


Fig. 3: Schematic of latent diffusion model's structure. Autoencoder is trained to reconstruct the training images. The diffusion model is trained as explained in Figure 2 on said encoded training images. During inference, random noise is input to the diffusion model to obtain a new sample which will be decoded back to the pixel domain

include geometric transformations, cropping, occlusion, intensity operations, noise injection and filtering [18].

D. Cost sensitive learning

In contrast with data augmentation, a data level approach towards dealing with data imbalance would be cost sensitive learning. It modifies the weight that each class has on the loss function so that a mistake made on the minority class is more punished, therefore balancing the lack of cases. The weight of each class can be made proportional to the imbalance present in the dataset [19] according to the formula:

$$w_j = \frac{n}{c * n_j} \tag{6}$$

where n is the total number of samples, c the number of classes and n_j is the number of samples in class j.

E. Performance evaluation

To individually assess the performance of each model, sensitivity and accuracy are employed. Sensitivity is defined as the fraction of the number of true positives and the sum of the true positives and false negatives. Accuracy is defined as the fraction of the sum of true positives and true negatives and the sum of true positives, true negatives, false positives and false negatives. To compare different models, a t-test is performed. P values lower than 0.05 are considered statistically significant.

III. Experiments

A. Materials

For this project, two datasets were used, the TotalSegmentator dataset [20] and the in-house dataset from the University Medical Center Utrecht. The TotalSegmentator dataset is an open source dataset comprising 1405 CT scans with segmentations of 104 body structures including bones, organs, muscles and vessels. These scans were obtained in a retrospective study, randomly sampled in 2012, 2016 and 2020 ensuring diversity in age, scanners, sites and sequences. The majority of these scans were thorax abdomen and pelvis CT though other scans such as heart CT were also present to a lesser extent [21]. The in-house longitudinal dataset contained 95 full body CT scans of 50 MM patients and their respective segmentations obtained by Payer (Coarse to Fine Vertebrae Localization and Segmentation with SpatialConfiguration-Net and U-Net) [22].

Two NVIDIA RTX A500 GPUs were employed for the training of all models.

B. Data preparation

All the operations were performed inside the MONAI framework [23].

Scans where the spine was not present were discarded as well as vertebras outside of C7-L4 since VCFs do not usually occur outside those bounds [1] [2]. Scans with incorrect segmentations were also discarded. Only in the case of VCF presence the incorrect segmentations were manually corrected. Scans with screws that had metal artifacts which highly compromised the visibility of the adjacent vertebras were also discarded.

Presence of VCFs was manually checked employing Genant semiquantitative assessment. This method classifies fractures according to vertebral height reduction into grade 0 ($< 20\%$ height loss) , grade 1 ($20 - 45\%$ height loss), and grade 2 ($> 40\%$ height loss) [24]. Vertebras classified as Genant grade 1 or 2 were labeled as VCF. Out of the 1405 scans in the TotalSegmentator dataset 756 were found to have correct segmentations. A total of 42 VCFs were found in 33 of the scans. The in-house dataset was also inspected to find a total of 355 VCFs out of 95 scans corresponding to 50 subjects.

Scans were preprocessed to have the same orientation and voxel dimensions 1x1x1 mm. They were also normalized between 0 and 1. Scans were cropped to obtain individual images of each vertebra of 112x112x80 voxels. These croppings were centered in the segmentation of each vertebra. The neighboring vertebras were visible to different extents.

The scans were split into the train, test and validation set in an 80-20-10 proportion. Due to the longitudinal nature of the in-house dataset the separation into train validation and test sets was done at the patient level.

C. Diffusion process

The previously explained latent diffusion model was trained to generate synthetic samples. It was trained with every VCF from every scan from every patient belonging to the train set which amounted to 383 VCFs. Sampling was executed to obtain images of 120x120x80 voxels. A manual inspection of the results was carried out to select the ones of sufficient sample quality. Out of the 3320 diffusion sampled images, 653 were selected. The selected ones were preprocessed in the same pipeline as the real samples.

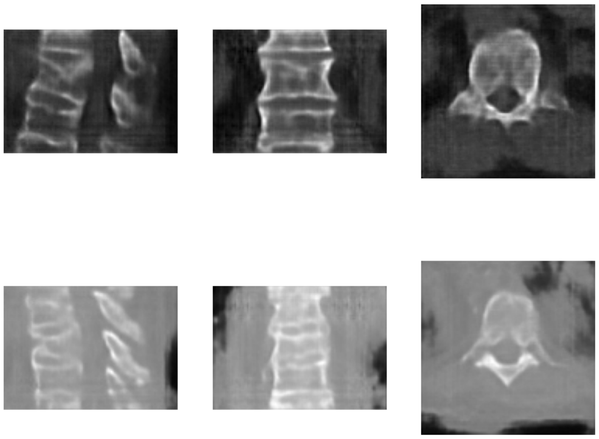


Fig. 4: Examples of selected diffusion sampled 3D images in their sagittal, frontal and coronal view from left to right

D. Data augmentation

As explained, data augmentation is a method used to artificially generate more samples by applying modifications which mimic changes in the acquisition process or the patient. Some of these techniques were applied. Geometric transformations include flipping the images in the sagittal axis, the only one where the spine has symmetry and cropping by applying random zooming. Intensity operations include gamma intensity correction and noise injections include gaussian noise. Lastly, filtering includes gaussian filters to smooth and sharpen the images. The results from these techniques can be observed in Figure 5.

E. Baseline detection

As a baseline detection model a CNN composed of two 3D convolutional layers each of them followed by a batch normalization layer and a maxpooling layer was employed. Dropout layers of 0.3 were also added for regularization. An Adam optimizer with *cross-entropyloss* was used. Early stopping evaluating sensitivity with patience of 20 epochs was implemented.

The performance of this baseline model was evaluated in different training conditions. Firstly, a gradual variation of the imbalance ratio in the training data was applied. Secondly, for each imbalance ratio the performance of the baseline model on its own as well as in combination of the different data balancing techniques was evaluated. The test set kept the imbalance ratio of the original dataset which was 0.03 (with 30 VCFs and 870 non VCFs) in all experiments.

