University of Sheffield

Improvement of confidence estimation of machine learning algorithms for cardiac MRI classification

Mateusz Grzybowski

Supervisor: Dr Haiping Lu

Dissertation submitted as the requirement for the BEng Bioengineering Research Project

Interdisciplinary Programmes in Engineering
May 12, 2022

Abstract

Reliable confidence estimation is one of the challenges for machine learning systems in applications where the risk of false predictions could cause severe repercussions. Classification of medical image data is one example of this and usually is accompanied by a class imbalancement. The main objective of this dissertation was to develop and test methods which could influence positively the confidence estimation of machine learning algorithms. For this study, SVC, logistic regression and ResNet-18 were considered where the first two were used as a part of the MPCA pipeline available in the PyKale library. Platt scaling was used as a calibration method which by itself has produced a better calibration of SVC and ResNet-18 algorithms. Additionally, a new method consisting of the Platt scaling and selective classification with the selective function was implemented and compared against calibration and imbalancement. As a third method, simple data augmentation techniques were used as a potential technique to improve confidence estimation without direct interference with a model. Results suggest that these methods are capable of improving the confidence estimation of predictive models. Although data augmentation and the new selective approach did not improve the calibration of the classical machine learning algorithms, they were still effective for the deep neural network. Furthermore, affine transformations and selective models were found effective in confronting the data imbalancement problem. In the end, interesting research directions were proposed which have a huge potential for an improvement of the research field hence increasing the reliability of the predictive systems.

Acknowledgements

I would like to thank my supervisor Haiping Lu for his guidance and indulgence whenever I was lost in the scope of the dissertation. His recommendations and explanations have accelerated my learning experience.

Many thanks to Lawrence Schöbs who helped me with the technical aspects of the project, and proposed many upgrades that yielded a higher quality content and an interpretation of an academic paper that became a substantial part of this research.

I am grateful for the support provided by my friends and family who were always around me whenever I was stressed or demotivated.

Author's Declaration

All sentences quoted in this final report from other's people work is specifically acknowledged by clear cross-referencing to the author, work and page(s). Any illustrations that are not the work of the author of this report have been used with the explicit permission of the originator and are specifically acknowledged. I understand that failure to do this amounts to plagiarism and will be considered grounds for failure in this project and the degree examination as a whole.

Signature:

Date: May 12, 2022

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1 Research Purpose

In 1895 the first medical modality was introduced by Wilhelm Conrad Röntgen, X-rays. For the first time, these allowed us to observe human skeletons without taking invasive measures. After this discovery, a variety of medical imaging techniques such as magnetic resonance (MRI) or computer tomography (CT) were developed and widely used for diagnosis. Different modalities have better performance and predispositions for capturing different parts of the human body [1]. In a 2015 study [2], it was estimated that one billion radiologic examinations are performed annually worldwide of which most are performed by radiologists. Understanding the nature of the classification problem requires appropriate facilities, specialist knowledge and experience. This implies sophistication of radiology and due to that fact, there might be a problem with the interpretability of images. Hence radiologist reports are more of a clinical consultation with conclusions produced based on evidence than a definite judgement [3]. MRI has been proven to be effective for depicting cardiovascular structures [4]. For decades, MRI moved from a research tool to an approved, safe and comprehensive imaging modality [5]. Its performance is particularly excellent for visualising soft tissues such as muscles that build the human heart [1].

However, even if radiologists are using world-class equipment, their classification is prone to be mistaken due to missed abnormalities which sometimes are difficult to detect with a human eye. Taking into account the rate of medical images taken annually, a massive amount of image data has been generated. These are the reasons why more deterministic methods are being investigated for producing more objective decisions. Undoubtedly, the most popular method is the application of machine learning for medical image analysis. In this context usually, it can be divided into two ways of application: supervised and unsupervised learning. This research is going to focus on the first as classification is a problem that requires labelling. Machine learning allows computers to learn the most important features and predict the label based on data whereas radiologists need a deeper understanding of a medical subject. In addition, algorithms in some cases are capable of noticing patterns beyond human perception [6]. On the other hand, in classification machine learning algorithms produce an integer that only answers a question of class belonging. For some applications, this might be sufficient but the medical diagnosis should not be regarded as a com-

prehensive decision. The reason is that a hypothetical patient could be classified with a lethal disease but the probability of this exact decision would be relatively low. Therefore, the main motivation of this dissertation is to address the importance of probability estimation for medical classification and examine techniques that could improve it for both classical machine learning and deep learning.

1.2 Report Structure

- **Chapter 2** introduces a theoretical knowledge and literature review of available methods used behind this research project.

- **Chapter 3** presents the requirements of the study and analysis of the dataset. Additionally, limitations to the research are going to be provided.

- **Chapter 4** establishes necessary methods for the experiments such as specific algorithms and metrics and brief motivation and explanation behind these particular choices.

- **Chapter 5** shows the results of the conducted experiments.

- **Chapter 6** delivers an interpretation of the results, challenges faced whilst their conduction and comparison to the state of the art. Moreover, a few suggestions for future research are provided.

- **Chapter 7** concludes the research.

- **Chapter 8** examines the management of the project.

- **Chapter 9** concludes the author's performance with a descriptive and honest self-review.

CHAPTER 2

Literature Review

2.1 Confidence Estimation

Whenever a machine learning model is trained well, it will classify data with a high accuracy which is a sufficient criterion for deployment in many applications. However, this approach brings severe consequences in fields relying on confidence of risk such as medical diagnosis. For this specific example, we would rather receive a prediction for our condition in terms of probabilities than a number denoting class belonging. It is better to receive information from a doctor "There is an 80% chance that you have a disease" than "You have been classified with a medical disease.". This is why confidence estimation not only improves the reliability of predictions but also the interpretability of results. If it is desired to rely on algorithms for making medical diagnostic decisions, then machine learning engineers should aim to make them as much reliable as possible. There have been defined two main use cases for confidence estimation: selective classification and confidence calibration [7]. So far, there is little research conducted on the improvement of confidence estimation in the medical imaging classification context.

2.1.1 Selective Classification

As presented in [7], selective classification (also known as a reject option) is an ability to recognise what we don't know which might support learning. The motivation behind it is to reduce the error rate by restraining from making a prediction when confidence estimation is too low depending on a selection threshold for a particular classifier. A selective classifier consists of two functions, a standard classifier and a rejection function defined by the following:

$$(f, g)(x) = \begin{cases} f(x), & \text{if } g(x) = 1. \\ restrain, & \text{if } g(x) = 0. \end{cases} \tag{2.1}$$

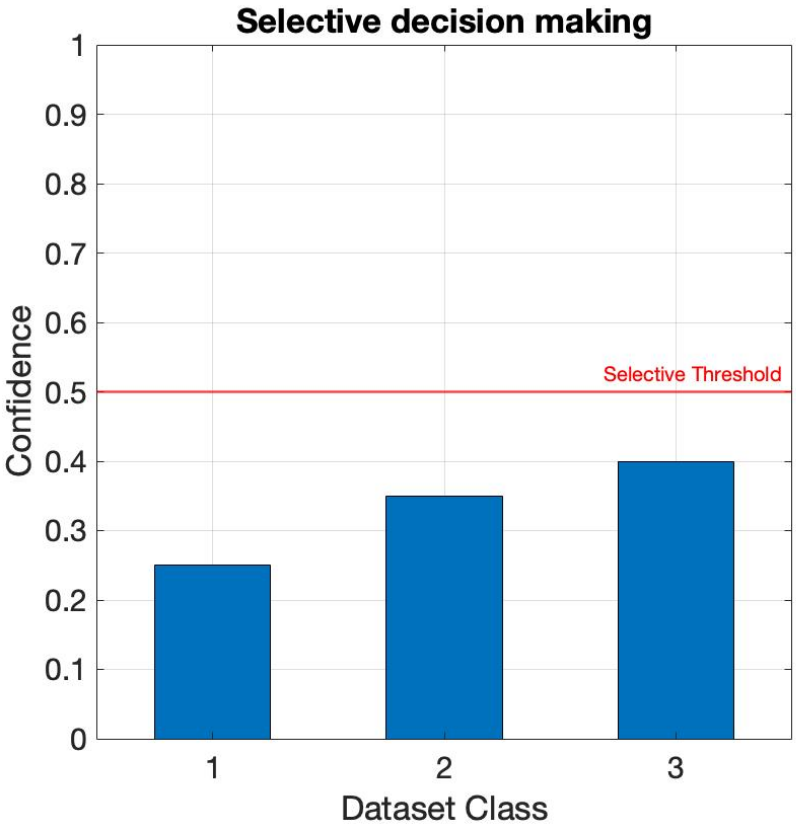


Figure 2.1: Bar plot of an exemplary confidence outputs for 3 class case. Selective threshold denotes a decision making value which unless exceeded, a prediction will be dropped.

A selective function $g(x)$ uses a predefined selective threshold θ to decide whether a prediction from a classifier should be dropped. For example, the prediction from Figure 2.1 has maximal confidence for class 3 but it is still below the selective threshold. In this case, the prediction would have abstained as a classifier was not certain enough to make a quality decision. Moreover, it is essential to determine an appropriate threshold to make the implementation effective. In [8], the selection with guaranteed risk (SGR) algorithm was established as a method to find the optimal threshold for selective function construction.

2.1.2 Confidence Calibration

Confidence calibration aims to make probabilistic predictions that match the perfect calibrated line as possible which is defined by:

$$P(\hat{Y} = Y|\hat{P} = p) = p \qquad \forall p \in [0, 1] \tag{2.2}$$

where $\hat{Y}$ is a predicted label and $\hat{P}$ is a confidence estimation. The meaning of Equation 2.2 is that expected accuracy is a function of confidence. For example, if there were 100 samples and a model was 50% confident in its predictions, then it would be expected to get only half of the samples classified correctly. Although this improves confidence estimation like selective prediction, it is unknown if they are connected [7]. The main difference between these is that confidence calibration is conducted after training and selective prediction is ideally performed during the training process which is not always possible as indicated in [8].

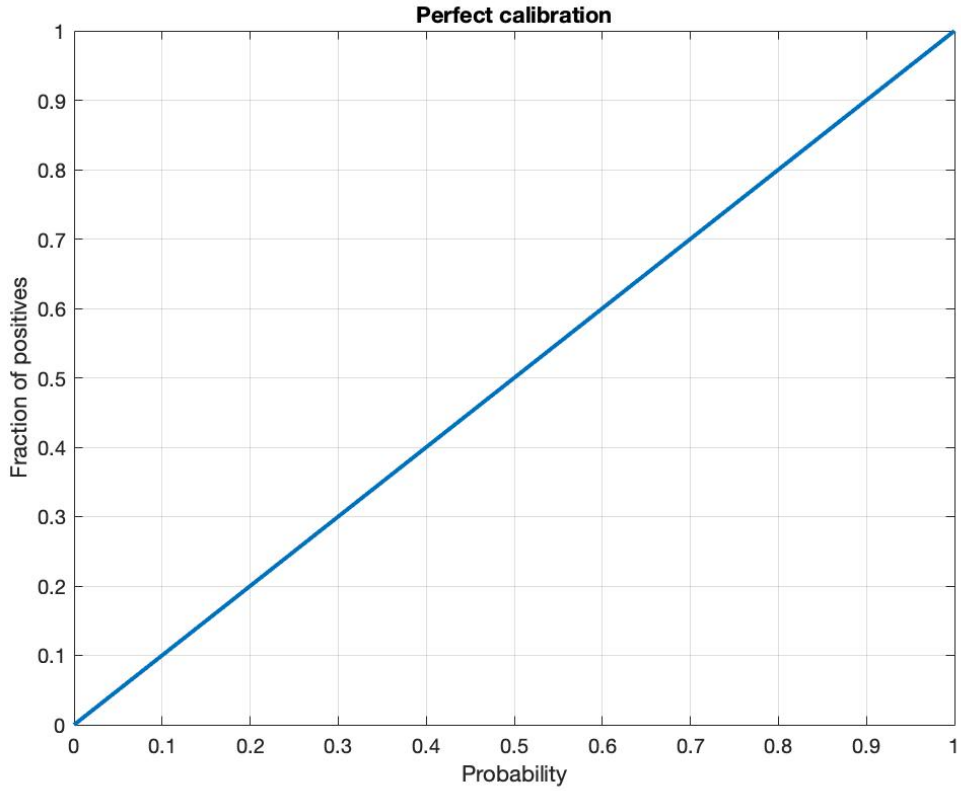


Figure 2.2: Perfect calibration curve which is desired to achieve when calibrating a machine learning model.

However, in real-world applications, achieving perfect calibration is impossible and potential reasons for this are that data usually is disturbed, noisy and might contain sampling bias and algorithms seldom predict perfectly. Also, test data may be out of distribution compared to train data.

Currently, papers on confidence calibration mostly research neural networks. The reason is that state-of-the-art deep learning architectures for classification such as ResNet tend to be outstandingly accurate but poorly calibrated [9]. Additionally, Thulasidasan et al. in [10] observed that deep neural networks (DNNs) transit from under-confidence at the beginning of training towards overconfidence at the end. Nota bene, classical machine learning methodologies such as SVM also tend to produce miscalibrated estimations. The only found exception is logistic regression which often produces well-calibrated probabilities [11]. These justify the goal of the thesis to explore methods for confidence improvement for both types of machine learning methodologies.

2.2 Machine Learning

Machine learning is a field in computer science which researches algorithmic methods for computers to learn without being explicitly programmed. This is regarded as a subfield of publicly known artificial intelligence. Over decades, machine learning solutions evolved drastically from simplistic perceptron models to modern deep learning architectures. Due to the lack of sufficient computational power, the research pace in this field was slow. Nowadays after acquiring better hardware the interest in machine learning revived and has become popular in public.

Although this study focuses more on techniques for confidence estimation, the choice of an appropriate set of algorithms and data processing methods is crucial to obtaining meaningful results. Thus, both classical machine learning and deep learning are considered.

2.2.1 Feature Extraction

Due to the small number of samples, the following preprocessing pipeline had to be applied for classical machine learning algorithms.

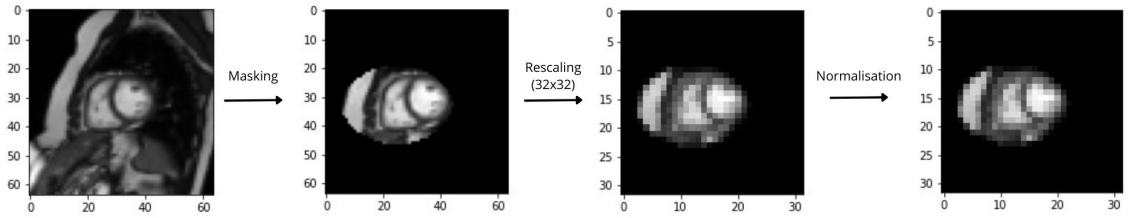


Figure 2.3: Preprocessing pipeline of MRI images which consists of i) masking ii) rescaling and iii) normalisation.

Masking is applied to make algorithms focus on areas that are the most important for classification. Rescaling was used to reduce the number of features that needed to be learnt and then normalised to optimise the computation of weights hence lowering the computational time and improving a model [12]. What is important, this preprocessing pipeline applies only to classical machine learning algorithms because deep learning architectures extract features automatically. After preprocessing, data could be implemented to train a model. However, this careless decision could with a high probability cause a model to overfit. This means a model would have a high variance to unseen data hence producing huge errors when making a prediction. Our goal is to create a model that will generalise data in the best way possible. One could think of generating more data to overcome the overfitting issue. Unfortunately, sometimes generation of more samples can be difficult or expensive. Moreover, if an engineer decided at some point to append an additional feature to a dataset, then the demand for data would drastically increase. This phenomenon is widely known as the curse of dimensionality. One way to cope with this problem is to select only these features that contain the most information to build an efficient model. Feature extraction specifically aims to reduce the dimensionality of data by keeping features that hold the most information [13]. This can be understood as a lossy compression since we are keeping the most essential parts of data.

Principal component analysis (PCA) is an unsupervised learning algorithm that is one of the most popular choices for extracting features since it is a white-box method and relatively easy to understand. However, it becomes more sophisticated to implement for tensor data such as MRI images. Specifically to handle tensor data, a multilinear PCA (MPCA) has already been implemented for cardiac MRI dataset which one data point had I x J x K shape [12]. This method has successfully extracted the most valuable features hence enabling the production of good predicting models. However, there are restrictions to this method. First of all, it is obligatory to scale data before feeding it to the PCA, because the algorithm is susceptible to different scales of features [13].

2.2.2 Feature Selection

Features selection methods sometimes are interchangeably called feature extraction. It is not correct as feature extraction creates a new feature subspace of lower dimension from an original

feature space. The feature selection task is to select a subset of variables without changing the original dimension of the dataset. It can be used to find objectively the best features. For this dissertation, a discriminative feature selection was implemented as a part of the MPCA pipeline.

2.2.3 Algorithms

Machine learning clearly during the last few decades has been divided into two sub-fields, classical machine learning and deep learning. The difference mainly is in the mathematical definition of algorithms. In classical machine learning, there are many different methodologies to achieve various tasks such as classification, clustering and regression based on mathematical and statistical modelling. Deep learning on the other hand relies on architectures that have been proved to be effective for desired tasks. For medical image classification, there is a debate on which type of algorithms should be used. One side states that deep learning is prone to overfit and due to the highly black-box nature it is difficult to understand what parts of a medical image made the biggest impact on classification decision [12]. Another argument against the use of deep learning is the computational time which could be longer than a manual prediction by experienced radiologists.

On the other hand, deep learning architectures have shown an accuracy comparable to humans in pattern recognition. Moreover, deep learning automatically extracts the most important features [14]. Also, there are techniques such as transfer learning and dropout that support deep learning models to avoid overfitting and achieve superb performance [15]. Therefore, this study applied calibration methods to both types of machine learning algorithms.

