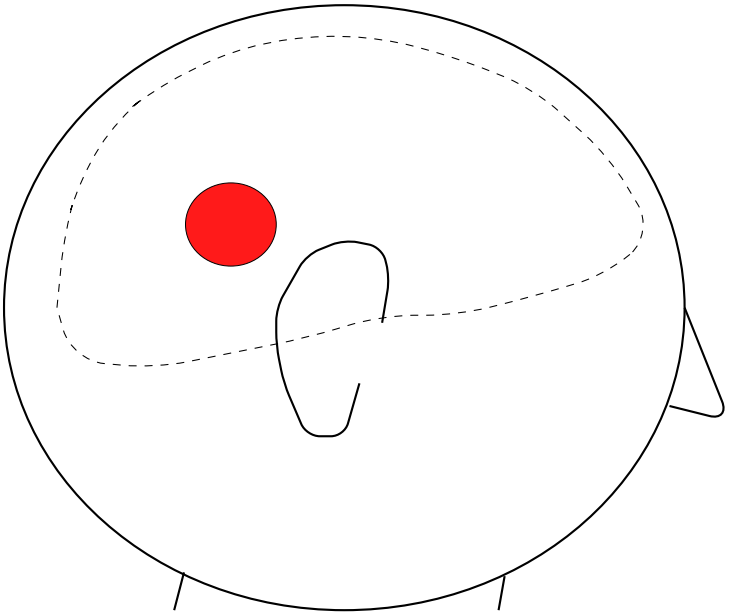


Predicting Position and Volume of Hemorrhagic Strokes

Using Machine Learning on Microwave Data

Master's thesis in Complex Adaptive Systems

EMMA NIRVIN

DEPARTMENT OF PHYSICS

CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2022
www.chalmers.se

Master's thesis 2022

Predicting Position and Volume of Hemorrhagic Strokes

Using Machine Learning on Microwave Data

EMMA NIRVIN

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Physics
Chalmers University of Technology
Gothenburg, Sweden 2022

Predicting Positon and Volume of Hemorrhagic Strokes
Using Machine Learning on Microwave Data
EMMA NIRVIN

Supervisor: André García Gómez, Ann-Sophie Hilkert, Medfield Diagnostics AB
Examiner: Kristian Gustavsson, Department of Physics

Master's Thesis 2022
Department of Physics
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: A brain bleed, image by author

Typeset in LaTeX
Printed by Chalmers Reproservice
Gothenburg, Sweden 2022

Predicting Position and Volume of Hemorrhagic Strokes
Using Machine Learning on Microwave Data
EMMA NIRVIN
Department of Physics
Chalmers University of Technology

Abstract

This study has explored different neural network methods for position and volume prediction of hemorrhagic strokes. Three different data sets of microwave data were used as input for the different networks. A simple diagnostic classifier was used as benchmark to help in evaluating success. The two largest challenges of the study was instrument variations in the data as well as the limited data available. A Multi-Source Adversarial Domain Adaptation (MSADA) network was introduced to lower the effect of the instrument variations, and a Divergence Based Domain Adaptation (DBDA) network was implemented to attempt to resolve the limited number of data samples.
The networks showed promising results for both position and volume in all three data sets used. The MSADA network successfully lowered the instrument specific noise when predicting volumes, but was concluded to be unnecessary for position classification. The DBDA network was not enough to remedy the lack of sufficient data.

Keywords: Deep neural networks, machine learning, adversarial domain adaptation, multi-source adaptation, microwave data, stroke detection.

Acknowledgements

First, I would like to thank my supervisors Ann-Sophie Hilkert and André García Gómez for all the support during this thesis. Your input and commitment has been greatly appreciated.
Next I want to thank Johanna Matero for all the support during this thesis. Being able to discuss everything and anything with you has been invaluable.
I would also like to thank Medfield Diagnostics AB for allowing me to finish my studies with them. It has been an interesting and informative end to my years at Chalmers.
Lastly I want to thank family and friends for all the times they've helped during my studies.

Emma Nirvin, Gothenburg, January 2022

Contents

List of Figures

List of Tables

1
Introduction

The use of machine learning and other forms of artificial intelligence (AI) is very wide-spread in today's society and is largely integrated in everyday life. In the medical areas where artificial intelligence has been introduced, it is showing very promising results and can sometimes outperform clinicians in diagnosing for example skin cancer [2].

1.1 Background

In 2019, more than 20 000 patients were diagnosed with a stroke at a Swedish hospital [3]. A stroke is therefore one of the most common causes of death [3]. How fast a stroke is diagnosed and treated largely affects the extent of the damage done to the brain [4]. This means that diagnosing and starting treatment early can be crucial for recovery. There are typically two different conditions referred to when talking about strokes. The first being the ischaemic stroke (IS), which means that a blood clot has gotten stuck in a blood vessel and blocks blood flow to some part of the brain. The other condition is the hemorrhagic stroke (HS), where a blood vessel instead bursts, allowing blood to build up and put pressure on the brain [5]. Since a stroke can be either hemorrhagic or ischaemic, an incorrect diagnosis can result in paramedics giving blood thinners to someone with a brain bleed thinking it was a clot, which will only aggravate the state of the patient.
Diagnosing a stroke can sometimes be difficult, especially in the early stages, with studies showing accuracies between 52% and 72% in prehospital diagnosis depending on the background of the medical personnel [6]. Today's process of making a prehospital diagnosis can include multiple tests and scans. These tests can be measurements of cholesterol, blood sugar levels, blood pressure and pulse to name a few [7]. The National Institute of Health Stroke Scale (NIHSS) is another tool used in the decision process, which examines 11 different items such as field of vision and facial palsy to determine the ability of the patient [8]. When arrived at the hospital, the patient can undergo scans for a magnetic resonance image (MRI) or a computed tomography (CT) [7].
To aid in the decision process in the ambulance, Medfield diagnostics AB have developed a classification tool called Strokefinder MD100 (see figure 1.1) [1]. It is a machine with eight antennas which are placed around the head of the patient. With the help of microwave data collected from these antennas, the instrument then classify the patient as either healthy or having a stroke. This information is to be used as a guide rather than a definitive answer. By further developing the