For each imbalance ratio the performance of the following models was observed:

- The baseline VCF detection model
- The baseline VCF detection model trained with additional diffusion sampled images
- The baseline VCF detection model trained with data augmentation
- The baseline VCF detection model trained with both diffusion sampled images and augmented images
- The baseline VCF detection model employing cost sensitive learning

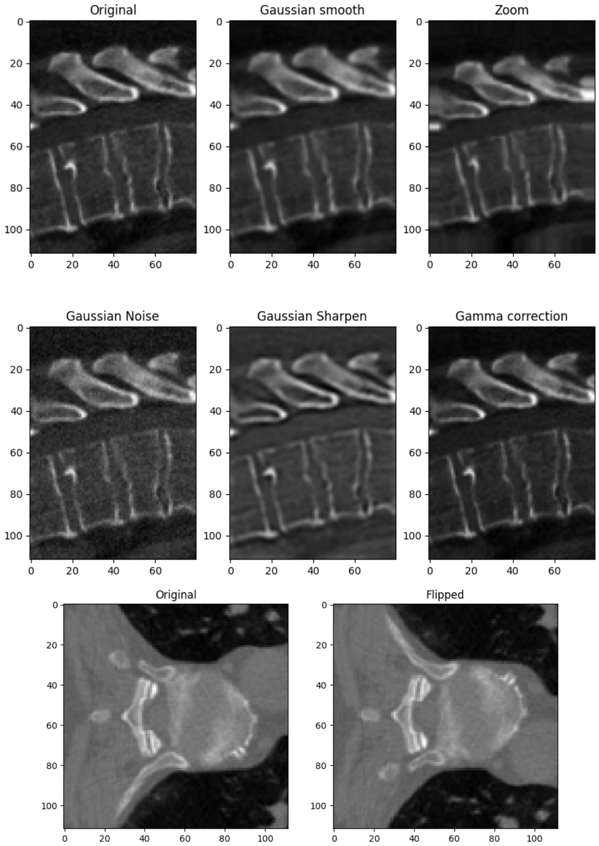


Fig. 5: Examples of all applied data augmentation techniques for the same image

The imbalance ratios can be simulated varying the selection of images from each class. This can be done in different ways which will be referred to as data matching and data equalizing. For data matching, starting with ratio 1 where the number of VCFs and non VCFs was the same, non VCF images were progressively added to reduce the ratio as observed in figure 5a). Then data augmentation and the diffusion sampled methods were employed to add as many images as needed to match the number of VCFs to non VCFs as illustrated in figure 5b). In this method all available VCFs were used but the number of images across ratios was not constant. For data equalizing the number of images was kept constant across ratios. The number of non VCF images would be reduced to keep the number of images stable as seen in figure 5 d). This method did not use all available VCFs.

Real training images, augmented images or diffusion sampled images to be used at each ratio, for each model were selected at random. To account for these variability and attempt to cover the possible data subsets bootstrapping was employed. 20 and 100 bootstraps of each model at each imbalance ratio

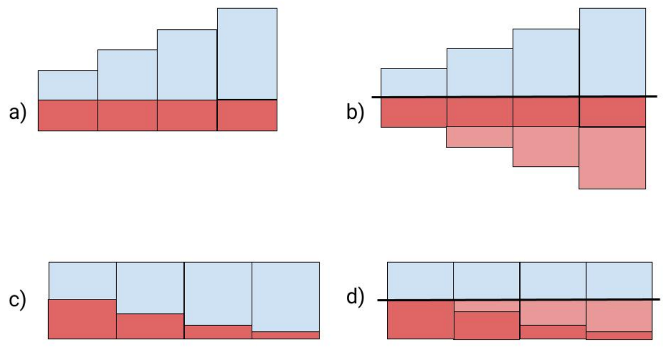


Fig. 6: Illustration of data matching and data equalizing, methods to simulate the imbalance ratios. In blue the non VCFs, in red the VCFs and in pink the synthetic VCFs whether diffusion sampled or augmented images.

were run using data matching and data equalizing respectively. Training was done with real and generated images (augmented and synthetic) and validation and testing on real images. A t-test was done to assess differences in performance in the test set with respect to the baseline.

IV. Results

For every experiment sensitivity and accuracy are reported below. Accuracy to account for the overall performance of the detection model and sensitivity to focus on the detection of the minority class.

A. Data matching

Table I presents the median sensitivity and interquartile ranges (IQR) of all models trained with data matching (number of images not constant). No method presented a statistically significant improvement. On ratio 0.256 (baseline median sensitivity 0.76 IQR [0.75-0.81]) there is a statistically significant decrease in sensitivity when employing data augmentation (median sensitivity 0.67 IQR [0.56-0.73]), diffusion (median sensitivity 0.69 IQR [0.62-0.73]) and both of them in combination (median sensitivity 0.69 IQR [0.62-0.76]) Table II presents the median accuracy and interquartile ranges (IQR) of all models trained with data matching (number of images not constant). Accuracy increased as the imbalance ratio increased for every model. No method presented a statistically significant improvement. Cost sensitive learning significantly decreases accuracy on ratio 0.256 with respect to the baseline (median accuracy 0.88 IQR [0.85-0.9]) with a median accuracy of 0.83 (IQR [0.77-0.87]. On ratio 0.5 there is a significant decrease in accuracy with respect to the baseline (median sensitivity 0.89 IQR [0.85-0.9]) when employing diffusion (median accuracy 0.82 IQR [0.78-0.87]) and diffusion in combination with data augmentation (median accuracy 0.85 IQR [0.73-0.86]).

B. Data equalizing

Table III presents the sensitivity of all models trained with data distributed with data equalizing. In all models sensitivity decreases as imbalance ratio increases. No data balancing technique increases performance in a statistically significant way. In ratios 0.06, 0.075 and 0.1 both diffusion and diffusion in combination with data augmentation cause a significant decrease in performance with respect to the baseline. The greatest decrease in sensitivity after applying a data balancing technique is seen in ratio 0.06 when employing diffusion. Median baseline sensitivity at the baseline is 0.67 IQR [0.57-0.76] and median diffusion sensitivity is 0.43 IQR [0.38-0.57].

Table IV presents the accuracy of all models trained with data distributed with data equalizing. In all models accuracy increases as imbalance ratio increases. All data balancing techniques decrease accuracy in a statistically significant way in the lower ratios (0.06, 0.075 and 0.1). Data augmentation and data augmentation in combination with diffusion also decrease performance significantly in ratio 0.5. The greatest decrease in performance is seen in ratio 0.075 for diffusion models with a median baseline accuracy of 0.92 IQR [0.89-0.95] and a diffusion median accuracy of 0.76 IQR [0.7-0.85].