Support Vector Machines

SVMs are one of the most commonly used classifiers in image classification because they have a good generalization capability [1]. Furthermore, the optimisation problem of SVM is convex and the solution is unique. Particularly this is an advantage over deep learning models in which optimisation is rather local [16]. In addition, SVM performs well on high dimensional data which is a valuable trait when approaching a problem related to medical imaging [17].

The SVM is built on the concept of decision boundary also known as a hyperplane and its margins. The goal is to maximise the margin denoted by ξ around a decision boundary by minimising the magnitude of the normal vector of the boundary which is denoted by the following equation:

$$w^T x + b = 0 \tag{2.3}$$

The distance from the hyperplane can be written as:

$$d_H(x) = \frac{|w^T x + b|}{||w||_2} \tag{2.4}$$

The optimal solution for SVM is mathematically denoted by:

$$w^* = \underset{w}{\operatorname{argmax}}[\min_n d_H(x_n)] \tag{2.5}$$

However, SVM like any algorithm has some disadvantages. The problem is the time complexity of optimisation of SVM which is $O(n^3)$ [18]. This implies that SVM is more computationally taxing whenever an analysed dataset becomes bigger. Although it is not a huge issue for medical imaging classification since datasets usually are relatively small to image datasets from other in-

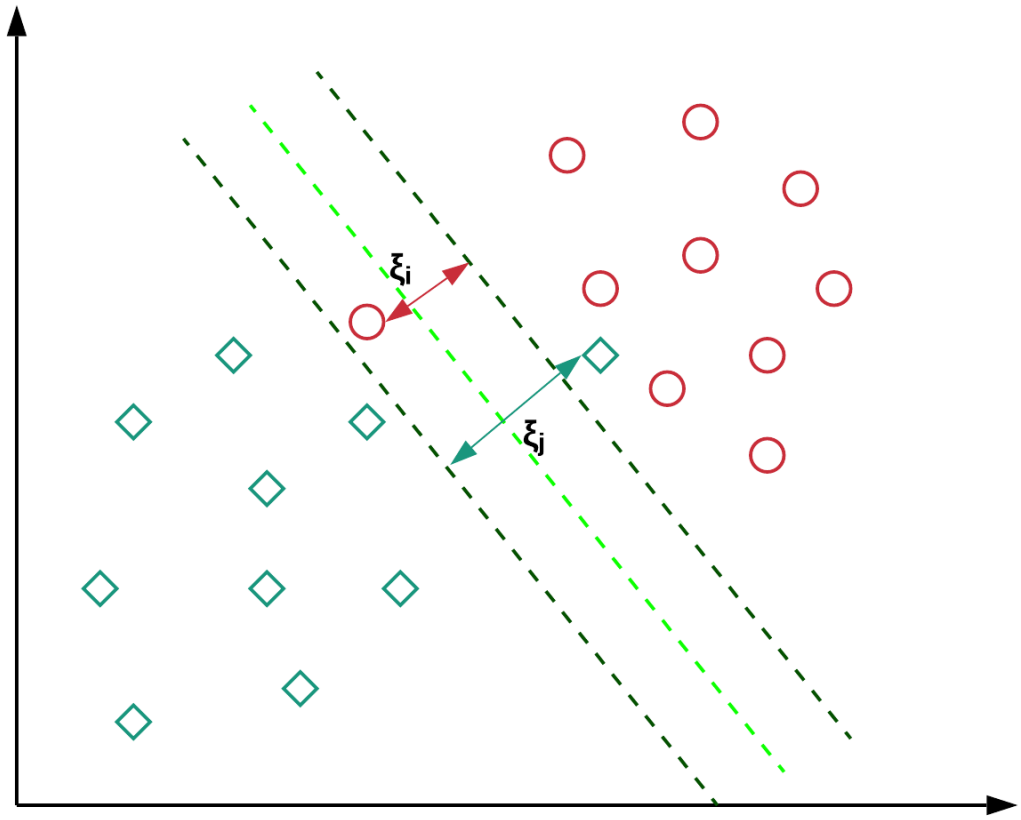


Figure 2.4: Diagram of SVM and its hyperplanes. Source: https://towardsdatascience.com/support-vector-machines-soft-margin-formulation-and-kernel-trick-4c9729dc8efe

dustries. SVM fails when a dataset is significantly imbalanced. This is a potential issue for this research when performing experiments with this classifier. Therefore, a cautious examination of the dataset and the algorithm should be performed against this particular problem. Lastly, SVM produces uncalibrated outputs that cannot be interpreted as probabilities [19]. It is a substantial issue for a reliable diagnosis, but calibration methods could potentially solve this which is going to be investigated in one of the experiments. Nevertheless, it is still a great choice for medical image classification.

Logistic Regression

Logistic regression is one of the oldest and most recognisable classical machine learning algorithms. One of the advantages of this algorithm is that it often automatically produces well-calibrated probabilities but it is not guaranteed in general [11]. The second advantage is the transparent nature of the algorithm which might be helpful to comprehend how decisions for classifications were made. Furthermore, the time complexity of training logistic regression is $O(nd)$, where n and d are the numbers of samples and features respectively. This means that this methodology is much cheaper computationally than SVM or deep neural networks. What is more, logistic regression predicts a label of data by multiplying it with trained weights and introducing this output to the sigmoidal function. Therefore, predictions are produced rapidly which is an advantage over manual diagnosis.

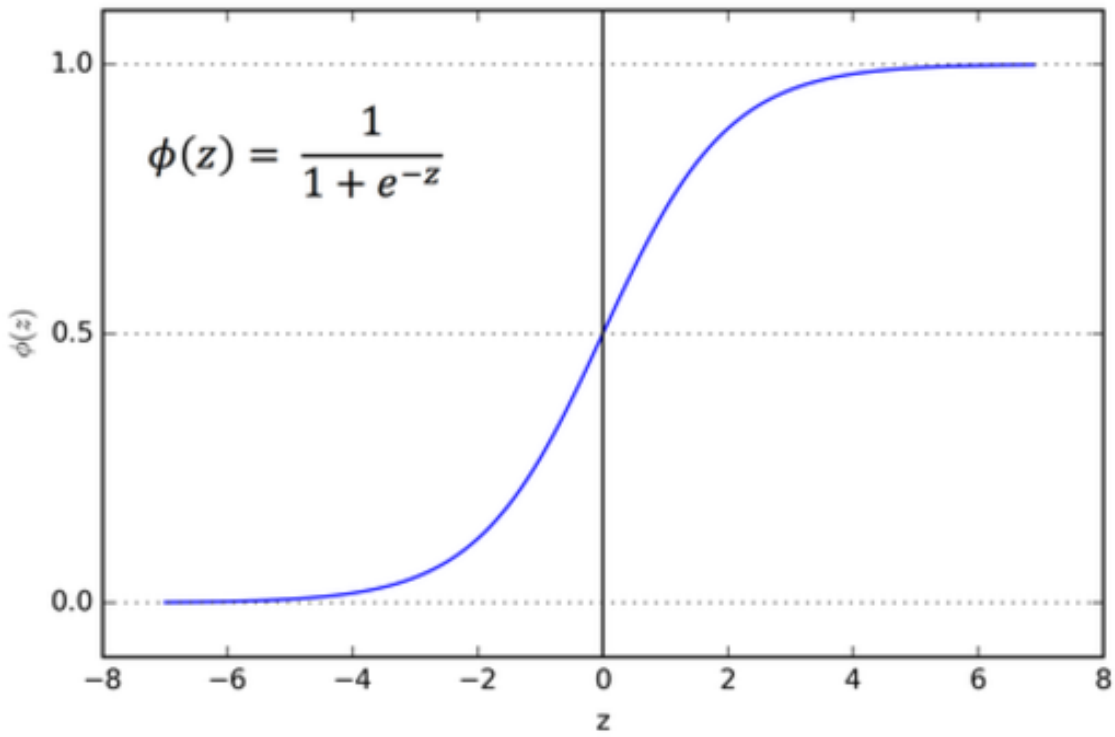


Figure 2.5: Sigmoidal function used as a prediction function in the logistic regression. Source: https://ai-master.gitbooks.io/logistic-regression/content/sigmoid-function.html

Each feature of data input is multiplied by its corresponding weight to form the following:

$$z = X^T\theta \tag{2.6}$$

Which is introduced to the sigmoidal hypothesis function:

$$h_\theta(X^T\theta) = \frac{1}{1 + e^{-X^T\theta}} \tag{2.7}$$

This produces an output from the range $< 0, 1 >$. During the training, a derivative of a binary cross-entropy cost function with respect to the weights is calculated. The solution of this operation is used for a subtraction from current weights causing their update which is described by:

$$\theta_j = \theta_j - \alpha \frac{\partial \mathcal{L}}{\partial \theta_j} \tag{2.8}$$

where α is a learning rate and the expression next to it is a partial derivative of the cost function with respect to a particular weight.

On the other hand, Logistic Regression is weak when data is highly dimensional. Fortunately, feature engineering in form of feature extraction and selection could minimise this issue. Also, logistic regression due to its linearity assumes a linear relationship between an input and an output. Given all of these advantages and disadvantages and research topics, logistic regression was claimed to be an interesting choice for this research.

Neural Networks

Neural networks are algorithms that at their origin were inspired by human neurons. A network consists of nodes to which a weighted input arrives and with an activation function, an output is produced and forwarded further within a network. Similarly, as for logistic regression, each feature is multiplied by an appropriate weight which sum of all is then used to determine an output. Depending on an activation function, the thresholds and range of values for an output differ. However, the main phenomenon behind perceptron and other neural-based models is a back-propagation algorithm that utilises an error of the output to update model weights hence improving a model.

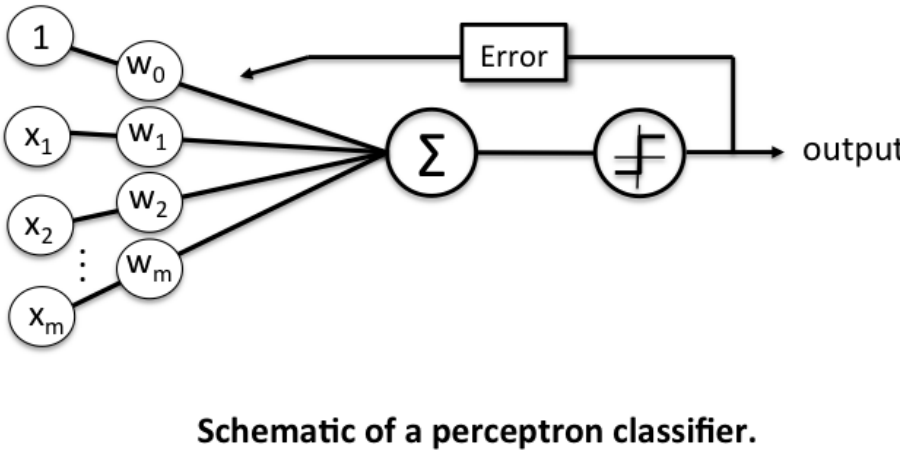


Figure 2.6: Perceptron - the first neural based machine learning classifier. Source: http://fiascodata.blogspot.com/2018/05/a-computer-program-is-said-tolearn-from.html

A huge disadvantage of the perceptron model is the incapability of solving XOR problems. This truth is because perceptron is a linear model and XOR classes are not linearly separable. To address this issue, artificial neural networks (ANN) were developed. The presence of the hidden layer in the middle in Figure 2.7 allowed neural networks to solve the XOR problems.

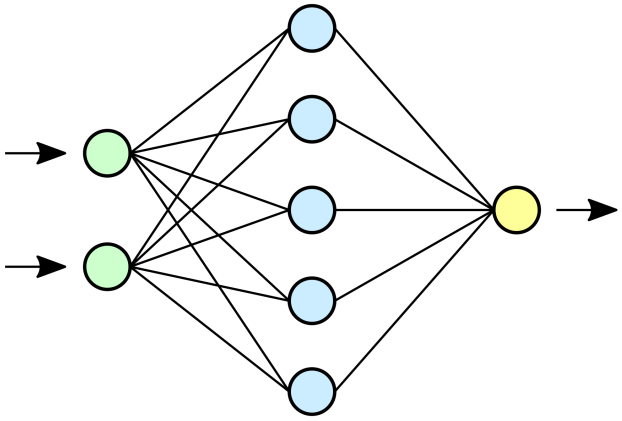


Figure 2.7: An example of Artificial Neural Network. Source: https://commons.wikimedia.org/wiki/Artificial_neural_network

The aforementioned back-propagation algorithm calculates the derivative of a loss function concerning input weights which is used to update them. For this complex ANN, it is possible due to the chain rule. On the other hand, there is a significant loss in the transparency of the model which is the reason why neural networks are regarded as black-box models.

Nowadays deep learning architectures are found more interesting to the research community. These refer to ANNs with multiple layers with numerous nodes and the term was coined by R.Dechter in [20].

2.2.4 Data Augmentation

Data augmentation stands for techniques of an upsampling dataset by artificially modified copies of its data samples. These techniques are meant to reduce an overfit of machine learning models and diversify a dataset without collecting new data. There are various types of strategies such as affine and colour-based transformations and the generation of new data through image transformation by generative adversarial networks (GANs). The term has been developed as the choice to train on similar but different examples to training data [21].

2.3 Deep Learning - Convolutional Neural Networks

Convolutional Neural Networks (CNNs) have become a standard model for image-related problems. For the first time, CNN has been proposed in [22] for the classification of handwritten digits. Due to the high accuracy in image classification, CNNs have become a common research subject which in consequence improved knowledge in machine learning and computer vision. One of the fields applying them is medical imaging which is used to solve problems related to classification, segmentation or registration [23], [24], [25]. This methodology has become so popular in the field that it has to be discussed and investigated in this research. CNN's consist of three main types of layers described below.

2.3.1 Convolution

Convolution in refers to the mathematical operation denotes with asterisk in the following way:

$$f[n] * g[n] = \sum_{k=-\infty}^{+\infty} f[k]g[n-k] \tag{2.9}$$

where f is the input and g is a filter which is also known as a kernel. The convolution layer takes an image patch as input and convolves it with a kernel producing an entry to a new feature map. Then, the image patch moves accordingly to the chosen value of steps also called strides. The process repeats until a new complete feature map is achieved which is going to be forwarded to the next layer.

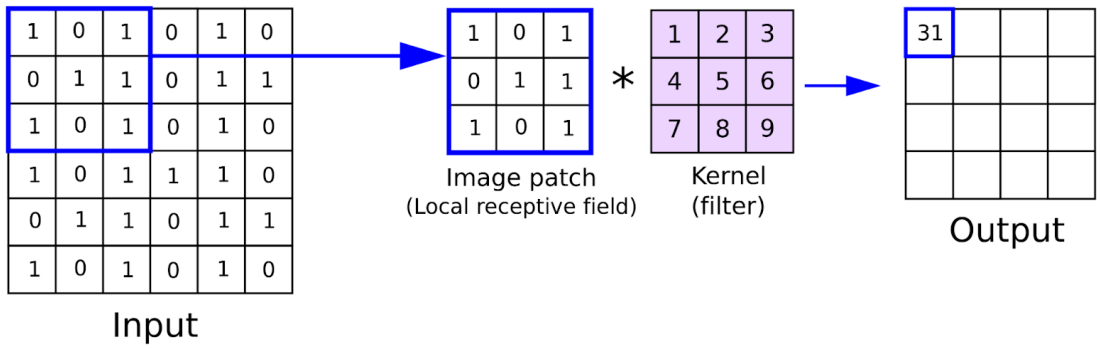


Figure 2.8: Convolution of an image input. Source: https://analyticsindiamag.com/what-is-a-convolutional-layer/

The choice of stride and kernel size is dependent on the data used so it is difficult to find their most optimal settings. It is important to keep in mind that by increasing the stride, the size of the feature map is going to be reduced.

2.3.2 Pooling

Pooling refers to a process of size reduction of feature maps. This is used to minimise the number of parameters to learn and computation time. Also, pooling produces values that are meant to represent the most important features. The most often types are averaging and max-pooling layers.

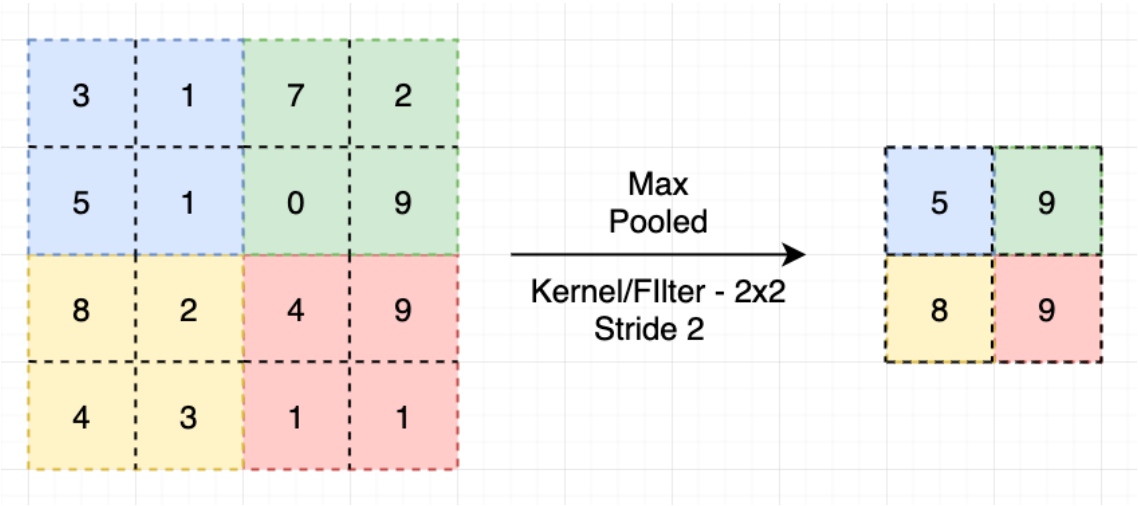


Figure 2.9: MaxPooling - a layer which takes a maximal value in a kernel placed on a feature map. Source: https://ai.plainenglish.io/pooling-layer-beginner-to-intermediate-fa0dbdce80eb

2.3.3 Fully Connected Layer

A fully connected layer essentially works similarly to a feed-forward neural net. The outputs of the previous layer are multiplied by weighting which produces a value that could be forwarded to the next layer or used to compute a final output when it is connected to some activation function such as softmax or sigmoid.

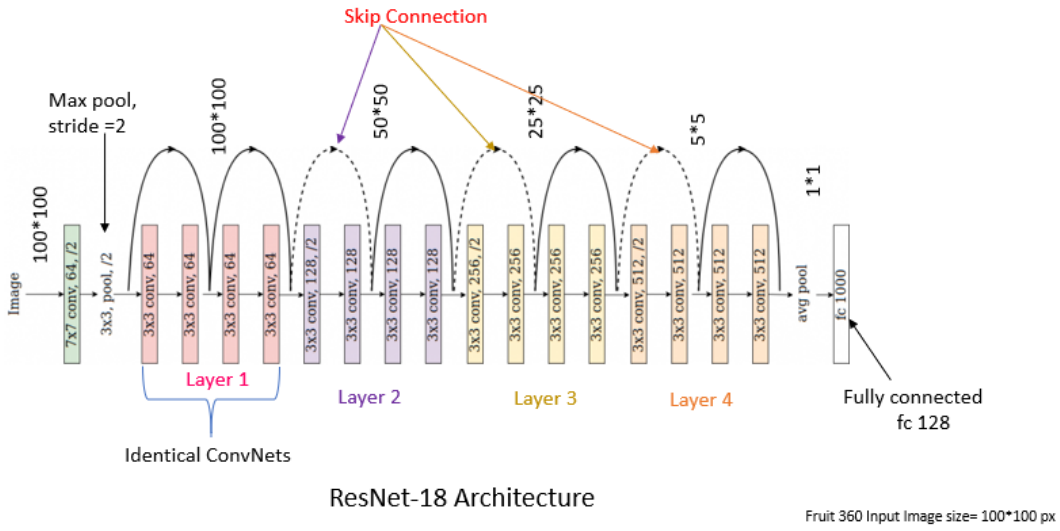


Figure 2.10: Examplary CNN architecture - ResNet-18. Source: https://www.pluralsight.com/guides/introduction-to-resnet

2.3.4 ResNet-18

The deep learning architecture from Figure 2.10 is an effective classification model of a residual neural network. The main advantage and characteristics of this architecture are residual blocks in which an input at the beginning is added to the output at their end. This specific implementation revolutionised the deep learning field because Kaiming He et al. introduced skip connections that allowed to train much deeper networks [26]. ResNet-18 is specifically considered as it was defined

as the smallest network hence that has the smallest number of parameters to be trained. In this study, the ResNet-18 will be implemented with a softmax and sparse categorical cross-entropy loss function.

2.3.5 Limitations

Although deep learning architectures have extraordinary accuracy, they are poorly calibrated. Modern architectures compared to older architectures are even more uncalibrated [9]. As a result, without an appropriate approach to improve their confidence estimation, probabilistic predictions of DNNs should not be trusted. However, deep learning models due to their structures enable techniques for confidence estimation development such as dropout that cannot be used for classical machine learning algorithms [7]. Another important limitation is the size of the dataset presented in the next chapter on which a DNN would be prone to overfit.

CHAPTER 3

Requirements and Analysis

3.1 Requirements

This research's aim and requirement is to improve the confidence estimation of MPCA pipelines available in the Pykale library [27], an open-source dedicated to interdisciplinary research. Due to the previous research conducted, SVC and logistic regression within the pipeline will be utilised. For convenience, every time one of these algorithms is mentioned, the entire pipeline is referred. MPCA consists of a classifier, feature extraction and selection methods introduced in the Literature Review.

Including the deep learning architecture in experiments and selective classification use cases are personal additions to the project owing to the literature review. The following objectives have been designed to extensively investigate available methods from which predictive models could benefit.

3.1.1 Calibrating confidence using Platt scaling

As discussed in the Literature Review chapter, Platt scaling is the most suitable method for calibrating models given this specific dataset and classifiers. To show the effect of calibration on the reliability of the algorithms, the models with Platt scaling will be compared to models without any calibration technique applied. This method is believed to improve confidence estimation of the algorithms because it was shown to be effective in the reviewed studies. The application of this method will be evaluated on how close predictions will imitate a perfect calibration line and calibration metrics such as expected calibration error (ECE) and negative log-loss.

3.1.2 Selective Classification with the SGR algorithm

The first approach is to compute a selective classifier by employing the SGR algorithm developed by Yonatan Geifman and Ran El-Yaniv in [8]. The algorithm searches for an optimal selection threshold based on confidence measures provided to a confidence function. When the threshold is obtained, a rejection function can be applied to improve the quality of predictions hence confidence. Although this technique was used for deep learning, it is believed that classical

machine learning models could benefit from it as well.

It is hypothesised that an implementation of a calibration technique with the selective approach with SGR will perform better than a separate calibrated or selective model. This is because calibration should make probabilities more appropriate hence the SGR would search for a more suitable selective threshold. Therefore, a rejection function would be more accurate in rejecting bad predictions yielding a better calibrated model. Thus, the SGR method will be combined with the Platt scaling as one of the experiments. The evaluation will be based on a comparison of the base, calibrated, selective (without calibration) and fully calibrated (selective and calibrated) models. In addition to metrics used for the calibration experiments, classification metrics such as accuracy and F1-score will be introduced as these could be affected significantly by the selective approach.

3.1.3 Data Augmentation for the confidence estimation

Predicting good probabilities relies on the training process of models. Thus, it is hypothesised that certain data augmentation techniques should have an impact on the calibration of models. Additionally, this method could help the models cope with a potential data imbalancement problem.

For experiments to test this hypothesis, three types of simple data augmentation are considered: affine, coloured and a combination of these two. Base and augmented models are going to be compared based on the confidence metrics and classification metrics such as accuracy and F1-score.

3.2 Analysis

3.2.1 Dataset

The dataset for this study was provided by The University of Sheffield for pulmonary arterial hypertension (PAH). It consists of 3850 cardiac MRI images of 179 patients. For each, 20 frames per cardiac cycle of a resolution 512x512 were taken and then resized to 64x64 [12].

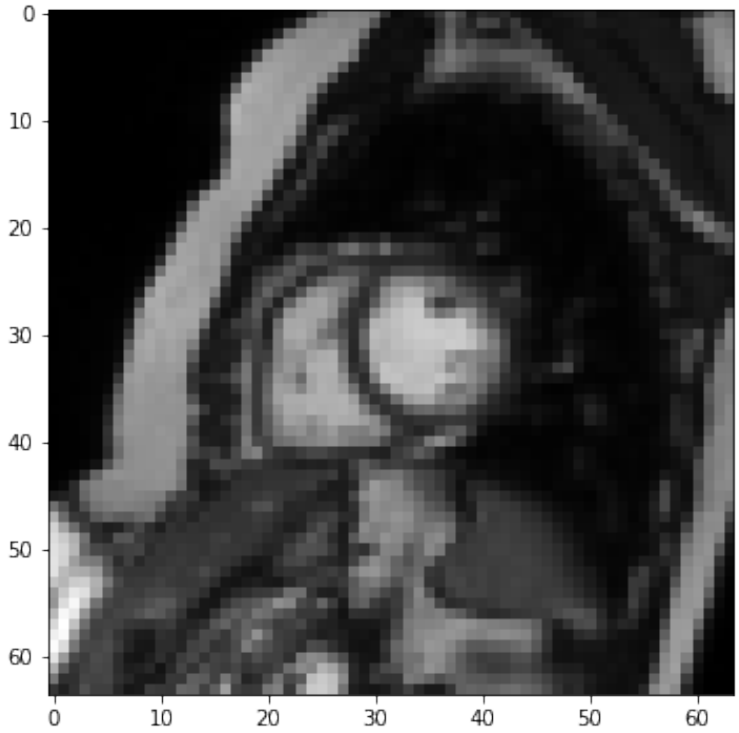


Figure 3.1: An example of a cardiac MRI scan from the PAH dataset.

Originally, the dataset has three classes, "0" denotes health control (no PH), "1" denotes idiopathic PAH (IPAH) and "2" labels PAH.

A major part of the results for the experiments described above is going to be represented as floats in an appropriate table as a mean value with a sum or subtraction of a standard error of all trials. These values will be used for statistical testing using the Student's t-test to investigate the significance of applied methods with 95% confidence. Thus, it is assumed that all of the results are i.i.d and follow a Gaussian distribution.

3.2.2 Limitations

Implementation is one of the biggest challenges for this research project. This is because it is planned to use TensorFlow for building the deep learning network. Additionally, the SGR algorithm will be implemented from the Github repository [8] that was created by Yonatan Geifman et al. However, there are numerous documentation and tutorials for TensorFlow and with supervision from Haiping Lu, I believe this problem can be overcome.

Gaps in knowledge is another issue that can have an impact on the rigour of this research as pointed out in my interim report. This is due to the fact of coming from a non-computer-science degree into the sophisticated subject in computation which is machine learning. As advised by Haiping Lu, I focused on a careful analysis of a good quality academic research and I attempted to study as much of the subject as I could.

Time needed for all of the planned experiments is significant. Given the complexity of the problems, it might be difficult to complete all of the objectives. The solution to that is to at first focus on the most crucial objectives of the research (i.e. calibration and selective prediction for MPCA pipeline classifiers) and then attempt to finish the optional goals like the implementation of CNN architecture and data augmentation methods.

Data size might influence the relevance of some experiments and models, especially the ResNet-18 application which is prone to overfit. However, with techniques discussed in the Methods an overcoming of this potential problem will be attempted.

CHAPTER 4

Methods

4.1 Introduction

As mentioned in Chapter 2, there are two main use cases for confidence estimation, calibration and selective classification. Both of them are going to be explored despite the fundamental differences between them. However, their aim is the same, to produce as many probabilistically reliable classifiers as possible.

The first method is about the application of Platt scaling to improve confidence estimation. According to the literature review and conclusions presented in Chapter 2, it is the best option for this particular choice of algorithms and data. Since the application of calibration is relatively easy if you have a huge amount of data, this method is optimal when data is scarce.

The second method is based on the SGR algorithm developed by Y. Geifman et al's work in [8] as mentioned before. It is a selective classification method that creates a selective classifier which abstains from predicting if the confidence of a prediction does not reach a selective threshold. In this method, it will be considered a combination of confidence calibration and SGR to improve further confidence estimation. To my best knowledge, this approach has not been done before hence this introduces a novelty to the field. A comparison with stand-alone methods is necessary to visualise the impact of this method.

The last method is a series of small experiments which as mentioned before cannot be categorised into any of the use cases above but still can impact the confidence of machine learning models. Common practice such as data augmentation will be discussed on a probabilistic basis and judged whether it is useful for a confidence improvement.

4.2 Method 1: Platt Scaling - Confidence Calibration

4.2.1 Motivation

There are parametric and non-parametric calibration techniques when used properly to improve the confidence estimation of our models [9]. The most recognisable parametric calibration method is called Platt scaling. This transformation has been experimentally proven to improve SVM reliability by Platt in [19]. However, there are two main issues with this solution. First is the potential introduction of unwanted bias by training a classifier and calibration technique on the same training sets. To cope with this, cross-validation or a separate validation set could be used. Secondly, researchers have shown that Platt scaling is most effective when confidence distortion has got a sigmoid shape which depends on the used classifier and dataset [28]. Thus, this technique is not necessarily effective every time. On the other hand, it is easy to implement and computationally cheap. The idea behind it is to learn a pair of parameters A and B that will minimise the cross-entropy error function:

$$J = min - \sum_i t_i log(p_i) + (1 - t_i) log(1 - p_i) \tag{4.1}$$

where

$$p_i = \frac{1}{1 + e^{Af(x_i)+B}} \tag{4.2}$$

where $f(x_i)$ is a prediction on a ith sample from a dataset (x_i, y_i) and y_i are labels. The t_i term denotes target probabilities which are produced from the original labels in the following way:

$$t_i = \frac{y_i + 1}{2} \tag{4.3}$$

The 4.2 is a sigmoidal function which is used to map the prediction of a model to a calibrated probability given the two optimisation parameters from 4.1.

Furthermore, researchers investigated different statistical ways to improve the reliability of predictive models. If one cannot identify the shape of the probability distortion, then non-parametric methods can be implemented. One of the simplest of them is histogram binning in which training examples are divided to their scores into B sets of the same size, bins. The problem with this method is that if a dataset is small or imbalanced, then an optimal number of bins is considered to be unlikely [29]. Zadrozny and Elkan have proposed another method, isotonic regression. This was developed to create a method that would calibrate any monotonic distortion. The way it works is in essence to produce an isotonic function h which will transform uncalibrated probabilities and minimize the square loss. The isotonic function itself is found by solving the following optimization problem:

$$\hat{h} = argmin \sum_{i=1}^{N} (y_i - f(x_i))^2 \tag{4.4}$$

However, it is more prone to overfitting if the calibration dataset is small than Platt scaling [28]. Both of the methods are suitable only for binary classification. The time complexity of Platt scaling and isotonic regression is almost similar [30]. Given that medical imaging datasets usually are small (including the one used for this research project) and the characteristics of the aforementioned methods, Platt scaling will be a wiser choice for calibration tasks.

4.2.2 Novel Contributions

Since this technique has been already applied in the industry, it is not providing any novelty to the field by itself. However, deep exploration and comprehension of this method and related experiments will be necessary to understand the second method. Additionally, there is little research done for the calibration of machine learning models in the medical image classification context which make a small contribution to the field.

4.2.3 Evaluation

The evaluation of confidence calibration is relatively simple and there is a consensus on the methods used. One of the most used is expected calibration error (ECE) denoted by the expression:

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{N} |acc(B_m) - conf(B_m)| \tag{4.5}$$

This metric calculates the absolute value of the difference between accuracy and confidence for a certain confidence bin. As mentioned in Chapter 2, the aim is to make our model as reliable as possible. If a model had the confidence of 20%, it would be expected to receive 20% of accurate predictions. Therefore, this metric should be minimised. Additionally, a negative log loss metric will be used to judge a calibration of a classifier. This is calculated by:

$$LogLoss = -\frac{1}{N} \sum_{i=1}^{N} y_i log(p(y_i)) + (1 - y_i) log(1 - p(y_i)) \tag{4.6}$$

This metric originates from the entropy formula. In essence, the lower the log loss value is the better a classifier becomes.

To include a visual evaluation of the calibration, reliability diagrams will be introduced. These graphs will provide information about at what exact confidence bin there is good or bad reliability of a given predictive model. Figure 4.1 illustrates an example of a reliability diagram. The dotted line denotes a perfect calibration curve which should be aimed to achieve when calibrating a model. Red bars illustrate a gap between perfect and actual accuracy. The bigger the gap, the further an accuracy per bin is from the accuracy for perfect calibration. The black bold horizontal line on the top or bottom of the bars denotes an expected accuracy. If the line is on a bottom of a bar, then the model is over-confident in this bin. Otherwise, the model is under-confident. At a bottom of a reliability diagram, is a jointed histogram of predictions per a confidence bin. The bold vertical line describes an accuracy of a model and the dotted line gives information about average confidence across all of the examples.

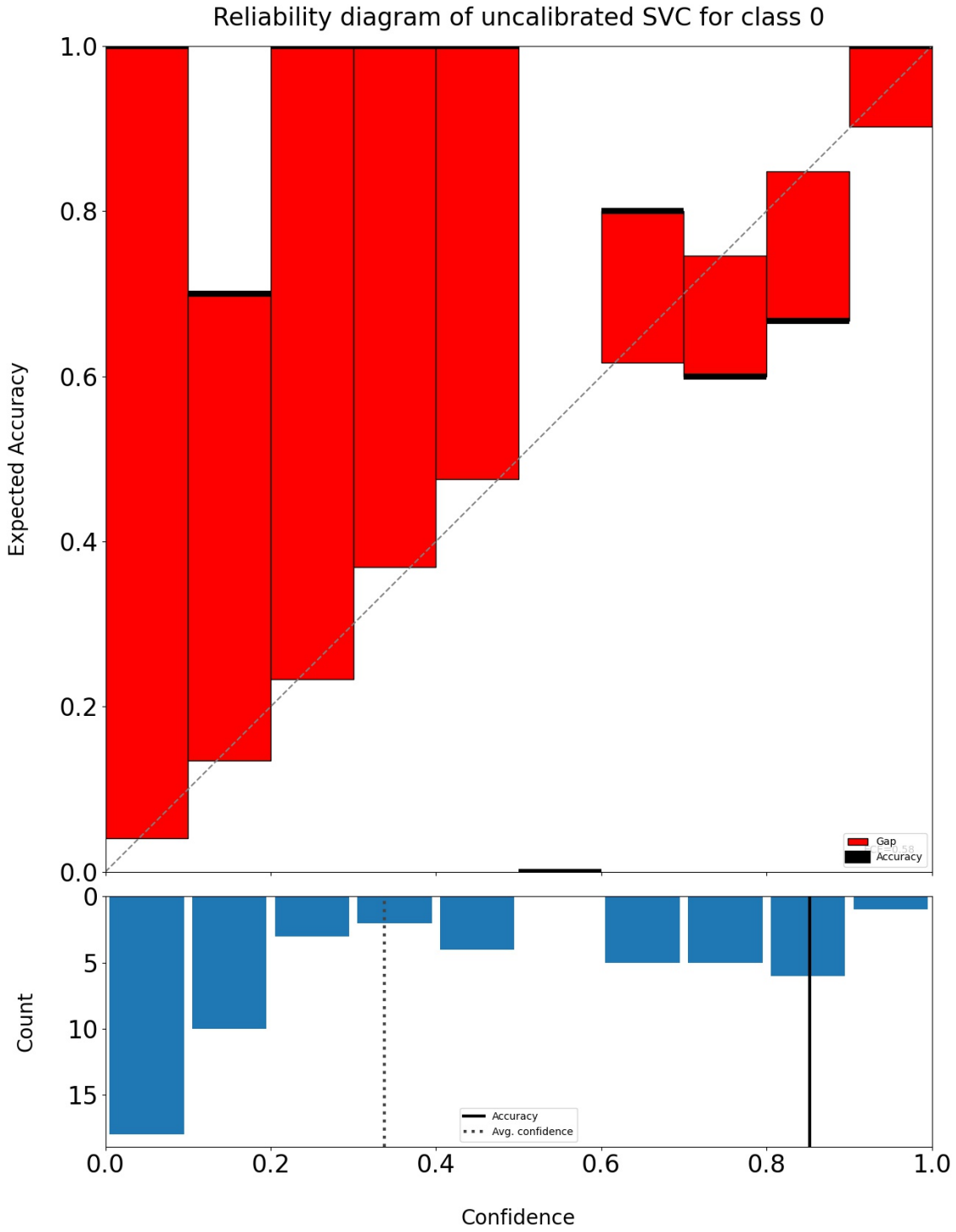


Figure 4.1: An example of a reliability diagram. Source: https://github.com/hollance/reliability-diagrams - code used and modified for the purpose of the dissertation.