V. Discussion

To sum up, data balancing techniques including, data augmentation, diffusion sampled images, a combination of data augmentation and diffusion sampled images and cost sensitive learning were compared. This was done in different imbalance ratios. These ratios were simulated in our training data employing two different methods, the key difference between them being that data equalizing kept the number of images constant in all ratios while decreasing the number of non VCFs trained on.

The first notable observation is the large variability in performance. One notable case is the baseline model trained with cost sensitive learning in ratio 0.06. There sensitivity ranged from 0.2 and 0.8 (See Figure 22 in the appendix). In most experiments for the sake of reproducibility and repeatability a seed is fixed. With this, the same subselections of data are used in every re-run. Having 100 bootstraps with random seeds allowed performance across these different subselections of data to be observed. Though the model should perform similarly in every subselection of data, we have seen that it is not always the case. Though the focus of this project was not to get good performances or particularly robust models, this calls our attention to how dependent a result can be on seed placement.

As for the performance of the model trained with the diffusion sampled images, it is surprising to see that not only it fails to surpass the baseline model but also it decreases performance in a significant way in both data matching and data equalizing. A possible reason for this is that the quality of the samples was not good enough. Though the selected samples presented the differentiating characteristics of VCFs some of the samples were lacking in aspects such as the realism of structures that

Ratio	Baseline	Augmentation	Diffusion	Augmentation+diffusion	CSL
0.256	**0.76[0.75-0.81]**	0.67[0.56-0.73]*	0.69[0.62-0.73]*	0.69[0.62-0.76]*	**0.76[0.76-0.80]**
0.5	**0.71[0.65-0.76]**	**0.71[0.67-0.76]**	**0.71[0.67-0.76]**	**0.71[0.61-0.77]**	**0.71[0.67-0.74]**
0.75	0.76[0.70-0.81]	0.71[0.65-0.77]	0.71[0.70-0.76]	0.71[0.55-0.76]	0.71[0.67-0.76]
1	**0.76[0.71-0.82]**	**0.76[0.70-0.81]**	0.74[0.67-0.81]	**0.76[0.71-0.81]**	**0.76[0.71-0.81]**

TABLE I: Data matching. Median sensitivity [Interquartile ranges (IQR)], CSL = cost sensitive learning. * marks statistically significant difference with respect to baseline $(p \leq 0.05)$.

Ratio	Baseline	Augmentation	Diffusion	Augmentation+diffusion	CSL
0.256	**0.88[0.85-0.9]**	0.87[0.84-0.9]	0.86[0.82-0.88]	0.87[0.83-0.92]	0.83[0.77-0.87]*
0.5	**0.89[0.85-0.9]**	0.85[0.82-0.88]	0.82[0.78-0.87]*	0.85[0.73-0.86]*	0.86[0.82-0.87]
0.75	0.82[0.75-0.85]	0.81[0.74-0.87]	0.79[0.78-0.84]	0.82[0.79-0.84]	**0.84[0.79-0.89]**
1	0.81[0.77-0.84]	0.8[0.76-0.83]	**0.83[0.8-0.85]**	0.78[0.76-0.8]	0.79[0.76-0.86]

TABLE II: Data matching. Median accuracy [IQR], CSL = cost sensitive learning. * marks statistically significant difference with respect to baseline $(p \leq 0.05)$.

Ratio	Baseline	Augmentation	Diffusion	Augmentation+diffusion	CSL
0.06	**0.67[0.57-0.76]**	0.62[0.46-0.71]	0.43[0.38-0.57]*	0.48[0.38-0.62]*	**0.67[0.52-0.76]**
0.075	**0.67[0.52-0.71]**	0.62[0.48-0.71]	0.5[0.38-0.57]*	0.52[0.43-0.62]*	**0.67[0.57-0.76]**
0.1	**0.67[0.57-0.76]**	0.64[0.52-0.76]	0.57[0.43-0.67]*	0.57[0.48-0.67]*	**0.67[0.57-0.76]**
0.25	**0.71[0.65-0.76]**	0.71[0.62-0.81]	0.67[0.57-0.71]	0.67[0.57-0.71]	**0.71[0.62-0.81]**
0.5	**0.76[0.67-0.81]**	**0.76[0.67-0.81]**	0.71[0.67-0.76]	0.71[0.67-0.76]	0.71[0.67-0.76]
0.75	**0.76[0.67-0.81]**	**0.76[0.67-0.81]**	**0.76[0.67-0.81]**	**0.76[0.67-0.81]**	0.71[0.67-0.76]
1	**0.76[0.71-0.81]**	**0.76[0.71-0.81]**	**0.76[0.71-0.81]**	**0.76[0.67-0.81]**	0.71[0.67-0.81]

TABLE III: Data equalizing. Median sensitivity [IQR], CSL = cost sensitive learning. * marks statistically significant difference with respect to baseline $(p \leq 0.05)$.

Ratio	Baseline	Augmentation	Diffusion	Augmentation+diffusion	CSL
0.06	**0.91[0.85-0.94]**	0.83[0.76-0.88]*	0.78[0.7-0.83]*	0.82[0.74-0.88]*	0.84[0.77-0.92]*
0.075	**0.92[0.89-0.95]**	0.82[0.76-0.86]*	0.76[0.7-0.85]*	0.83[0.76-0.88]*	0.85[0.76-0.91]*
0.1	**0.92[0.86-0.94]**	0.84[0.78-0.89]*	0.79[0.69-0.84]*	0.84[0.78-0.87]*	0.86[0.81-0.9]*
0.25	**0.87[0.8-0.9]**	0.79[0.74-0.85]	0.81[0.76-0.87]	0.82[0.75-0.86]	0.83[0.77-0.88]
0.5	0.84[0.79-0.88]	0.8[0.74-0.85]*	0.79[0.72-0.84]*	0.8[0.76-0.84]	**0.85[0.8-0.88]**
0.75	0.81[0.75-0.85]	0.81[0.75-0.87]	0.8[0.75-0.85]	0.79[0.72-0.84]	**0.82[0.78-0.86]**
1	0.8[0.74-0.84]	0.8[0.75-0.85]	0.8[0.74-0.84]	0.79[0.76-0.84]	**0.81[0.75-0.85]**

TABLE IV: Data equalizing. Median accuracy [IQR], CSL = cost sensitive learning. * marks statistically significant difference with respect to baseline $(p \leq 0.05)$.

were not bones or the overall blurriness of the image. In cases where the imbalance ratio was high the model would be learning mostly from the sampled images rather than the real ones. This means that the model might have been taking a shortcut, learning to differentiate between real and sampled images rather than healthy and VCF. This scenario would explain the low performances in the test set despite having a balanced dataset. Furthermore, the imbalance ratios were the performance was significantly worse for data equalizing seen in figure were in the lower ratios 0.06, 0.075, 0.1 and 0.25. These ratios are the ones where the number of generated training images are higher than the real training images. Thus, in those scenarios the model would be learning more from synthetic data than from real data.

On the other hand, the widely used technique data augmentation also did not increase performance in a significant way. This could be because it is not a technique that works well with spine images. Garcea et al showed that recent studies (2018-2022) have employed data augmentation successfully. However, out of these more than 300 studies only three of them were using spine images. Two of them where employing MR images and failed to improve the baseline (without data augmentation) and the remaining study using CT images did not account for the performance without data augmentation. Even though it is a widespread technique that is usually effective, it might not be suitable for every application. On the other hand, it might also be the case that the chosen parameters for the transformation resulted in changes that were too subtle. When doing so the purpose was to obtain realistic images but it might have resulted in changes that might not introduce enough variability.

As for cost sensitive learning, it did not increase nor decrease performance for the most part. The weights were derived from the number of instances of each class meaning the greater the imbalance the greater the action of the weights. In contrast with this simple approach Nguyen et al treated these weights as trainable parameters inside the model instead of fixing them [25]. This might be more appropriate approach since it is not only the imbalance ratio that determines the weight suitable to correctly detect the minority class.

Moreover, accuracy does not increase significantly with any technique. There is a trade-off between sensitivity and specificity and accuracy depends on both of them. However, the weight each of them hold will depend on prevalence. With a low prevalence, accuracy will reflect changes in specificity more than in sensitivity. As mentioned, due to the objective of improving the detection of VCFs, the focus was on increasing sensitivity. As a result of the trade-off between sensitivity and specificity and the higher number of non VCF images in the test set the overall accuracy decreased. Thus, this metric is not necessarily correlated to a good performance in a way that is meaningful to this context. This also serves as an explanation for the increasing of accuracy with higher imbalance ratios. Though the imbalance ratio in the test set is not changed during the experiments there is an inherent imbalance, and there are less VCFs than non VCFs. As the imbalance ratio increases in the training set, the model's ability to identify non VCFs increases as well improving specificity and therefore accuracy.

As for limitations, it should be noted that to balance all ratios the same pool of diffusion sampled images that were trained on all VCFs in the training set were used. When employing data equalizing, the only ratio where the experiment was depicted in a realistic way was ratio 1. This is because when simulating small ratios the diffusion process should have been trained only on the supposedly available VCFs. However training the diffusion model on 26 images would not yield good results. Thus though it was not technically realistic due to the lack of images the experiments were performed this way. In contrast, data matching does not present this issue. However, the comparison of performance in different ratios is biased. The reason for this is that the lower ratios have a larger quantity of training data. Though both methods have their pitfalls, a combination of both results enabled us to draw conclusions since in both of them adding the diffusion sampled images hindered the performance of the detection model.

Furthermore, it is worth mentioning that for the diffusion sampling process, less than 20% of the total diffusion samples were of acceptable quality. Out of the available evalutation metrics used for diffusion [26] only metrics that compare images individually were applicable. These were Structural Similarity Index or Mean Absolute Error. They were both tested as possible measures to filter out the lower quality samples. Each image was compared to a real one. However, there was no correlation between the scores obtained and the quality of the samples. As a consequence, a time consuming manual inspection was carried out. This was a subjective process influenced by external factors dependent on the examiner. To avoid this, more evaluation metrics that could bring an increased objectivity to this step would be useful asset to the field of generative AI.

It would be interesting to continue this project by employing explainable AI to observe the attention of the VCF detection model when employing diffusion sampled images. This way, confirmation that the sensitivity decreases due to the model taking shortcuts and learning to predict sampled or real instead of VCF or non VCF could be obtained. In that case the problem would be the quality of the samples. This would also explain why there was no consensus in the previous studies that employed diffusion models to balance data. Gladh et al found no improvement when using diffusion samples. The quantity of training data (hundreds of images) they employed to train the diffusion models was similar to our dataset. On the other hand Suh et al used different datasets that had 10 to 100 times more data (thousands and tens of thousands of images) and they found performance improvements when using diffusion sampled data. Therefore, our experiments could be repeated employing a greater quantity of data to create better quality samples. We could also assess the minimum number of images needed to successfully sample synthetic images realistic enough to avoid the shortcut taking behaviour. The diffusion sampling process could be trained with increasingly less images to observe the threshold number of cases there needs to be to obtain sufficiently good samples.

Lastly, in future works it would also be interesting to include other data balancing techniques such as undersampling or oversampling.

VI. Conclusions

The initial aim of comparing data balancing techniques in different imbalance ratios was achieved. As a result no data balancing technique was found to improve performance in a statistically significant way in the VCF detection model. Both data augmentation and diffusion significantly decreased performance mostly on the higher imbalance ratios. For diffusion the reason behind this might be a low quality in the sampled images enticing shortcut taking in the detection model. Cost sensitive learning did not increase nor decrease the performance of the VCF detection model significantly. Out of the tested methods none of them were effective when balancing the the VCF spine CT dataset. Future works include training the diffusion model with more images and employing other variations of cost sensitive learning.

References

[1] Mehmet Sonmez, Tulin Akagun, Murat Topbas, Umit Cobanoglu, Bircan Sonmez, Mustafa Yilmaz, Ercument Ovali, and Serdar Bedii Omay. Effect of pathologic fractures on survival in multiple myeloma patients: a case control study. *Journal of Experimental & Clinical Cancer Research*, 27:1–4, 2008.

[2] Jacob A Miller, Andrew Bowen, Megan V Morisada, Konstantinos Margetis, Daniel Lubelski, Isador H Lieberman, Edward C Benzel, and Thomas E Mroz. Radiologic and clinical characteristics of vertebral fractures in multiple myeloma. *The Spine Journal*, 15(10):2149–2156, 2015.

[3] Hester Zijlstra, Nienke Wolterbeek, Rosalin W Drost, Harry R Koene, Henk Jan van der Woude, Wim E Terpstra, Diyar Delawi, and Diederik HR Kempen. Identifying predictive factors for vertebral collapse fractures in multiple myeloma patients. *The Spine Journal*, 20(11):1832–1839, 2020.