4.3 Method 2: Selection with Guaranteed Risk (SGR) supported by calibration

4.3.1 Motivation

Geifman et al. in [8] have proposed an interesting method for the selective classification use case. The approach is based on the SGR algorithm that performs a binary search operation in

Algorithm 1. The main objective of it is to find the value of a selective threshold θ that will be used to create a rejection function g that is defined in the following:

$$g(x) = \begin{cases} 1, & \text{if } k_f(x) \geq \theta. \\ 0, & \text{otherwise} \end{cases} \tag{4.7}$$

where k_f is a confidence function. The choice of the function is not an obvious task as many different ones could be picked. The rule of thumb is that it should reflect the true loss of monotonicity. Meaning, that given a dataset (x_i, y_i), a confidence function should be sorted before starting the algorithm. For deep neural networks, softmax response is used by practitioners but this research also considers probabilistic outputs of the classical machine algorithms. Nevertheless, a confidence function k_f is described by the following:

$$k_f = max(f(x|m)) \tag{4.8}$$

Where x is an image input. This function chooses the highest value of confidence from each probabilistic prediction.

Algorithm 1 Selection with Guaranteed Risk

Input: k$_f$,δ, r^*, S_m ▷ Updated algorithm from [8]

$i_{min} = 1$
$i_{max} = m$
for $j = 1$ to $\lceil log_2(m) \rceil$ **do**
    $i = \frac{i_{min}+i_{max}}{2}$
    $\theta = k_f(x_i)$
    $\hat{r_j} = \hat{r}(f, g_j|S_m)$
    $b^* = B^*(\hat{r_j}, \frac{\delta}{\lceil log_2(m) \rceil}, g_i(S_m))$
    if $b^* < r^*$ **then**
        $i_{max} = i$
    else
        $i_{min} = i$
    end if
end for
return θ, b^*

It is questionable whether the softmax layer represents actual probabilities and this is going to be addressed later in the Discussion chapter. However, the confidence function k_f task is to rank outputs. Regarding linear machine learning classifiers, actual probabilities can be used for that. Selective classifier for a given set S_m has to be designed in a way that satisfies the following:

$$Pr(R(f, g) > r^*) < \delta \tag{4.9}$$

Hence, we want that the probability of the selective risk being higher than the target risk to be lower than the chosen confidence parameter δ. The first parameter is defined as an expected value of loss of model f multiplied by a rejection function divided by the expected value of the rejection function.

$$R(f, g) = \frac{E_p[\mathcal{L}(\hat{y}, y)g(x)]}{E_p[g(x)]} \tag{4.10}$$

Nota bene, it is assumed that the distribution P is unknown. Although it would be difficult to compute this expression since there is assumed no information about the distribution. Thus, by fixing P, an empirical selective risk is defined as:

$$\hat{r} = \frac{\frac{1}{m}\sum_{i=1}^{m} \mathcal{L}(f(x_i), y_i) g(x_i)}{\frac{1}{m}\sum_{i=1}^{m} g(x_i)} \tag{4.11}$$

For these given parameters, the SGR algorithm finds a bound risk b^* which is going to guarantee the required target risk with a given confidence. The authors claim that a selective classifier is going to be returned from the algorithm. However, technically the only useful output produced is a threshold value θ which allows one to try different selective functions to form a selective classifier. In this research, the simple function introduced defined in Equation 4.7 is going to be considered. The risk bound is optimally computed from the expression B^*:

$$B^* = \sum_{j=0}^{m\hat{r}(f|Sm)} \binom{m}{j} b^j (1-b)^{m-j} \tag{4.12}$$

The result from 4.12 is then compared to the value of the risk target. As results in [8] showed, the algorithm produces a risk bound very close to the target value. Therefore as claimed in another paper, this method is capable of producing probabilistically-calibrated selective classifiers [31].

4.3.2 Novel Contribution

As mentioned in Method 1, Platt scaling is going to be substantial for this method as well. The objective of it is to combine calibration and selective classification to obtain significantly improved confidence estimation of a machine learning model by a presumably more accurate choice of an appropriate selective threshold. To my knowledge, this implementation has never been used before and could make classifiers more reliable, not only in the medical imaging problem cases.

4.3.3 Evaluation

Similarly to Method 1, reliability diagrams and ECE can be used owing that selective classification also tries to make the classifier calibrated better. These metrics will be particularly useful because they will be necessary for comparison to judge the efficacy of the proposed implementation. Additionally, F1-score, precision and recall will be used to track the method's influence on a performance against a potential imbalacement. Accuracy will also be taken into account to visualise a general performance.

4.4 Method 3: Data augmentation with respect to confidence estimation

4.4.1 Motivation

The CMRI dataset is an example of a medical imaging dataset which are prone to imbalancement regarding the number of samples per class. In this case, the count of examples per class seems to be a bit imbalanced.

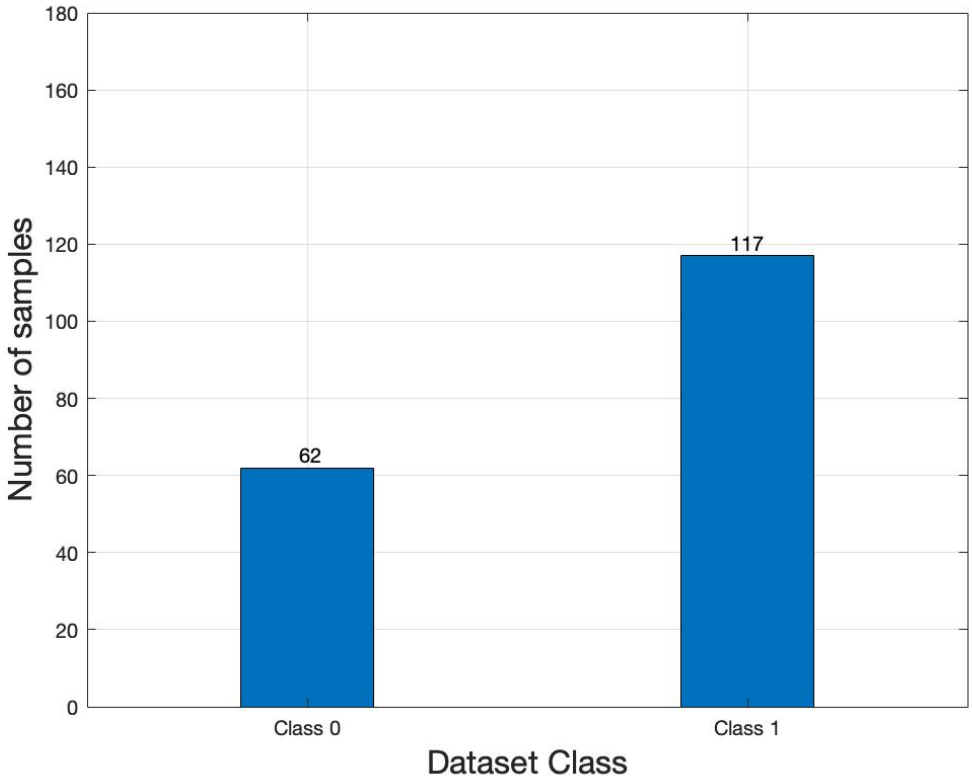


Figure 4.2: Class count for the CMRI dataset.

Data augmentation which is used in situations with a risk of an imbalanced dataset can impact a confidence estimation of a predictive model. If the amount and quality of data influence the quality of predictions, then the addition of artificial data to a training set is going to affect a model's confidence estimation. In [32] Yuval Bahat et al. have discussed the impact of image transformation on confidence estimation of popular deep learning architectures. The most frequently used list of image transformations consists of horizontal and vertical flips, zooming, brightness manipulation and rotation. More complex data augmentation techniques which potentially have a higher impact on the model performance are for instance gamma correction. In this part of this research, a series of experiments with the usage of different data augmentation methods are going to be investigated in terms of confidence estimation and its improvement. Affine transformations will consist of horizontal, vertical flips and a rotation. Colour transformations in this study refer to a gamma correction, hue, contrast and brightness. Lastly, a combination of the two above will be considered a separate type of transformation.

4.4.2 Novel Contribution

There are only a few papers that discussed data augmentation of medical image datasets with simultaneous attention paid to the model's confidence estimation capability. As claimed by Yuval Bahat et al., image transformations can be dependent on a particular dataset. Therefore, the results obtained with this method are going to be particularly useful for future research in the development of machine learning methodology for cardiac MRI datasets. This study demonstrates data augmentation as another technique for improving confidence estimation.

4.4.3 Evaluation

The evaluation of the performance of particular transformations on confidence estimation will be tracked by ECE and negative log loss. Additionally, precision, recall and F1-score will be used to track the performance against a potential imbalancement problem. Accuracy will be tracked for the general performance of models.

CHAPTER 5

Results

For all of the experiments, the data split was 70/30 for classical machine learning models with cross-validation for training and test set respectfully. Regarding the ResNet-18, the data split was 70/15/15 for the training, validation and testing sets respectfully.

5.1 Confidence Calibration - Experiments

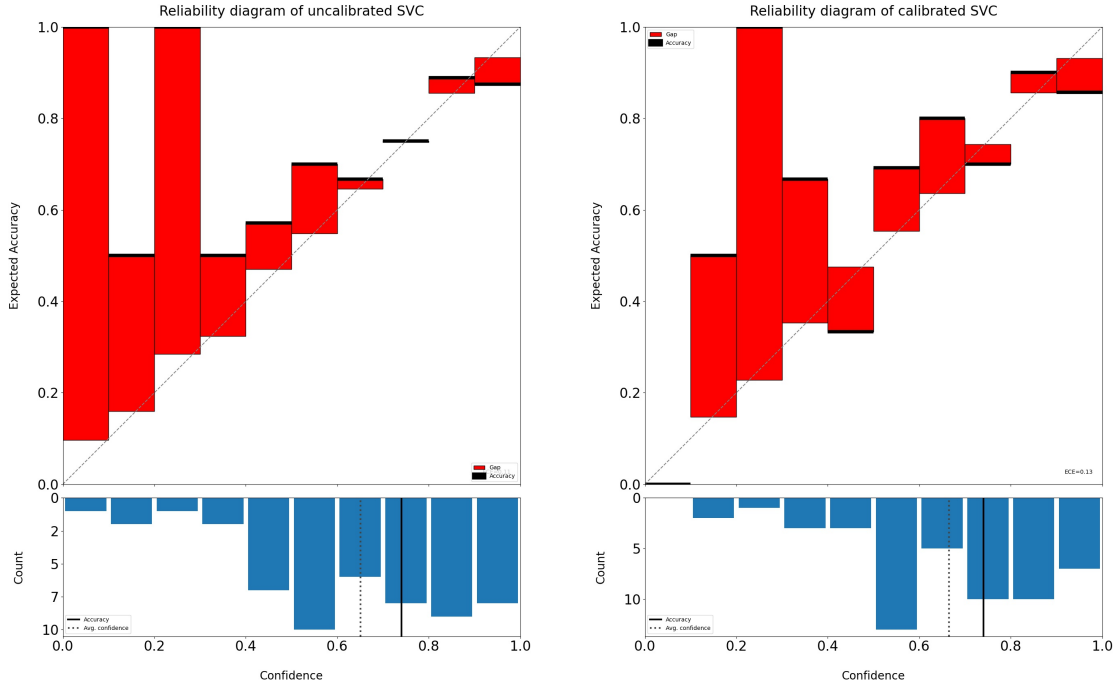
To calibrate the algorithms, a calibration with 10-fold cross-validation was implemented on both the logistic regression and the SVC with a *CalibratedClassifierCV* function from the sklearn library. A separate validation set was created only to calibrate the ResNet-18 which was necessary as DNN models cannot be introduced to the aforementioned function hence Platt scaling had to be developed originally. In addition, the ResNet-18 was trained on 15 epochs with a learning rate of 0.01 with a softmax layer applied as following the approach in [8]. Experiments consisted of 10 trials for each model for which a mean and a standard error were calculated. The results of confidence calibration have been presented in Table 5.1.

SVC	**Uncalibrated**	**Calibrated**
Expected Calibration Error	0.3422 ± 0.0104	**0.3348 ± 0.0098**
Log Loss	1.1066 ± 0.0417	**1.0110 ± 0.0133**
Logistic Regression		
Expected Calibration Error	0.3741 ± 0.0152	0.3741 ± 0.0152
Log Loss	1.3788 ± 0.1513	1.3788 ± 0.1513
ResNet-18		
Expected Calibration Error	0.4354 ± 0.0115	**0.2183 ± 0.0274**
Log Loss	1.7323 ± 0.0857	**0.7170 ± 0.0302**

Table 5.1: Platt scaling calibration results for SVC, logistic regression and ResNet-18 architecture. The lower the value of both metrics, the more calibrated a model becomes.

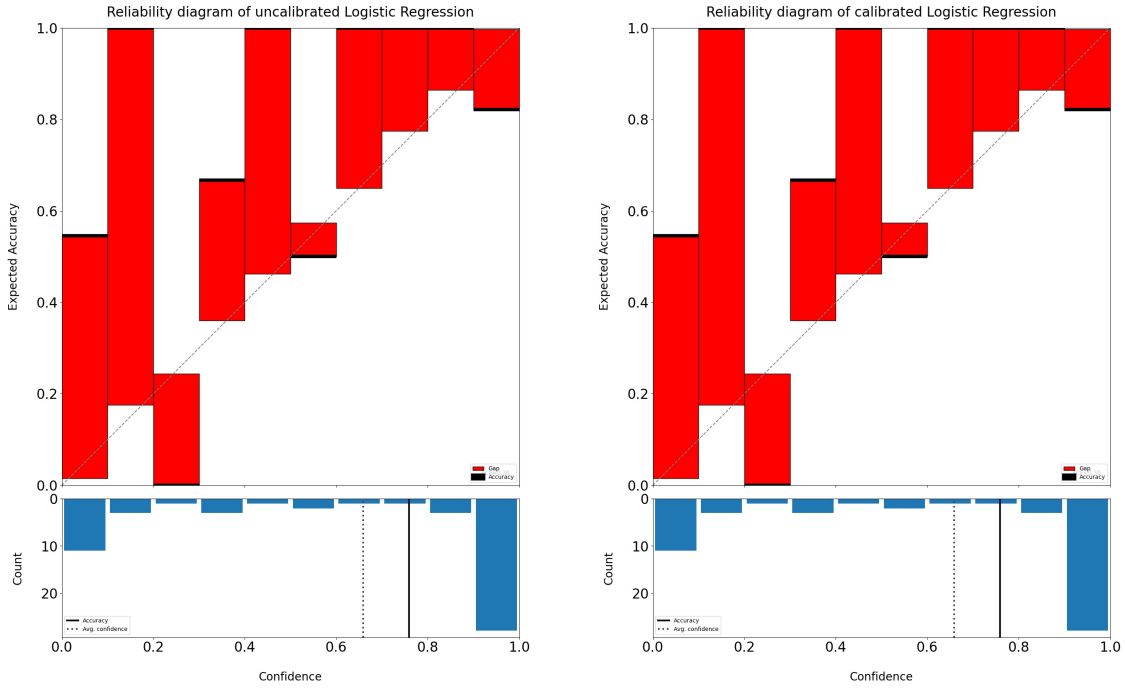
For the pipeline with SVC, a slight improvement in expected calibration error and negative log loss has been noted. According to the results, log loss value changed with a statistical significance whereas ECE did not. As for the logistic regression pipeline, no change in the metrics was observed

which was expected due to that the algorithm often produces well-calibrated probabilities. On the other hand, calibration of the ResNet-18 yielded a significant improvement in both metrics.



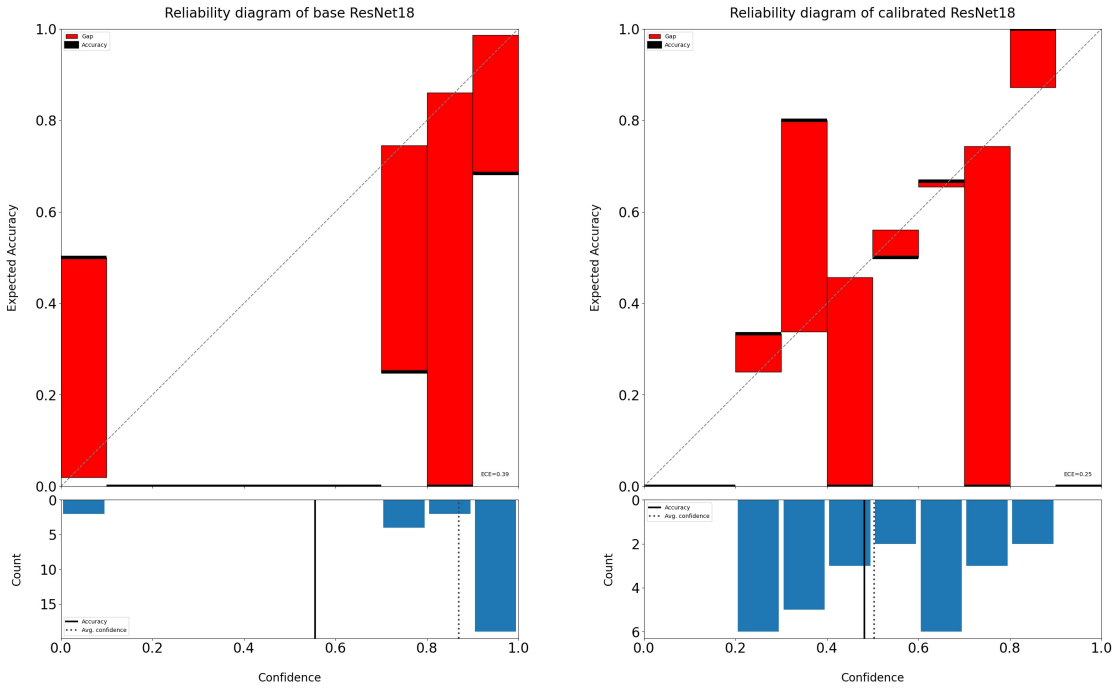
(a) Uncalibrated (b) Calibrated

Figure 5.1: Reliability diagrams of SVC pipeline before and after Platt scaling.



(a) Uncalibrated (b) Calibrated

Figure 5.2: Reliability diagrams of logistic regression pipeline before and after Platt scaling.



(a) Uncalibrated (b) Calibrated

Figure 5.3: Reliability diagrams of ResNet-18 before and after Platt scaling.

According to reliability diagrams visualised in Figure 5.1, Platt scaling has changed confidences for a couple of examples which is showed by a change of counts per bins in histograms. For this particular example, the error gaps have slightly increased for a few bins. Regarding the Figures 5.2, there was no change in the number of examples or accuracy per confidence. Thus, Platt scaling did not affect logistic regression overall. Contrary to this, calibration of ResNet-18 caused a significant change in error gaps and number of predictions per confidence bin. Before calibration, most of the examples were assigned to the most confident bin. Moreover, black line for the high confidence bins denote that the model was over-confident for these predictions. After Platt scaling, the predictions were scattered across most of the confidence bins. Additionally, unlike the uncalibrated model, for most bins the calibrated ResNet-18 was under-confident in its predictions but with mostly lower error gap. Furthermore, for the 50% confidence bin where accuracy of the model is placed there is an error gap much larger than expected.

5.2 Selective Classification - Experiments

5.2.1 SGR algorithm implementation

For all trials in this set of experiments, hyperparameters for the SGR algorithm which are confidence parameter δ and risk target were set to 0.01 and 0.2 respectively. For each type of classifier three models were considered. Base model to which no calibration or selective approaches were applied. Base selective models were modified with the SGR algorithm and for calibrated selective models additionally Platt scaling was used. Additionally, for a comparison of a calibration performance, an additional model with only Platt scaling was added to show the true impact of the proposed methods. For calibrating particular models, the approach is the same as in the previous experiments. Regarding the SGR implementation, a training set was used to derive a selective threshold. In Table 5.2 the results of 10 trials per model are presented as means and standard errors per each setup. Paramount part of the implementation of the calibrated selective models is

that the Platt scaling was applied before the SGR algorithm.

SVC	**Base**	**Calibrated**	**Base + Selective**	**Calibrated + Selective**
ECE	0.3422 ± 0.0104	**0.3348 ± 0.0098**	0.3830 ± 0.0122	0.3805 ± 0.0156
Log Loss	1.1066 ± 0.0417	**1.0110 ± 0.0133**	1.1455 ± 0.0370	1.4177 ± 0.0470
Accuracy	0.7741 ± 0.0101	0.7741 ± 0.0101	**0.8132 ± 0.0121**	0.8121 ± 0.0125
Logistic Regression				
ECE	**0.3741 ± 0.0152**	**0.3741 ± 0.0152**	0.3943 ± 0.0144	0.3943 ± 0.0144
Log Loss	**1.3788 ± 0.1513**	**1.3788 ± 0.1513**	1.5556 ± 0.2886	1.5556 ± 0.2886
Accuracy	0.7648 ± 0.0155	0.7648 ± 0.0155	**0.772 ± 0.0164**	**0.772 ± 0.0164**
ResNet-18				
ECE	0.4354 ± 0.0115	**0.2183 ± 0.0274**	0.4369 ± 0.0096	0.2225 ± 0.0280
Log Loss	1.7323 ± 0.0857	**0.7170 ± 0.0302**	1.6335 ± 0.2371	0.7331 ± 0.0411
Accuracy	0.5741 ± 0.0198	0.5778 ± 0.0332	0.5692 ± 0.0171	**0.5913 ± 0.0376**

Table 5.2: Comparison of performance of base models, calibrated models with Platt scaling, base selective models with a selective function and calibrated selective models with both selective function and Platt scaling. **Bolded** results outline the best results for a particular classifier. ECE and log loss is aimed to be minimised whereas accuracy is desired to be as close to 1 as possible.

Logistic regression results of base selective and calibrated selective models are likewise. This observation is similar to the one presented in the confidence calibration experiment. Although ECE and log loss were worsened a slight improvement in accuracy was observed but without any statistical significance according to the Student's t-test. For the SVC, the base model noted the best ECE and log loss within the group of the models. However, slight differences between base selective and calibrated selective SVC are present. The base model with the selective function had the best accuracy score which is better compared to the base and calibrated models but insignificantly different from the calibrated selective model. According to the results for the ResNet-18 group of models, the model with calibration only had the best values for the calibration metrics. However, these results compared to the model with an additional selective function are not statistically significant. On the other hand, these models perform much better than models without calibration. In calibrated ResNet-18 there is a slight unexpected change in its accuracy score compared to the base model but the value is too small to be meaningful.

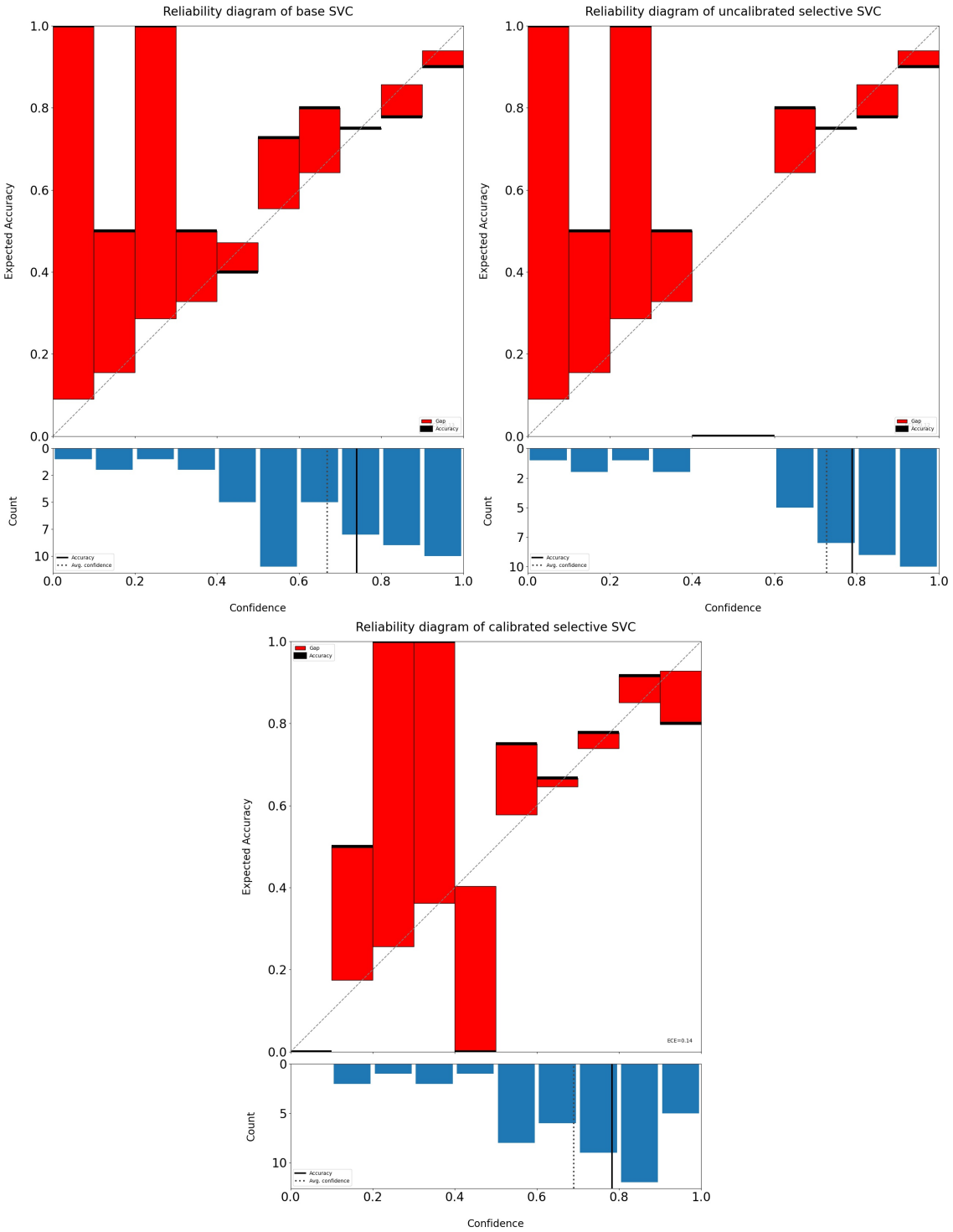


Figure 5.4: Reliability diagrams of SVC base, base selective and calibrated selective models.

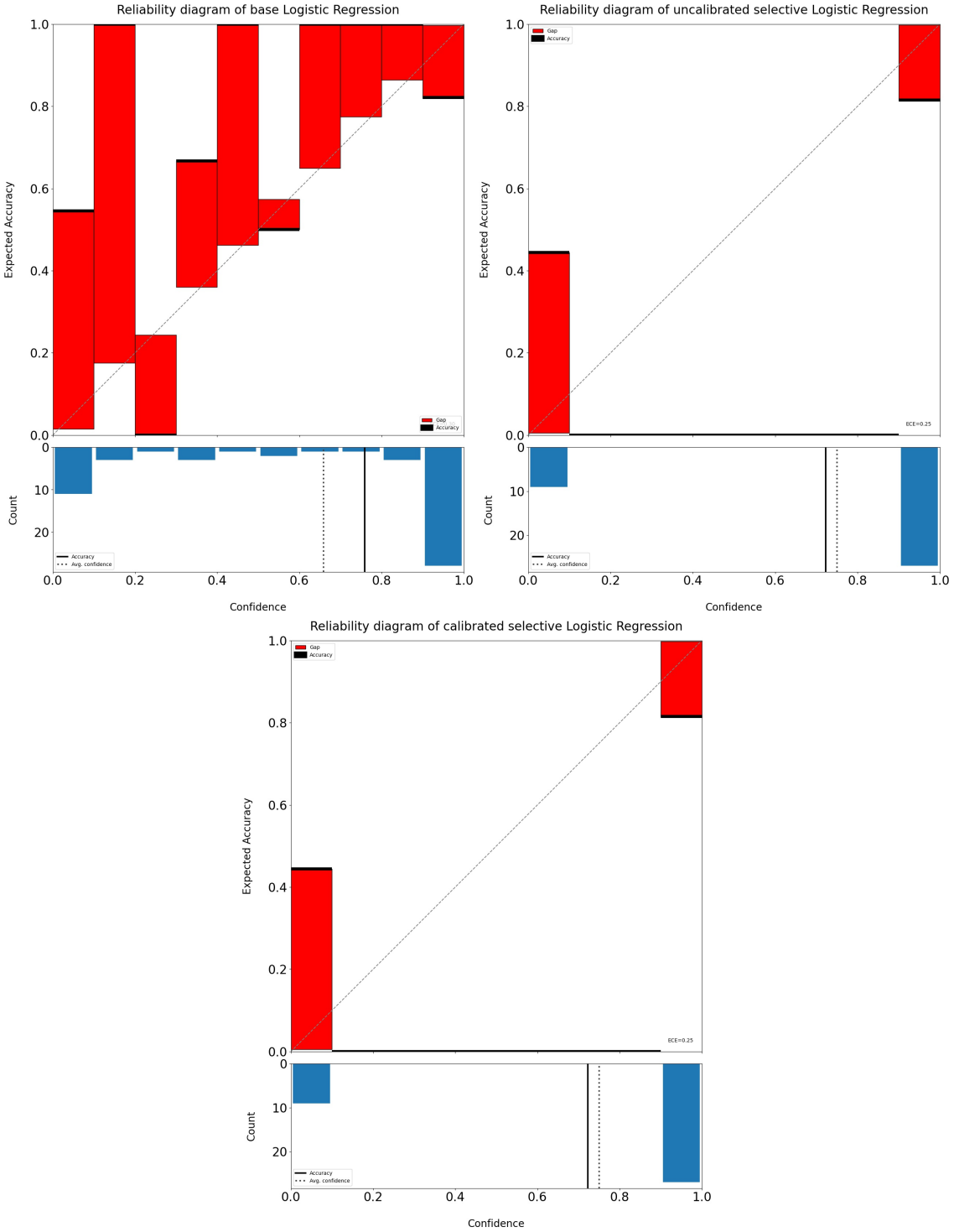


Figure 5.5: Reliability diagrams of logistic regression base, base selective and calibrated selective models.

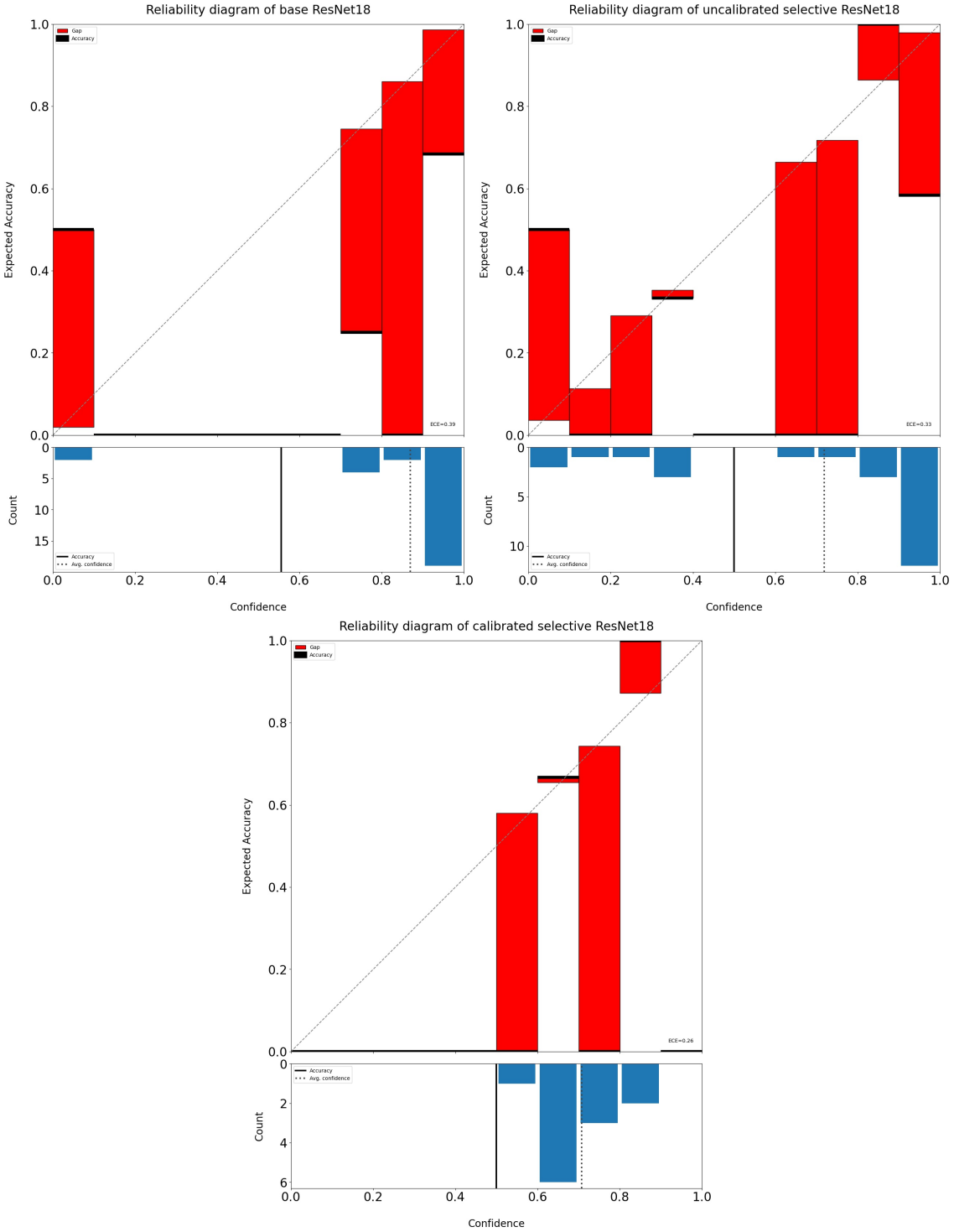


Figure 5.6: Reliability diagrams of ResNet-18 base, base selective and calibrated selective models.