[4] Sankaran Iyer, Arcot Sowmya, Alan Blair, Christopher White, Laughlin Dawes, and Daniel Moses. A novel approach to vertebral compression fracture detection using imitation learning and patch based convolutional neural network. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 726–730. IEEE, 2020.

[5] Amir Bar, Lior Wolf, Orna Bergman Amitai, Eyal Toledano, and Eldad Elnekave. Compression fractures detection on ct. In *Medical imaging 2017: computer-aided diagnosis*, volume 10134, pages 1036–1043. SPIE, 2017.

[6] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11461–11471, 2022.

[7] Julia Wolleb, Florentin Bieder, Robin Sandkühler, and Philippe C Cattin. Diffusion models for medical anomaly detection. In *International Conference on Medical image computing and computer-assisted intervention*, pages 35–45. Springer, 2022.

[8] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[9] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 8780–8794. Curran Associates, Inc., 2021.

[10] Hanqun Cao, Cheng Tan, Zhangyang Gao, Yilun Xu, Guangyong Chen, Pheng-Ann Heng, and Stan Z Li. A survey on generative diffusion model. *arXiv preprint arXiv:2209.02646*, 2022.

[11] Zhiwei Qin, Zhao Liu, Ping Zhu, and Yongbo Xue. A gan-based image synthesis method for skin lesion classification. *Computer Methods and Programs in Biomedicine*, 195:105568, 2020.

[12] David Kupas and Balazs Harangi. Solving the problem of imbalanced dataset with synthetic image generation for cell classification using deep learning. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 2981–2984. IEEE, 2021.

[13] Rogers Aloo, Atsuko Mutoh, Koichi Moriyama, Tohgoroh Matsui, and Nobuhiro Inuzuka. Ensemble method using real images, metadata and synthetic images for control of class imbalance in classification. *Artificial Life and Robotics*, 27(4):796–803, 2022.

[14] Sungho Suh, Haebom Lee, Paul Lukowicz, and Yong Oh Lee. Cegan: Classification enhancement generative adversarial networks for unraveling data imbalance problems. *Neural Networks*, 133:69–86, 2021.

[15] Marcus Gladh and Daniel Sahlin. Image synthesis using cyclegan to augment imbalanced data for multi-class weather classification, 2021.

[16] Yann LeCun et al. Lenet-5, convolutional neural networks. *URL: http://yann. lecun. com/exdb/lenet*, 20(5):14, 2015.

[17] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[18] Phillip Chlap, Hang Min, Nym Vandenberg, Jason Dowling, Lois Holloway, and Annette Haworth. A review of medical image data augmentation techniques for deep learning applications. *Journal of Medical Imaging and Radiation Oncology*, 65(5):545–563, 2021.

[19] Gary King and Langche Zeng. Logistic regression in rare events data. *Political analysis*, 9(2):137–163, 2001.

[20] Jakob Wasserthal. Dataset with segmentations of 104 important anatomical structures in 1204 ct images, 2022.

[21] Jakob Wasserthal, Hanns-Christian Breit, Manfred T Meyer, Maurice Pradella, Daniel Hinck, Alexander W Sauter, Tobias Heye, Daniel T Boll, Joshy Cyriac, Shan Yang, et al. Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images. *Radiology: Artificial Intelligence*, 5(5), 2023.

[22] Christian Payer, Darko Stern, Horst Bischof, and Martin Urschler. Coarse to fine vertebrae localization and segmentation with spatialconfiguration-net and u-net. In *VISIGRAPP (5: VISAPP)*, pages 124–133, 2020.

[23] M Jorge Cardoso, Wenqi Li, Richard Brown, Nic Ma, Eric Kerfoot, Yiheng Wang, Benjamin Murrey, Andriy Myronenko, Can Zhao, Dong Yang, et al. Monai: An open-source framework for deep learning in healthcare. *arXiv preprint arXiv:2211.02701*, 2022.

[24] Harry K Genant, Chun Y Wu, Cornelis Van Kuijk, and Michael C Nevitt. Vertebral fracture assessment using a semiquantitative technique. *Journal of bone and mineral research*, 8(9):1137–1148, 1993.

[25] Nguyen Thai-Nghe, Zeno Gantner, and Lars Schmidt-Thieme. Cost-sensitive learning methods for imbalanced data. In *The 2010 International joint conference on neural networks (IJCNN)*, pages 1–8. IEEE, 2010.

[26] Ajay Bandi, Pydi Venkata Satya Ramesh Adapa, and Yudu Eswar Vinay Pratap Kumar Kuchi. The power of generative ai: A review of requirements, models, input–output formats, evaluation metrics, and challenges. *Future Internet*, 15(8):260, 2023.

VII. Appendix

A. Extended results

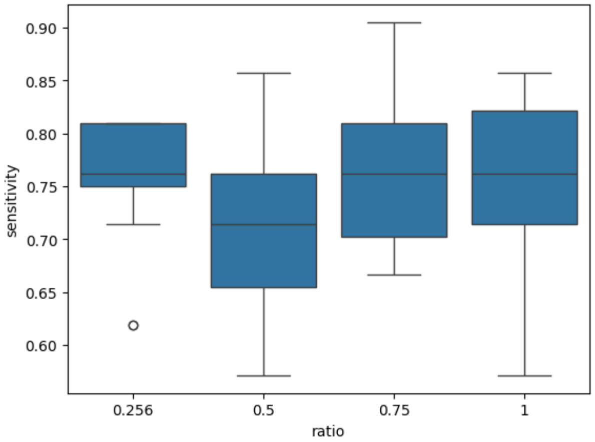


Fig. 7: Sensitivity of baseline (blue) for each of the imbalance ratios. Data matching.

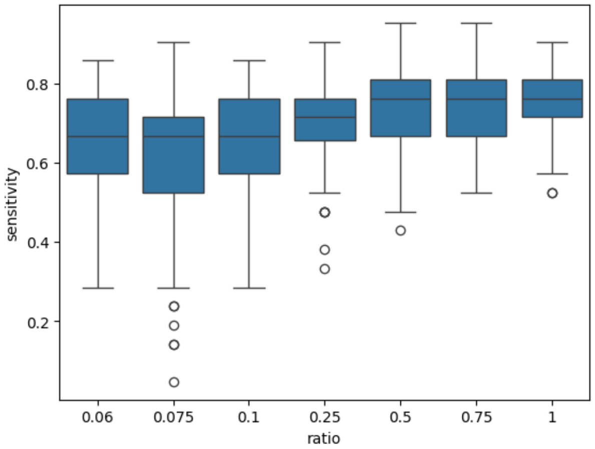


Fig. 8: Sensitivity of baseline (blue) for each of the imbalance ratios. Data equalizing.