The main consequence of using a selective function according to the Figures 5.4 - 5.6 is a drop of values for some confidence bins. If the error gap for a particular confidence bin was large and then was dropped then a model should benefit from this in terms of a calibration as false predictions were removed. Depending on the number of predictions with a certain confidence level, a different number of confidence bins could be dropped as when Figures 5.4 and 5.5 are compared. The reason for this is that SGR algorithm when using a binary search looks for an optimal value among given probabilistic values. Therefore, if a model had all of the predictions within one confidence bin, all of them would be dropped if a selective function was utilised.

However, due to the risk of the imbalancement, it is necessary to track the performance of the selective implementations with respect to F1-score, precision and recall. The following tables show the results of the methods in the discourse separately for class 0 and class 1 to observe and analyse the magnitude of imbalancement.

SVC	**Base / Calibrated**	**Base + Selective**	**Calibrated + Selective**
Precision	0.7289 ± 0.0284	0.7398 ± 0.0423	**0.7585 ± 0.0425**
Recall	**0.5951 ± 0.0439**	0.5814 ± 0.0827	0.5700 ± 0.0661
F1-score	**0.6478 ± 0.0341**	0.6189 ± 0.0631	0.6258 ± 0.0510
Logistic Regression			
Precision	0.6898 ± 0.0211	**0.7232 ± 0.0351**	**0.7232 ± 0.0351**
Recall	0.6094 ± 0.0247	**0.6236 ± 0.0229**	**0.6236 ± 0.0229**
F1-score	0.6444 ± 0.0194	**0.6645 ± 0.0192**	**0.6645 ± 0.0192**
ResNet-18			
Precision	0.3278 ± 0.0672	0.3621 ± 0.0667	**0.5563 ± 0.0778**
Recall	0.2099 ± 0.0747	**0.4533 ± 0.0737**	0.2804 ± 0.0418
F1-score	0.2559 ± 0.0576	**0.4026 ± 0.0690**	0.3729 ± 0.0431

Table 5.3: Comparison of performance of base models, calibrated models with Platt scaling, base selective models with a selective function and calibrated selective models with both selective function and Platt scaling against imbalancement for class 0. **Bolded** results outline the best results for a particular classifier. For all the metrics, the closer to 1, the better the model's performance.

SVC	**Base / Calibrated**	**Base + Selective**	**Calibrated + Selective**
Precision	0.7826 ± 0.0261	**0.8254 ± 0.0311**	0.8180 ± 0.0267
Recall	0.8688 ± 0.0155	0.9096 ± 0.0192	**0.9125 ± 0.0195**
F1-score	0.8214 ± 0.0183	**0.8619 ± 0.0199**	0.8594 ± 0.0171
Logistic Regression			
Precision	0.7802 ± 0.0245	**0.8159 ± 0.0178**	**0.8159 ± 0.0178**
Recall	0.8416 ± 0.0099	**0.8756 ± 0.0179**	**0.8756 ± 0.0179**
F1-score	0.8082 ± 0.0164	**0.8433 ± 0.0137**	**0.8433 ± 0.0137**
ResNet-18			
Precision	0.6332 ± 0.0341	**0.7025 ± 0.0205**	0.5908 ± 0.0242
Recall	0.6914 ± 0.0551	0.5531 ± 0.0564	**0.7258 ± 0.0646**
F1-score	**0.6610 ± 0.0366**	0.6189 ± 0.0391	0.6514 ± 0.0333

Table 5.4: Comparison of performance of base models, calibrated models with Platt scaling, base selective models with a selective function and calibrated selective models with both selective function and Platt scaling against imbalancement for class 1. **Bolded** results outline the best results for a particular classifier. For all the metrics, the closer to 1, the better the model's performance.

According to the results in Table 5.3 and 5.4, there are huge differences between metric values between classes. Base and calibrated models share the same results for the imbalancement metrics which was expected because Platt scaling does not affect label prediction. The introduced methods

in most cases improved precision, recall and F1 scores. For all classifiers, there is an improvement in precision compared to the base models. Logistic regression noted a significant improvement in the metrics for both classes. On the other hand, the SVC base model has performed the best regarding the recall and F1-score for class 0. However, class 1 methods with the SGR applied outperformed it. Regarding the ResNet-18, there is a huge development of the metrics in the selective models compared to the base and calibrated models. Although the calibrated selective model did significantly better in precision, the recall and F1-score values dumped compared to the base selective model but still, it has higher results than the base and calibrated for the class 0. According to Table 5.4, each metric has its highest value for separate models for class 1.

5.2.2 Influence of varying risk target

Computations for appropriate selective threshold θ depend on the probabilistic and hyperparameter values introduced to the SGR algorithm. Whenever one chooses a different δ or risk target value, a distinct θ will be obtained. For this specific dataset assuming δ to be constant, generally the smaller value of risk target yield a higher value of θ. To investigate the effect of this particular hyperparameter on the confidence estimation of models, 10 trials per model for a given risk value were conducted. Due to insignificant changes for logistic regression models against previous experiments, only SVC and ResNet-18 are considered in this part of the study. The specific range of risk target was chosen owing to that the most significant changes in θ hence in the metrics were noticed within it.

In all of the figures below, calibrated selective and base selective SVC seems to follow each other in all of the metrics for all tested values of the risk target. On the other hand, the difference between base selective and calibrated selective ResNet-18 models is substantial in that the latter always outperformed the first one in terms of the ECE and the negative log-loss. According to the accuracy Figure 5.9, this was a case only for lower risk target values. SVCs compared to ResNet-18 models have higher standard error for most measurements. The exception applies for Figure 5.8 where calibrated selective ResNet-18 has comparably low standard errors to SVCs curves for risk target values from 0.2 to 0.24. There is no significant difference between SVCs models for all of the metrics. However, there is a downtrend relationship between the risk target value and SVCs accuracy. Optimal values for each metric vary between the models.

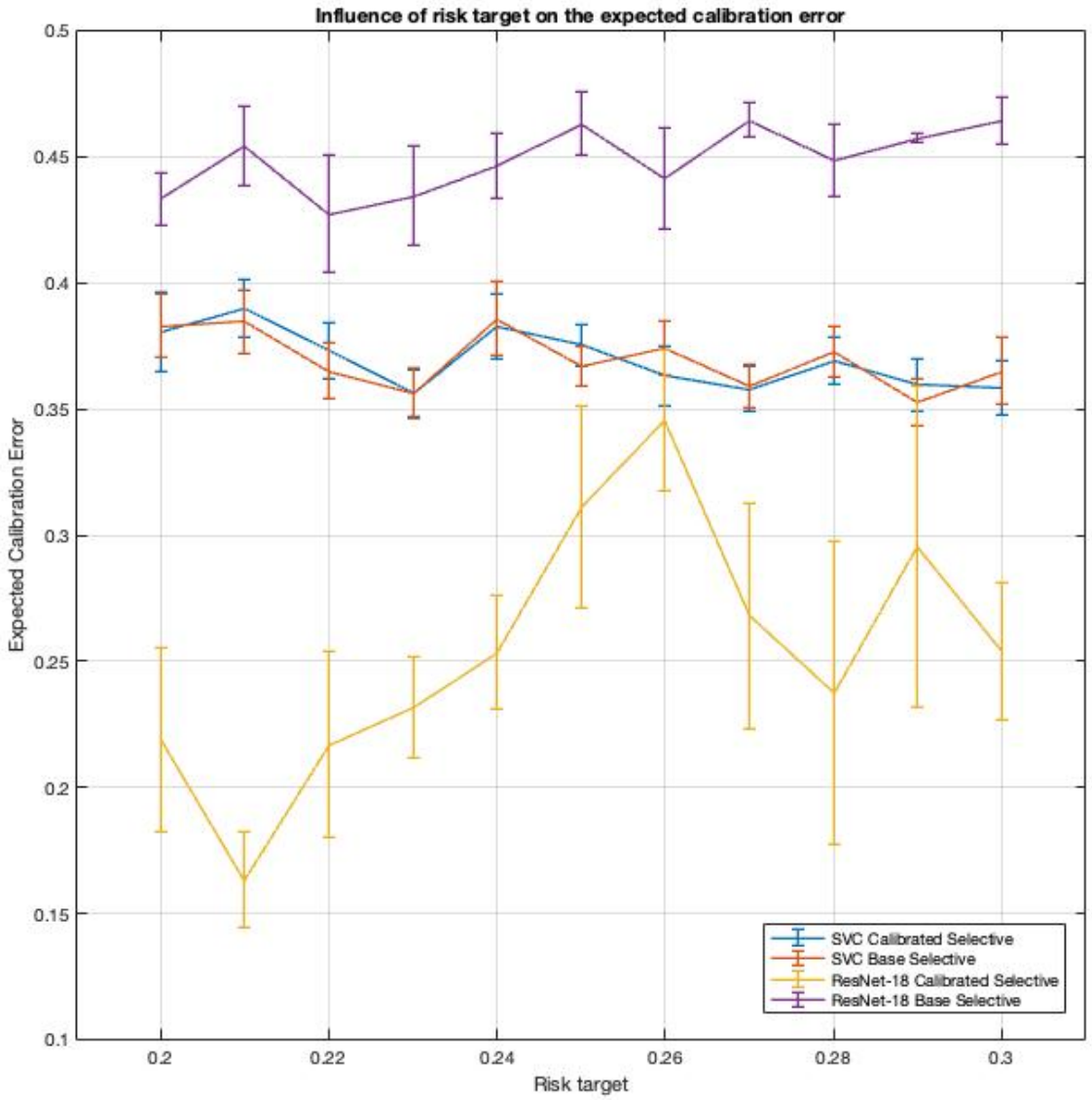


Figure 5.7: Impact of varying risk target value on the expected calibration error for base selective and calibrated selective SVC and ResNet-18 models. Each point is a mean value of 10 trials and has a corresponding standard error.

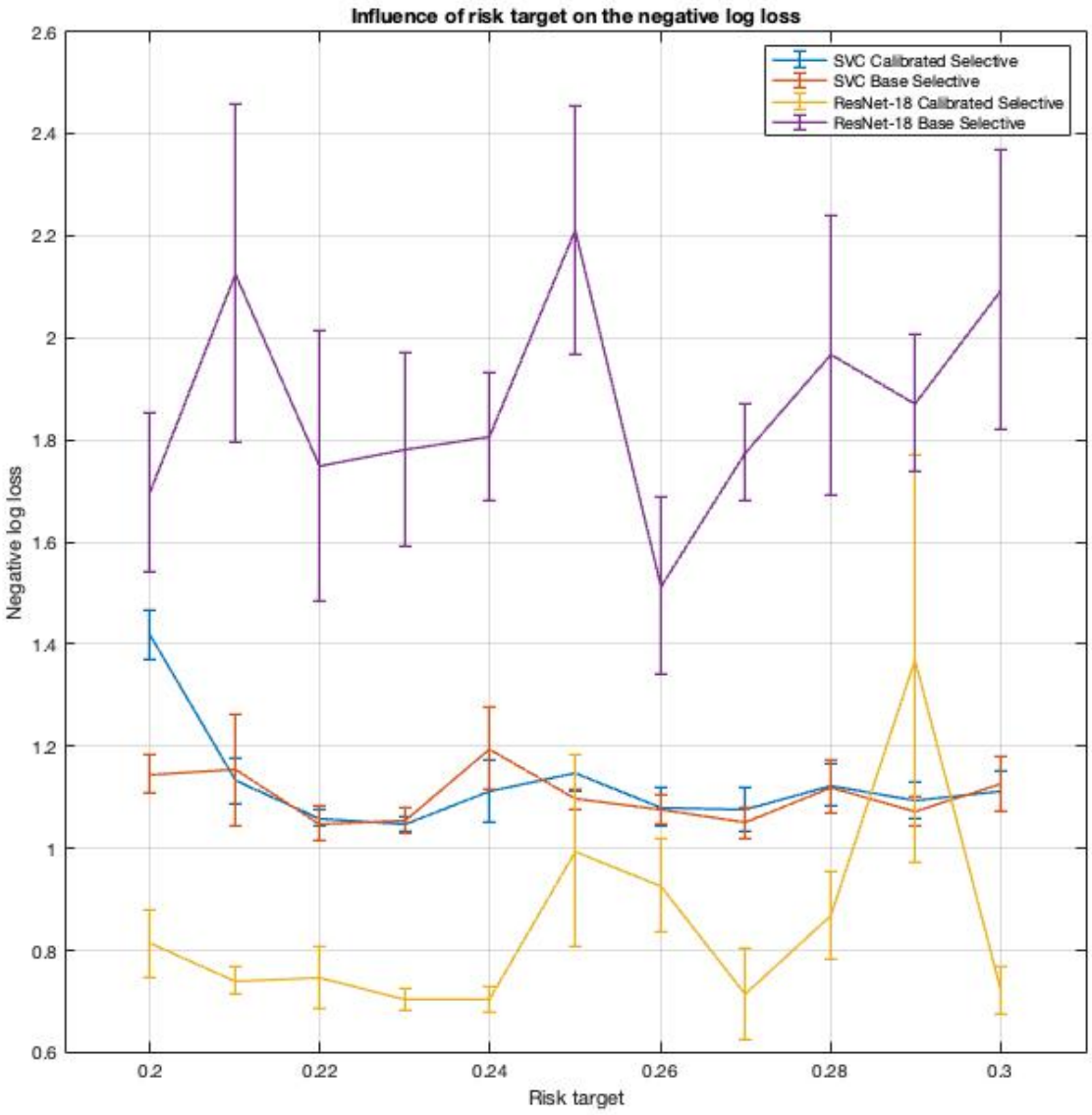


Figure 5.8: Impact of varying risk target value on the negative log loss for base selective and calibrated selective SVC and ResNet-18 models. Each point is a mean value of 10 trials and has a corresponding standard error.

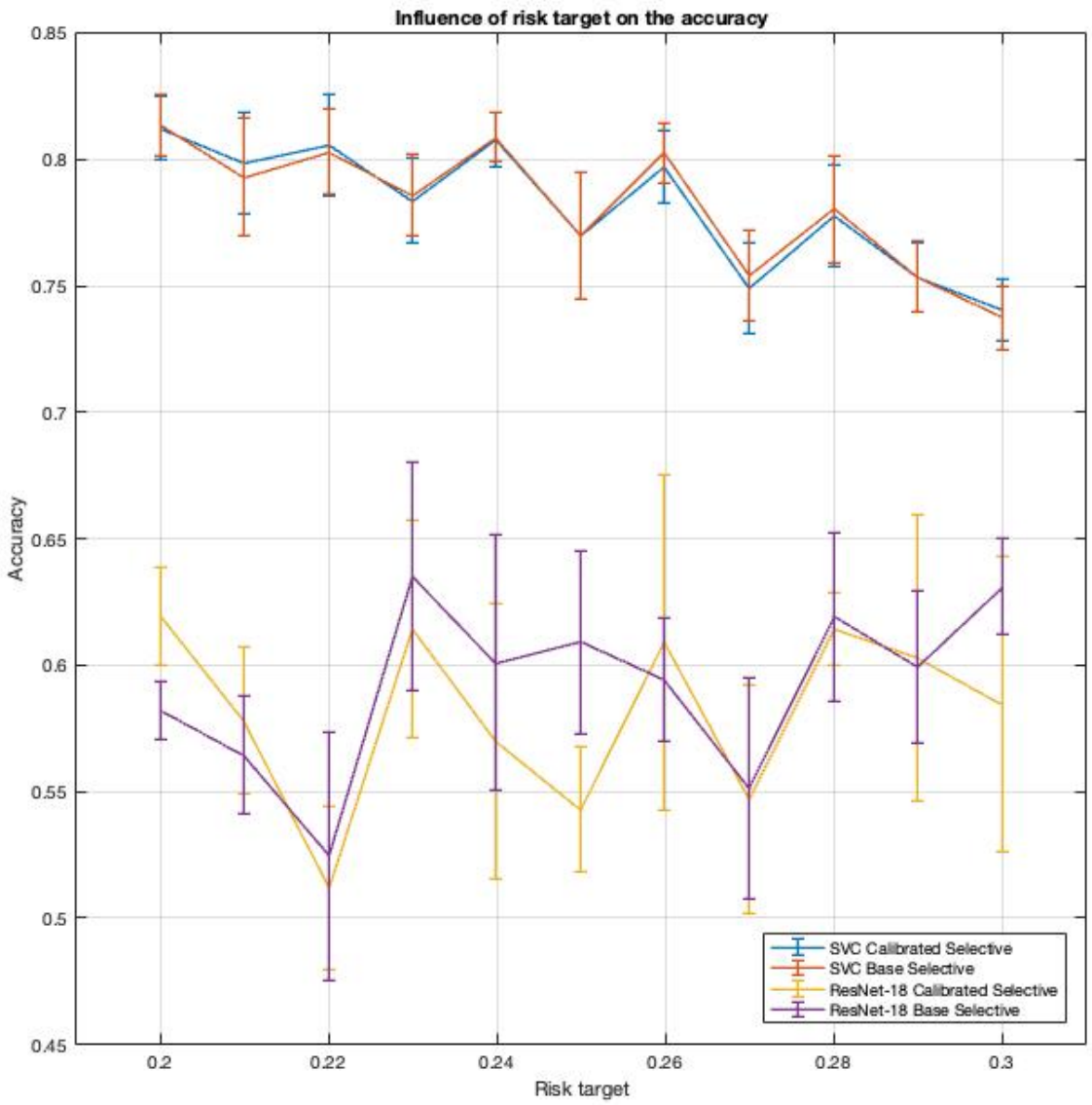


Figure 5.9: Impact of varying risk target value on the accuracy for base selective and calibrated selective SVC and ResNet-18 models. Each point is a mean value of 10 trials and has a corresponding standard error.

5.3 Data Augmentation - Experiments

The algorithms in this experiment were introduced to basic data augmentation techniques. To cope with the imbalancement issue illustrated by Figure 4.2, 50 artificial images of class 0 were added to the training set. Affine transformations consisted of a horizontal and vertical flip as well as a rotation. Coloured transformations were focused on gamma change, brightness, hue, saturation and contrast. Combined transformations contained all of the above. Similarly as for the previous experiments, for each algorithm and type of data augmentation, 10 trials were conducted after which a mean and standard error were calculated.

Table 5.5 demonstrates the results of the application of all of the data augmentation methods. Regarding the SVC and logistic regression, the only statistical significance was shown for the change of accuracy whereas ECE and negative log loss remained relatively similar. On the other hand, ResNet-18 benefited remarkably from coloured and combined transformations. The latter data augmentation methodology yielded the best accuracy score whereas purely coloured transformations produced the best ECE and negative log loss values. For the deep learning architecture, each type of transformation improved accuracy where combined methods produced significantly better results compared to the others.

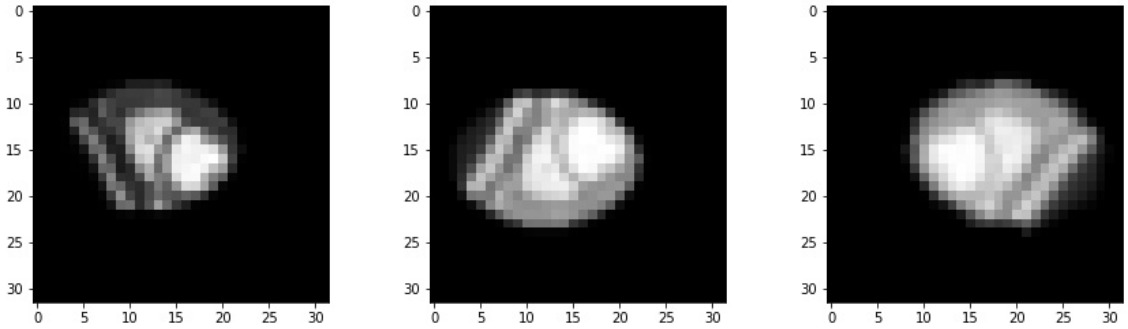


Figure 5.10: Examples of augmented data for the SVC and Logistic regression pipelines. From the left, rotated, brighter and combined.

A trend of distortion due to the affine transformations regarding the confidence metrics is observable for all of the classifiers. Notably, SVC and logistic regression affected by colour transformations had their accuracies dumped. Additionally, for all models, combined transformations always outperformed colour methods in terms of classification accuracy. The biggest benefit of data augmentation was received by the ResNet-18 for which the metrics improved significantly when colour or combined transformation were applied. Logistic regression had log loss and accuracy at the highest value for the base model and ECE was the lowest for colour transformations. Lastly, the SVC pipeline slightly benefited from coloured transformations in terms of calibration but the accuracy was significantly lower than the base model for which it was the highest among all of the SVC models.

Regarding the performance against the imbalancement problem, data augmentation methods showed diverse outcomes. Colour and combined transformations caused a decrease of all the metrics closely or exactly to zero for class 0 which was unexpected given the fact that there were more training examples for this particular class. However, affine transformations have improved the model's performance by increasing precision, recall and F1 scores. For SVC and logistic regression, base models had the best precision but superior scores for the other metrics were obtained after combined data augmentation was used. For all of the models, coloured transformations produced the worst results for the class 0. In Table 5.7 the classical machine learning algorithms' best performance was observed when no data augmentation was applied according to the recall and F1 scores. However, combined transformations significantly improved the precision score for SVC but for logistic regression the difference is not statistically significant.

SVC	**Base**	**Affine**	**Colour**	**Combined**
ECE	0.3446 ± 0.0164	0.3301 ± 0.0794	**0.3299 ± 0.0133**	0.3501 ± 0.0115
Log Loss	1.0385 ± 0.0353	1.042 ± 0.0314	**1.006 ± 0.0333**	1.0650 ± 0.0468
Accuracy	**0.7574 ± 0.0173**	0.7093 ± 0.01637	0.6629 ± 0.0175	0.7296 ± 0.0216
Logistic Regression				
ECE	0.3862 ± 0.0145	0.3789 ± 0.0205	**0.3620 ± 0.2232**	0.3907 ± 0.0111
Log Loss	**1.5864 ± 0.2868**	1.7135 ± 0.2926	1.6161 ± 0.2955	1.5692 ± 0.1423
Accuracy	**0.7685 ± 0.0132**	0.7315 ± 0.0126	0.6500 ± 0.0135	0.7222 ± 0.0117
ResNet-18				
ECE	0.4413 ± 0.0170	0.4420 ± 0.0091	**0.3670 ± 0.0343**	0.4324 ± 0.0140
Log Loss	1.7114 ± 0.1773	1.8363 ± 0.1344	**1.2243 ± 0.0938**	1.3788 ± 0.0772
Accuracy	0.5741 ± 0.0230	0.6074 ± 0.0183	0.6370 ± 0.0146	**0.7037 ± 0.0117**

Table 5.5: The effect of different data augmentation techniques on the classifiers performance. Base column stands for results for models without data augmentation whereas affine and coloured for basic data augmentation techniques. Combined transformations are a mixture of these two together. **Bolded** results outline the best results for a particular augmentation methodology. ECE and log loss is aimed to be minimised whereas accuracy is desired to be as close to 1 as possible

SVC	**Base**	**Affine**	**Colour**	**Combined**
Precision	**0.7289 ± 0.0284**	0.5799 ± 0.0628	0.5717 ± 0.0248	0.6681 ± 0.0225
Recall	0.5951 ± 0.0439	0.5847 ± 0.0750	0.5783 ± 0.0305	**0.6449 ± 0.0415**
F1-score	0.6478 ± 0.0341	0.5749 ± 0.0656	0.5662 ± 0.0154	**0.6481 ± 0.0243**
Logistic Regression				
Precision	**0.6898 ± 0.0211**	0.5906 ± 0.0661	0.5720 ± 0.0352	0.6775 ± 0.0215
Recall	0.6094 ± 0.0247	0.5445 ± 0.0679	0.5251 ± 0.0348	**0.6349 ± 0.0239**
F1-score	0.6444 ± 0.0194	0.5602 ± 0.0638	0.5381 ± 0.0248	**0.6499 ± 0.0129**
ResNet-18				
Precision	0.3278 ± 0.0672	**0.3901 ± 0.0726**	0.0 ± 0.0	0.05 ± 0.0476
Recall	0.2099 ± 0.0747	**0.2782 ± 0.0686**	0.0 ± 0.0	0.0571 ± 0.0542
F1-score	0.2559 ± 0.0576	**0.2647 ± 0.0299**	0.0 ± 0.0	0.0533 ± 0.0500

Table 5.6: The effect of different data augmentation techniques on the classifiers performance against imbalancement for class 0. Base column stands for results for models without data augmentation whereas affine and coloured for basic data augmentation techniques. Combined transformations are a mixture of these two together. **Bolded** results outline the best results for a particular augmentation methodology. For all the metrics, the closer to 1, the better the model's performance.

SVC	**Base**	**Affine**	**Colour**	**Combined**
Precision	0.7286 ± 0.0261	0.7997 ± 0.0296	0.7739 ± 0.0199	**0.8174 ± 0.0229**
Recall	**0.8688 ± 0.0155**	0.8324 ± 0.0211	0.7653 ± 0.0244	0.8291 ± 0.0164
F1-score	**0.8214 ± 0.0183**	0.8096 ± 0.0136	0.7655 ± 0.0148	0.8204 ± 0.0133
Logistic Regression				
Precision	0.7802 ± 0.0245	0.7561 ± 0.0297	0.7232 ± 0.0267	**0.7829 ± 0.0215**
Recall	**0.8416 ± 0.0099**	0.8357 ± 0.0219	0.7604 ± 0.0348	0.8153 ± 0.0155
F1-score	**0.8082 ± 0.0164**	0.7867 ± 0.0136	0.7372 ± 0.0182	0.7965 ± 0.0133
ResNet-18				
Precision	0.6332 ± 0.0341	0.6400 ± 0.0203	0.6333 ± 0.01522	**0.6509 ± 0.026**
Recall	0.6914 ± 0.0551	0.6909 ± 0.0745	**1.0 ± 0.0**	0.9800 ± 0.0189
F1-score	0.6610 ± 0.0366	0.6348 ± 0.0489	0.7744 ± 0.0116	**0.7763 ± 0.0130**

Table 5.7: The effect of different data augmentation techniques on the classifiers performance against imbalancement for class 1. Base column stands for results for models without data augmentation whereas affine and coloured for basic data augmentation techniques. Combined transformations are a mixture of these two together. **Bolded** results outline the best results for a particular augmentation methodology. For all the metrics, the closer to 1, the better the model's performance.

For the class 1, the deep learning architecture had the best results for coloured and combined transformations which is the opposite of the results presented in Table 5.6 for the same algorithm. Nonetheless, the differences between them were not proved to be statistically significant.

CHAPTER 6

Discussion

All of the proposed methods for confidence estimation enhancement were implemented and tested. The results from Platt scaling experiments indicate it is capable of developing the confidence of some algorithms. For SVC and logistic regression, the implementation was through cross-validation whereas for the ResNet-18 by using a separate validation set to remove an unwanted bias.

6.1 Confidence Calibration

The most effective improvement was observed for ResNet-18 for which ECE and logarithmic loss were minimised which indicates a confidence estimation improvement. Regarding its reliability diagrams in Figure 5.3, Platt scaling has translated extreme confidences from the lowest and highest confidence bins towards the centre. Calibrating the SVC model was less effective as there is only a slight improvement in the calibration metrics but without a statistical significance for the ECE. Although the scaling has slightly reduced the ECE for some confidence bins according to Figure 5.1. There was no difference observed between pre and post-calibration of the logistic regression model.

6.1.1 Baseline

Naeini et al. in [33] showed an improved calibration of the SVM with Platt scaling on their main dataset. However, the authors noted a deterioration of logistic regression after implementing this calibration technique. In [28], Niculescu et al. tested the efficacy of calibration techniques including Platt scaling by tracking the squared error (SE) and log-loss. Support vector machine after being calibrated with Platt's method, the logarithmic loss declined which is a comparable outcome to the one shown in Table 5.1. The same experiment authors performed on the logistic regression which as well did not show any significant difference compared to their base models.

Pollastri et al. in [34] attempted to calibrate deep learning architectures such as ResNet-18. The authors made a comparison of the models before and after being introduced to the Platt scaling. For this study, they used pre-trained models which is a significant difference compared

to the dissertation because they had an opportunity to compute weights appropriately. Although, in their study, a calibration impact is regarded separately for each class which leads to divergent results in each metric, especially for the ECE which improved for the parietal class and decayed for the mesangial class. However, if an average of these two results are considered, then the ECE value dropped compared to the uncalibrated model indicating a successful calibration. A similar outcome of decreased metrics after the calibration was observed in Table 5.1.

6.1.2 Contributions

There is a small number of research papers considering an improvement of confidence calibration through calibration for a cardiac MRI dataset specifically. Therefore, this study contributes to the field by describing the effect of Platt scaling on the algorithms for this particular type of dataset.

6.1.3 Comparison to the state of the art

It is impossible to make a valid direct comparison of results between studies due to the dissimilarity of datasets or models used. However, trends of having at least a slight improved confidence estimation after Platt scaling for the SVC and ResNet-18 have remained which matches the results obtained in Table 5.1. Regarding the logistic regression, Platt scaling either did not influence or degraded its confidence estimation capability. It is possible that the results from Table 5.1 and other researches empirically suggest that logistic regression already produces well-calibrated predictions hence application of the calibration technique might not affect it or introduce some noise for the given dataset.

6.1.4 Suggestions for a future research

Although results produced mostly calibrated models, there are a few suggestions to be considered for potential research in the future.

- Larger dataset of medical images could potentially produce better results. Although when used both cross-validation and validation sets were to remove bias, they used a small amount of data hence limiting the effect of Platt scaling.
- An exploration through other calibrating techniques such as Bayesian binning into quantiles (BBQ) might be helpful to determine their effectiveness in the medical imaging context.

6.2 Selective Classification

A rejection option with the SGR algorithm has been developed to create a selective classifier consisting of a selective function and a predictive model. A novel method of combined SGR and Platt scaling has been implemented and tested. Unfortunately, the SGR algorithm to produce an appropriate selection threshold requires more data than it could be delivered by a separate validation set. Thus, a training set has been used which makes the implementation prone to a potential bias.

6.2.1 Baseline

Owing to the novelty of the SGR and Plattscaling combined methodology, to my knowledge, no study could be compared to this part of the project. Moreover, the selective approach is not a widely known subfield in confidence estimation. Therefore, this specific combination of selective classification and confidence calibration is a subject for further discussion and testing.

6.2.2 Contributions

The major disadvantages of this study are a class imbalancement of the dataset and a small number of images which caused some models to overfit. For the classifiers included in the MPCA pipeline the overfitting issue was overcome which was shown by results from Tables 5.2 - 5.4 due to the accuracy scores on testing data being relatively good. Another weakness of this study is a potential biased introduction due to a choice of a threshold based on a training set, not cross-validation or a validation set. Unfortunately, there were miniature but noticeable differences between F1 scores of both classes suggesting a slight imbalancement. This specific issue was more severe for the ResNet-18 models that can be explained by insufficient well-trained parameters in the network. Therefore, the selective approach was examined under these issues. The results from the experiments show that the selective approach is capable of a slight improvement of model performance against the imbalance owing to the most improved metrics visualised in Table 5.3 and Table 5.4. This means that the selective function in most cases rejects invalid predictions which resulted in a significant improvement of imbalancement scores of calibrated selective ResNet-18 model for class 0 with a huge emphasis on a precision metric. In fact, Platt scaling by transforming confidence estimates results in a selection of different selective thresholds to a model without it ultimately resulting in a dissimilar number of rejected predictions.

Additionally, the selective approach made the deep learning model perform better at classifying images of the minority class according to the imbalancement metrics. This might be due to the that a rejection function successfully dropped incorrect predictions for this class as the model did not have enough training data to be capable and confident when predicting them. It is believed that this approach could be effective generally for deep learning models because they tend to push predictions toward extreme confidence as visualised in Figure 5.6. However, in this study, the architecture did not have a sufficient amount of data to be trained on hence this hypothesis should be tested on a larger dataset. SVC and logistic regression have benefited from the methods with a selective function when handling the imbalancement problem. Thus, the rejection mechanism works as expected.

A combination of the SGR algorithm with the Platt scaling showed enhancement in some experiments for calibration. In Table 5.2 an impressive improvement in the calibration of ResNet-18 was achieved. When compared to the results with only Platt scaling applied in Table 5.1, the calibrated selective approach produced an even lower ECE value but a slightly higher score for the negative log loss. Unfortunately, base selective and calibrated selective approaches were not successful in enhanced calibrating of the SVC and logistic regression. Selective function stopped predictions hence accuracies per certain confidence bins became more distant from the values ideal to a perfect calibration. Regarding the negative log loss, a selective function by abstaining from making certain predictions which did not contain much information causes an increment of this metric.

Although this method did produce similar results for the classical machine learning algorithms. Actually, according to the ECE and log loss metrics, selective and fully calibrated approaches miscalibrated the models but simultaneously improved their accuracy. In Figure **??** - **??** a drop in counts with confidence around the centre of the confidence domain was visualised as caused by a rejection of predictions with too low confidence. High maximal confidence for the opposite class is a likely reason why some predictions with low confidence remained for ResNet-18 and SVC. Logistic regression produced more gradual confidence which the confidence function used to produce

values which after being introduced to the SGR algorithm could produce lower threshold values compared to the other models given the same risk target and confidence parameter.

Figures 5.7 - 5.9 could give insights on optimal risk target values for this particular problem. However, these results are not necessarily generalisable for other machine learning models. In this part of the project, a few trends were visualised such as a downtrend of the accuracy of SVC algorithms or convergence of ECE for calibrated selective and base selective SVC models. The higher risk target value usually resulted in a lower selective threshold value which could explain the downtrend as more incorrect predictions would not be dropped. For the calibration metrics, steady-state values were reached which might suggest that SVC models are already calibrated for the given dataset because the change of the selective threshold does not result in its change.

6.2.3 Comparison to the state of the art

Geifman et al. in [8] proposed the SGR algorithm against a different metric, the risk-coverage curve. However, the authors did not explain how this curve would track an improvement of the confidence estimation of a machine learning model. Additionally, their pseudocode for the SGR is misleading as it states that a selective classifier is an output where in reality a selective threshold is produced. Therefore, the results cannot be compared as each study tracks separate aspects. To my current knowledge, no study assesses an application of this specific algorithm. Moreover, a selective classification is rarely a topic of discourse even if it is regarded as one of the use cases for confidence estimation [7]. Perhaps this research could benefit from an investigation of the proposed methods with respect to the risk-coverage metric.