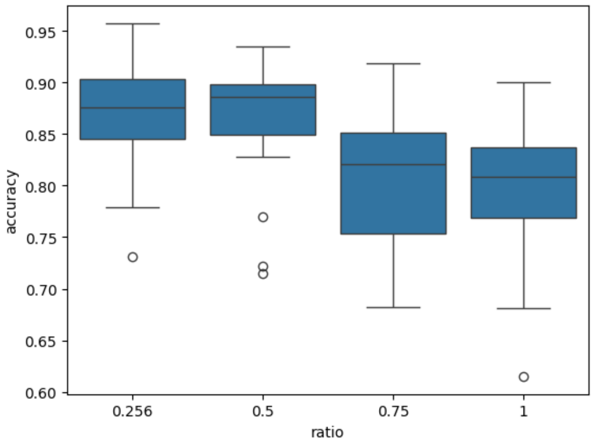


Fig. 9: Accuracy of baseline (blue) for each of the imbalance ratios. Data matching.

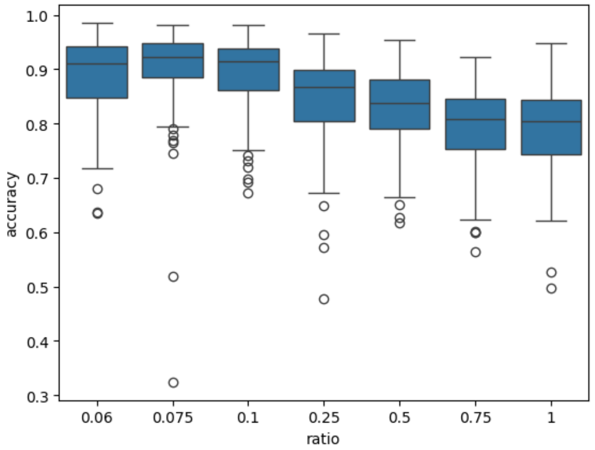


Fig. 10: Accuracy of baseline (blue) for each of the imbalance ratios. Data equalizing.

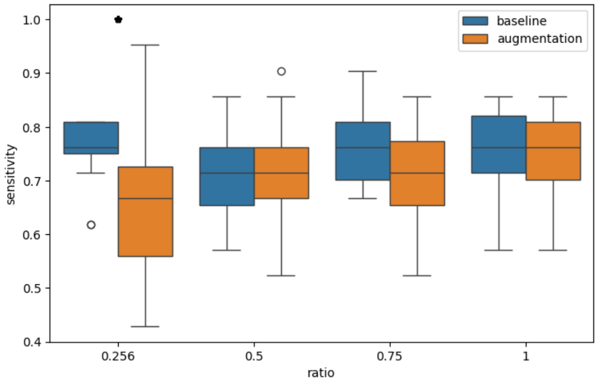


Fig. 11: Comparison of sensitivity of baseline (blue) and baseline trained with augmented images (orange) for each of the imbalance ratios. Data matching. Imbalance ratio 0.256 where there is a statistically significant difference is marked with a star

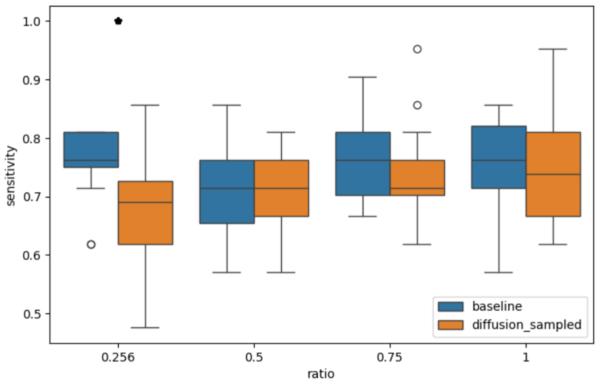


Fig. 12: Comparison of sensitivity of baseline (blue) and baseline trained with diffusion sampled images (orange) for each of the imbalance ratios. Data matching. Imbalance ratio 0.256 where there is a statistically significant difference is marked with a star

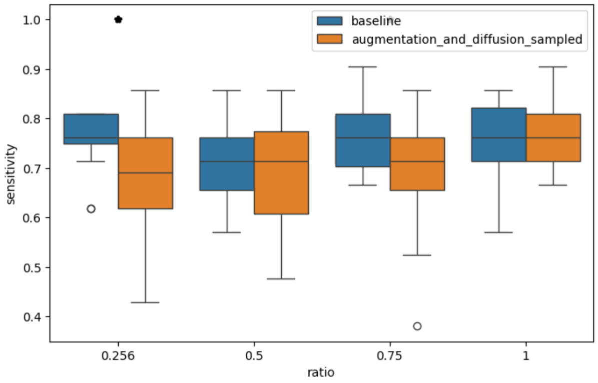


Fig. 13: Comparison of sensitivity of baseline (blue) and baseline trained with both diffusion sampled and augmented images (orange) for each of the imbalance ratios. Data matching. Imbalance ratio 0.256 where there is a statistically significant difference is marked with a star

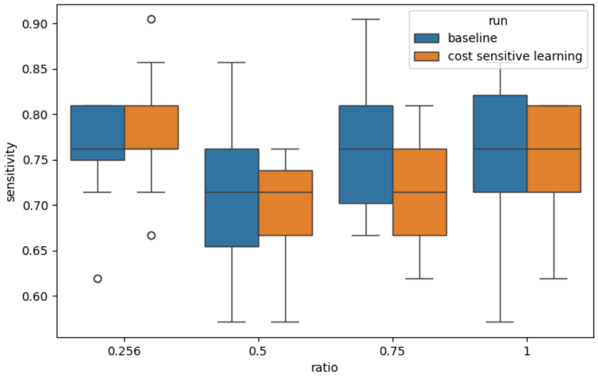


Fig. 14: Comparison of sensitivity of baseline (blue) and baseline trained with cost sensitive learning (orange) for each of the imbalance ratios. Data matching.

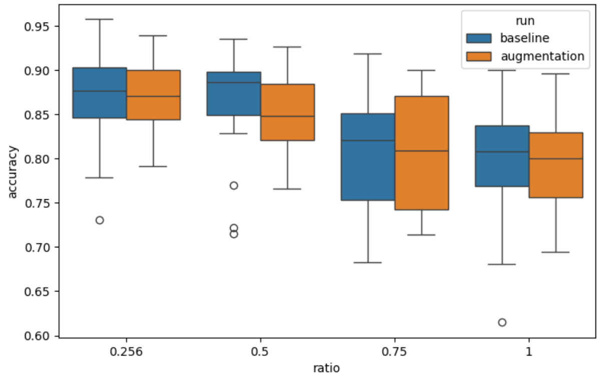


Fig. 15: Comparison of accuracy of baseline (blue) and baseline trained with augmented images (orange) for each of the imbalance ratios. Data matching.

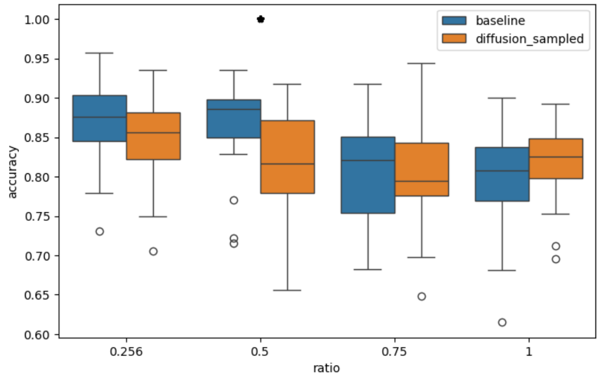


Fig. 16: Comparison of accuracy of baseline (blue) and baseline trained with diffusion sampled images (orange) for each of the imbalance ratios. Data matching. Imbalance ratio 0.5 where there is a statistically significant difference is marked with a star

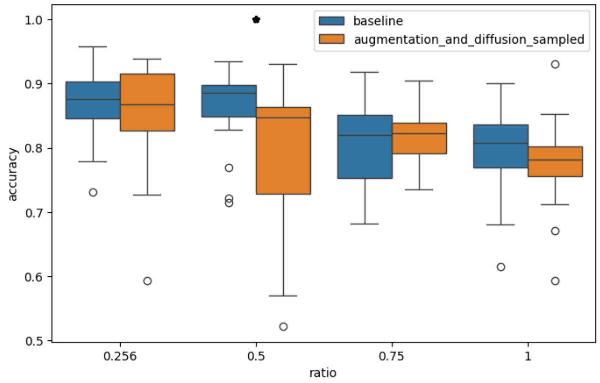


Fig. 17: Comparison of accuracy of baseline (blue) and baseline trained with both diffusion sampled and augmented images (orange) for each of the imbalance ratios. Data matching. Imbalance ratio 0.5 where there is a statistically significant difference is marked with a star

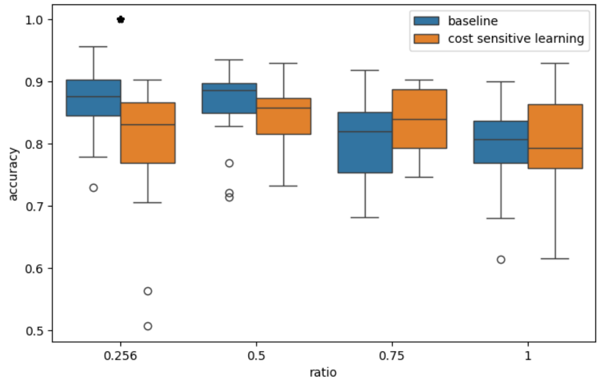


Fig. 18: Comparison of accuracy of baseline (blue) and baseline trained with cost sensitive learning (orange) for each of the imbalance ratios. Data matching. Imbalance ratio 0.256 where there is a statistically significant difference is marked with a star

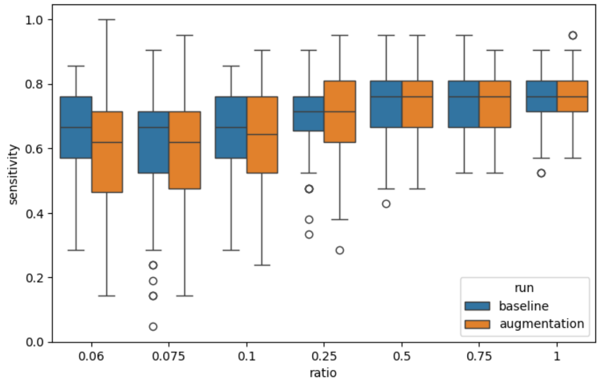


Fig. 19: Comparison of sensitivity of baseline (blue) and baseline trained with augmented images (orange) for each of the imbalance ratios. Data equalizing.

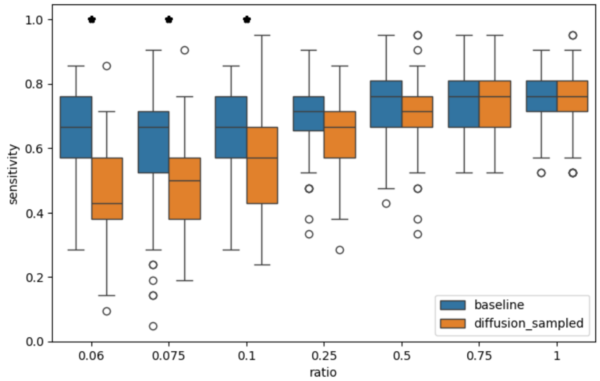


Fig. 20: Comparison of sensitivity of baseline (blue) and baseline trained with diffusion sampled images (orange) for each of the imbalance ratios. Data equalizing.

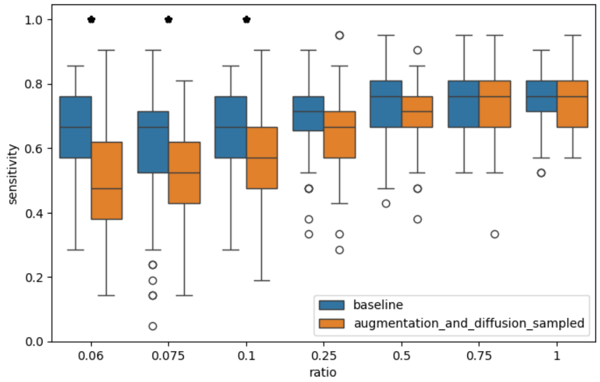


Fig. 21: Comparison of sensitivity of baseline (blue) and baseline trained with both diffusion sampled and augmented images (orange) for each of the imbalance ratios. Data equalizing. Imbalance ratios 0.06, 0.075 and 0.1 where there is a statistically significant difference are marked with a star.