6.2.4 Suggestions for a future research

- Using a larger dataset to create a sufficiently large validation set or implementing a specific cross-validation technique that could protect the implementation of the SGR from a potential bias.
- Exploring different selective functions could produce perhaps better results. For example, a function with an additional constraint on extreme confidence that if a selective threshold is going to be very high, then it might reject all of the predictions which is an undesirable case.
- If various calibration techniques are considered, then testing their combinations with a selective prediction would be interesting to see.
- Researching a new method for computation of an appropriate selective threshold θ. Perhaps this method could perform a different kind of search over a list of confidences to find an optimal solution.

6.3 Data Augmentation

Simple transformations of the training data were used to improve confidence estimation and remove the imbalancement of the models. Validation and testing data remained unchanged as it is crucial to test machine learning models with samples that are as close to real-world examples.

6.3.1 Baseline

Data augmentation in the context of confidence estimation is a subfield under development. To my knowledge, no study considered simple data augmentation techniques for such as affine

transformations in the context of improving confidence estimation of a predictive model. However, the Mixup data augmentation method is a frequent research topic in calibration [10], [35]. In addition, seems that researchers are more interested in testing more sophisticated data augmentation techniques for this particular problem [36].

6.3.2 Contributions

This study investigated the effect of affine, colour and their combination transformations on the aforementioned problems. As expected, ResNet-18 with more images performed better than the base model due to having more data to be trained on. In addition, a significant difference between colour and combined transformations is present. Surprisingly, for these transformations, the imbalancement metrics for class 0 were drastically decreased even though only images of this class were augmented to the training set. Thus, this result is opposite to what was expected. Moreover, for class 1 the recall is almost perfect which overall by no means defines a good classifier. Regarding the affine transformations, this drastic change was not noticed. These transformations instead improved metrics for class 0 scores and the other changes mostly were not statistically significant. In this case, the colour transformations disturbed the model's ability to recognise negative class correctly. This could indicate that these transformations are not suitable for the MRI image data owing that the correctness of a prediction of a medical image relies on small details that could be concealed by them.

Classical algorithms did not face the same issue. In terms of calibration, logistic regression had only improved ECE when coloured transformations were applied. For SVC, best ECE and log-loss were observed for the same data transformations. Also, combined transformations improved recall and F1 scores for class 0 of these models. However, data augmentation led to a much worse accuracy which compared to a positive change of a few metrics cannot be regarded as a worthwhile trade-off.

6.3.3 Comparison to the state of the art

As mentioned before, the research is focused on more complex transformations. Patel et al. [36] investigated methods such as Mixup, confidence enhancing data augmentation (CEDA) and on-manifold adversarial data augmentation (OMADA) and their effect on calibration. The results showed on CIFAR-10 and CIFAR-100 empirically prove that data augmentation is capable of a calibration improvement for some algorithms and datasets. The effects obtained, for instance, on DenseNet on CIFAR-10 are much more influential than the other ones conducted in this particular research. Although it is not quite correct to directly compare the dissertation with this study but both indicate that there is room for improvement of confidence estimation with data augmentation.

6.3.4 Suggestions for a future research

Even though proposed data augmentation techniques were not successful in calibrating models, there are a few suggestions for future research:

- Testing more sophisticated data augmentation techniques such as mixup, CEDA or even GANs.
- Data augmentation Mixup conducted concerning a difference between accuracy and confidence of a sample has been proposed in [37] which could be an interesting topic for potential research to investigate further with an aim for confidence estimation improvement.

CHAPTER 7

Conclusion

A slight improvement in confidence estimation has been achieved but with a lot of room for advancement. Although the positive outcomes mostly related to the deep learning architecture which was prone to overfitting, some small improvements for classical machine learning algorithms were obtained. The aims and objectives of this research project have been met but what is more important the results are reckoned to motivate further research to be more focused on the potential of a selective prediction and a data augmentation for an improvement of confidence estimation.

Data augmentation is especially interesting for the classification of medical images as the data already discussed in the dissertation can often be hard to get. Also, the medical dataset usually is imbalanced which makes this subject even more important.

Selective classification has some potential on producing more calibrated models but there is not much literature currently available. Therefore, there are a lot of opportunities to improve currently available methods.

The main goal of empirically proving the possibility of confidence estimation enhancement of machine learning models with other techniques than widely-known calibration has been achieved.

CHAPTER 8

Project Management

Project planning was subject to many changes during its duration because whenever more knowledge was acquired, the more significant changes were applied. After receiving the feedback on the interim report and revisiting the Gantt Chart, some objectives listed in Figure 8.1 appeared to be not relevant to the scope of the project. For instance, when the project started, the objective was to use many different algorithms for a comparison study such as random forests and gaussian-naive classifier. However, since the main theme of the dissertation was confidence estimation it was necessary to focus more on methods related to it than machine learning algorithms. Thus, despite of SVC, logistic regression and ResNet-18 models, all of the algorithmic implementations were dropped. When the second semester started, the Gantt chart was visited frequently and updated with respect to new findings to make the planning more effective. The new plan presented in Figure 8.2 dropped irrelevant objectives and introduced new during the project which were chosen carefully so that they would be realistic to complete.

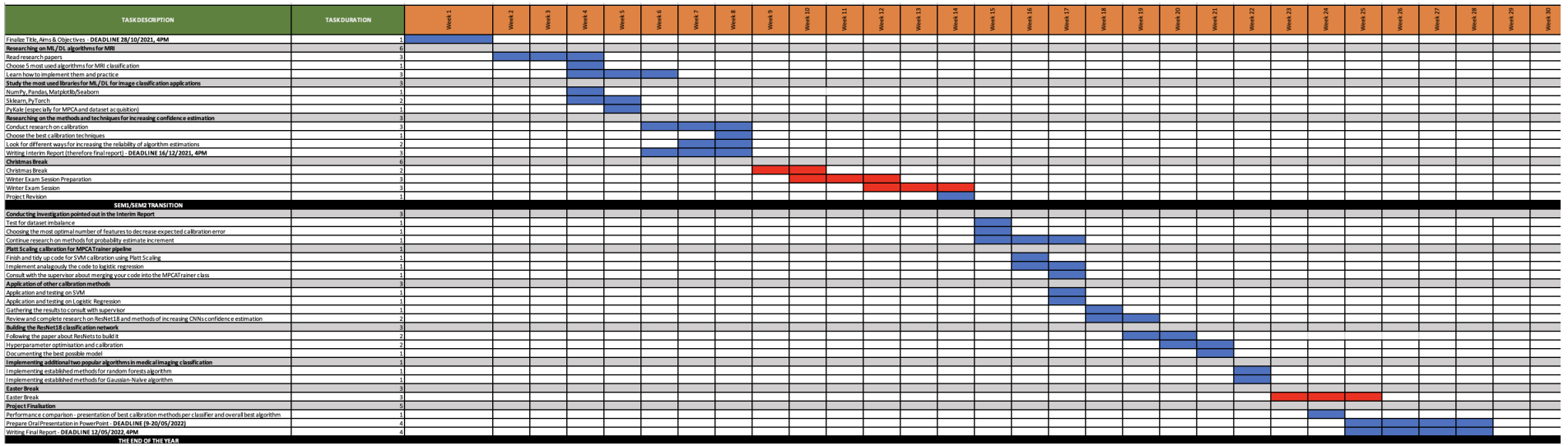


Figure 8.1: Initial Gantt Chart from the Interim Report.

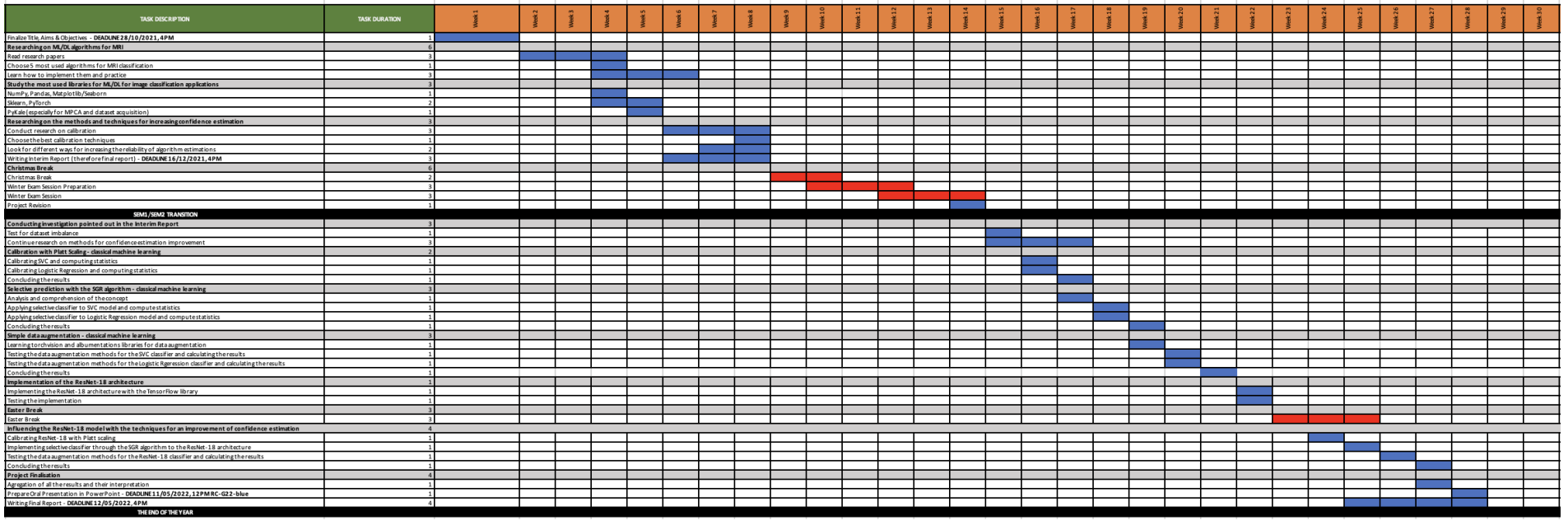


Figure 8.2: Final Gantt Chart for the project planning.

CHAPTER 9

Self-Review

When I commenced this project at the beginning of the first semester I was afraid of not delivering a quality project and hence failing it. Due to that, I did not come from a computer science background, and starting a project in machine learning was a tough decision to make given all of the possible risks involved with it. However, I was and after finishing this project I am still passionate about the implementation of artificial intelligence in the fields related to human health and neurotechnology.

During the first semester, I was struggling with learning about the fundamentals behind machine learning and confidence estimation. Before the Christmas break began I have acquired most of the necessary knowledge to start experiments that were planned more professionally. The more knowledge I gained, the more I was realising that my lack of understanding correctly the field led me to produce inefficient and in some ways irrelevant objectives. However, this was a good sign that I was truly learning when working on this final year project.

After the winter examination session, I devoted myself fully to completing this project. Although the objectives have changed and I had much less time in the end I can claim that all of the aims have been completed by the deadline of the dissertation submission. Some of the results are not that spectacular but there are also a few which make the entire project much more interesting. I am aware that in some places this study is not perfect but I have pushed myself out of the boundaries that I thought are my actual limits. I have learnt more than I could have imagined and I am more motivated than ever to improve myself in the field. All of this is thanks to the opportunity I have used to accelerate my growth in the field of my interest. Without any doubt, I have learnt a lot when I compare my knowledge before commencing this project.

Bibliography

[1] E. Miranda, M. Aryuni, and E. Irwansyah, "A survey of medical image classification techniques," in *2016 International Conference on Information Management and Technology (ICIMTech)*. IEEE, 2016, pp. 56–61.

[2] M. A. Bruno, E. A. Walker, and H. H. Abujudeh, "Understanding and confronting our mistakes: the epidemiology of error in radiology and strategies for error reduction," *Radiographics*, vol. 35, no. 6, pp. 1668–1676, 2015.

[3] P. Robinson, "Radiology's achilles' heel: error and variation in the interpretation of the röntgen image." *The British Journal of Radiology*, vol. 70, no. 839, pp. 1085–1098, 1997.

[4] J. P. Earls, V. B. Ho, T. K. Foo, E. Castillo, and S. D. Flamm, "Cardiac mri: recent progress and continued challenges," *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 16, no. 2, pp. 111–127, 2002.

[5] C. B. Marcu, A. M. Beek, and A. C. Van Rossum, "Clinical applications of cardiovascular magnetic resonance imaging," *Cmaj*, vol. 175, no. 8, pp. 911–917, 2006.

[6] B. J. Erickson, P. Korfiatis, Z. Akkus, and T. L. Kline, "Machine learning for medical imaging," *Radiographics*, vol. 37, no. 2, pp. 505–515, 2017.

[7] Y. Ding, J. Liu, J. Xiong, and Y. Shi, "Evaluation of neural network uncertainty estimation with application to resource-constrained platforms," *ArXiv*, vol. abs/1903.02050, 2019.

[8] Y. Geifman and R. El-Yaniv, "Selective classification for deep neural networks," *CoRR*, vol. abs/1705.08500, 2017. [Online]. Available: http://arxiv.org/abs/1705.08500

[9] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1321–1330.

[10] S. Thulasidasan, G. Chennupati, J. A. Bilmes, T. Bhattacharya, and S. Michalak, "On mixup training: Improved calibration and predictive uncertainty for deep neural networks," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[11] M. Kull, T. M. Silva Filho, and P. Flach, "Beyond sigmoids: How to obtain well-calibrated probabilities from binary classifiers with beta calibration," *Electronic Journal of Statistics*, vol. 11, no. 2, pp. 5052–5080, 2017.

[12] A. J. Swift, H. Lu, J. Uthoff, P. Garg, M. Cogliano, J. Taylor, P. Metherall, S. Zhou, C. S. Johns, S. Alabed *et al.*, "A machine learning cardiac magnetic resonance approach to extract disease features and automate pulmonary arterial hypertension diagnosis," *European Heart Journal-Cardiovascular Imaging*, vol. 22, no. 2, pp. 236–245, 2021.

[13] S. Raschka and V. Mirjalili, *Python machine learning*, 2nd ed., 2019.

[14] M. Kim, J. Yun, Y. Cho, K. Shin, R. Jang, H.-j. Bae, and N. Kim, "Deep learning in medical imaging," *Neurospine*, vol. 16, no. 4, p. 657, 2019.

[15] C. Chen, C. Qin, H. Qiu, G. Tarroni, J. Duan, W. Bai, and D. Rueckert, "Deep learning for cardiac image segmentation: a review," *Frontiers in Cardiovascular Medicine*, vol. 7, p. 25, 2020.

[16] L. Auria and R. A. Moro, "Support vector machines (svm) as a technique for solvency analysis," 2008.

[17] W.-J. Lin and J. J. Chen, "Class-imbalanced classifiers for high-dimensional data," *Briefings in bioinformatics*, vol. 14, no. 1, pp. 13–26, 2013.

[18] D. Barber, *Bayesian reasoning and machine learning*. Cambridge University Press, 2012.

[19] J. Platt *et al.*, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in large margin classifiers*, vol. 10, no. 3, pp. 61–74, 1999.

[20] R. Dechter, "Learning while searching in constraint-satisfaction problems," 1986.

[21] P. Y. Simard, Y. A. LeCun, J. S. Denker, and B. Victorri, "Transformation invariance in pattern recognition—tangent distance and tangent propagation," in *Neural networks: tricks of the trade*. Springer, 1998, pp. 239–274.

[22] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel, "Handwritten digit recognition with a back-propagation network," *Advances in neural information processing systems*, vol. 2, 1989.

[23] Q. Li, W. Cai, X. Wang, Y. Zhou, D. D. Feng, and M. Chen, "Medical image classification with convolutional neural network," in *2014 13th International Conference on Control Automation Robotics Vision (ICARCV)*, 2014, pp. 844–848.

[24] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 Fourth International Conference on 3D Vision (3DV)*, 2016, pp. 565–571.

[25] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca, "An unsupervised learning model for deformable medical image registration," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9252–9260.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[27] H. Lu, X. Liu, R. Turner, P. Bai, R. Koot, S. Zhou, M. Chasmai, and L. Schobs, "Pykale: Knowledge-aware machine learning from multiple sources in python," *arXiv:2106.09756 [cs.LG]*, 2021.

[28] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in *Proceedings of the 22nd international conference on Machine learning*, 2005, pp. 625–632.

[29] B. Zadrozny and C. Elkan, "Transforming classifier scores into accurate multiclass probability estimates," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2002, pp. 694–699.

[30] X. Jiang, M. Osl, J. Kim, and L. Ohno-Machado, "Calibrating predictive model estimates to support personalized medicine," *Journal of the American Medical Informatics Association*, vol. 19, no. 2, pp. 263–274, 2012.

[31] Y. Geifman and R. El-Yaniv, "Selectivenet: A deep neural network with an integrated reject option," in *International Conference on Machine Learning*. PMLR, 2019, pp. 2151–2159.

[32] Y. Bahat and G. Shakhnarovich, "Classification confidence estimation with test-time data-augmentation," *arXiv preprint arXiv:2006.16705*, 2020.

[33] M. P. Naeini, G. Cooper, and M. Hauskrecht, "Obtaining well calibrated probabilities using bayesian binning," in *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[34] F. Pollastri, J. Maroñas, F. Bolelli, G. Ligabue, R. Paredes, R. Magistroni, and C. Grana, "Confidence calibration for deep renal biopsy immunofluorescence image classification," in *2020 25th International Conference on Pattern Recognition (ICPR)*, 2021, pp. 1298–1305.

[35] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[36] K. Patel, W. Beluch, D. Zhang, M. Pfeiffer, and B. Yang, "On-manifold adversarial data augmentation improves uncertainty calibration," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 8029–8036.

[37] Y. Wen, G. Jerfel, R. Muller, M. W. Dusenberry, J. Snoek, B. Lakshminarayanan, and D. Tran, "Improving calibration of batchensemble with data augmentation," *TWorkshop on Uncertainty and Ro-Bustness in Deep Learning*, 2020.