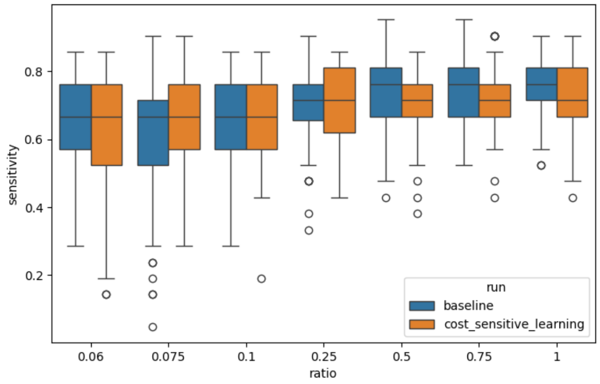


Fig. 22: Comparison of sensitivity of baseline (blue) and baseline employing with cost sensitive learning (orange) for each of the imbalance ratios. Data equalizing. Imbalance ratios 0.075 and 0.1 where there is a statistically significant difference are marked with a star.

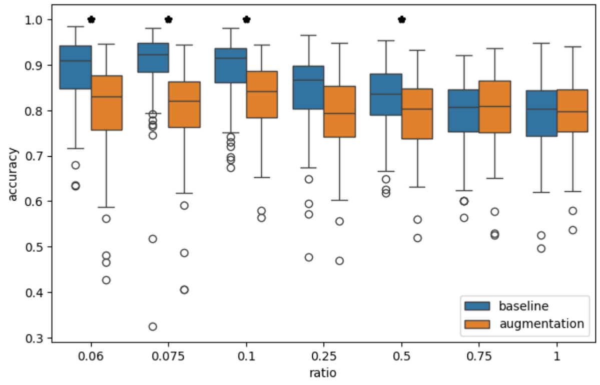


Fig. 23: Comparison of accuracy of baseline (blue) and baseline trained with augmented images (orange) for each of the imbalance ratios. Data equalizing. Imbalance ratios 0.06, 0.075, 0.1 and 0.5 where there is a statistically significant difference are marked with a star.

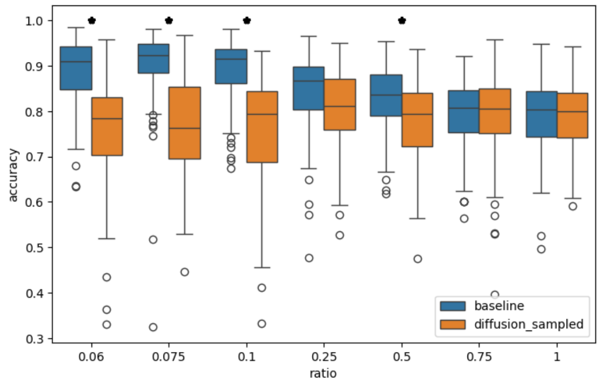


Fig. 24: Comparison of accuracy of baseline (blue) and baseline trained with diffusion sampled images (orange) for each of the imbalance ratios. Data equalizing. Imbalance ratios 0.06, 0.075, 0.1 and 0.5 where there is a statistically significant difference are marked with a star.

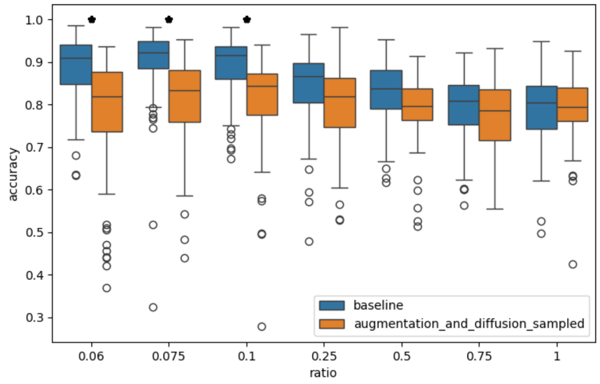


Fig. 25: Comparison of accuracy of baseline (blue) and baseline trained with both diffusion sampled and augmented images (orange) for each of the imbalance ratios. Data equalizing. Imbalance ratios 0.06, 0.075 and 0.1 where there is a statistically significant difference are marked with a star.

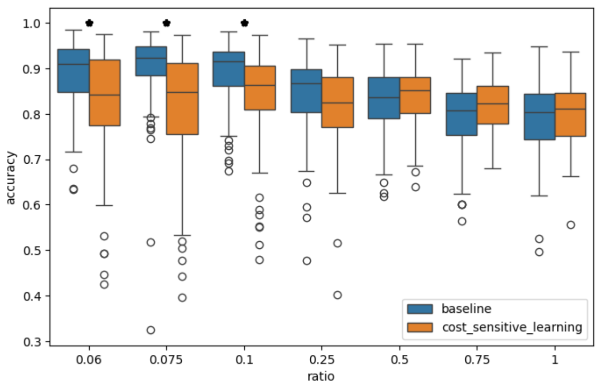
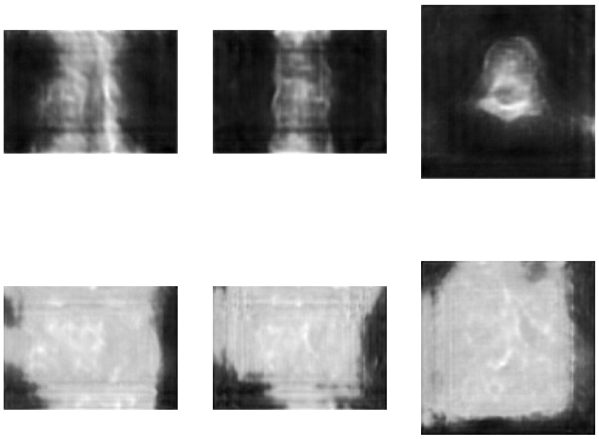


Fig. 26: Comparison of accuracy of baseline (blue) and baseline employing with cost sensitive learning (orange) for each of the imbalance ratios. Data equalizing. Imbalance ratios 0.06, 0.075 and 0.1 where there is a statistically significant difference are marked with a star.

B. Examples of discarded diffusion sampled VCFs



C. Examples of diffusion sampled VCFs