Strokefinder, the hope is that patients with a stroke will always be brought to a hospital with the correct equipment to treat them. The idea behind this thesis is that by using the knowledge of position and volume of the hemorrhages used during training, the classifier may learn to predict the patients situation more accurately. By knowing the volume of the bleed the network could potentially differ between smaller bleeds in a way that may be confusing to a classifier which counts a stroke with 100 ml blood the same as that with 0.3 ml. Similarly, by knowing the position the classifier could possibly learn the difference between a bleeding close to the left temple and one behind the right ear for example.

Figure 1.1: The Strokefinder MD100 [1]

1.2 Scope

The aim for this project is to create an AI that can predict position and volume of hemorrhagic strokes. It is also interesting to see if knowing position and volume can increase the accuracy when classifying a sample as either healthy or hemorrhaging. In this thesis, we will work with three different data sets, which will be called A, B and C in this report. How these were collected and the main differences between them are explained in section 3.1. First, data set A is used to test the theory. Then, a switch to data set B will be done to see how the theory transfer to the instruments, where noise and instrument specific properties are present. Lastly, data set C is used to see how good the performance is when the data has different types of noise and with a non-uniform distribution of position as well as volume.
Data set B is collected from different instruments of an old version of the Strokefinder MD100, and therefore have different input distributions due to small differences in the hardware and how it is assembled. These differences could lead to the network learning the instrument-specific noise, rather than the diagnosis. This in turn could make the network perform worse when tested on data from an instrument it hasn't been trained on compared to if the instrument was part of the training. Similarly,

data set C is collected from many different instruments from different revisions. In later revisions of the instrument these variations are much better controlled, but to enable the use of older data an effort was spent in mitigating these effects. A Multi-Source Adversarial Domain Adaptation (MSADA) model is introduced to examine whether such effects of instrument variations can be removed and if that improves the performance. This will be further explained in section 2.1.1. Another difficulty is the number of samples, especially in data set C. As this study only considers hemorrhagic strokes, where the position or volume have been specified, only a small subset of the data collected by the company is used. This limited and non-uniform data leads to a high risk of overfitting on the training samples and the network may not be able to generalise.
In both MRI and CT, the data collected from the instruments is transformed into an image. On such an image, an AI can perform image segmentation to find a stroke and therefore show classification, position and volume directly [9, 10]. Creating such an image is currently not possible to do with the data from the Strokefinder MD100, and is therefore not considered for this project.
Another limitation to this project is that data describing ischaemic strokes is not used because of the differences in how the two different types of strokes affects the brain. The network will therefore not learn the differences between a bleed and a clot and can instead focus on the differences present in hemorrhagic strokes.
The data outputted from the instruments is complex valued. While works such as [11] has found that keeping the complex nature of the data can improve the performance, work done in [12] suggests the complex networks require new and better solutions in areas such as activation functions and regularisation before it can be widely used in deep learning and signal processing problems. Therefore, this thesis will map the complex values into real numbers as described in 3.2.

1.3 Related studies

EM Tensor and EM Vision are companies with similar work as Medfield Diagnostics AB. EM Tensor has created a portable imaging machine called EMTensor BrainScanner [13]. This allows for rapid imaging, but the images needs to be viewed by a radiologist or neurologist for diagnosis rather than letting a machine learning algorithm find a possible stroke. EM Vision also have a hardware solution for imaging of the brain, but they introduce a machine learning algorithm for classification for diagnosis [14]. Articles published from the company, such as [15], show promising results on simulated data in classifying between IntraCranial Haemorrhages (ICH) and IS. Important to note is that they do not present any results showing performances in classifying healthy samples. Another problem they may not have considered is instrument variations, which could lead to a decrease in performance.
Abdulrahman S.M. Alqadami *et al.* developed a flexible electromagnetic cap with 16 antennas situated in a circular array on the cap [16]. The localisation of the bleed was done by computing the differences in the signals from a head without a bleeding compared to a head with a bleeding. They used both computer simulations and a realistic head phantom to collect their data. The position of the bleeding was also moved around when collecting the results and their imaging algorithm showed good

results in multiple positions, but had some difficulties when the bleeding was placed deep in the head.
The brain has a right-left symmetry in the same way as a face, meaning it only has some smaller dissimilarities. This was both verified and utilised in the paper by Aida Brankovic *et al.* [17]. They generated statistical fields in the brain by combining 3 or 4 different antennas and recorded the area each combination was affected by. Each area is then given a similarity score between 0 and 1.The score was compared to that of a reference medium of a homogeneous phantom with similar dielectric properties to a healthy head. The area covered by the statistical field is divided into regions by a mesh and the expected value of that area is computed from all fields covering that same area. The results showed promising results on larger strokes that are not too deep into the brain, but had difficulties on smaller strokes that were deeper in the brain.
Training on simulated data and testing on experimental data usually gives poor results, partially due to the noise introduced in the experimental data [11]. Ahmed Al-Saffar *et al.* proposed a complex convolutional network with domain adaptation to allow training a network on simulated data [11]. To localise the bleed, the imaging domain was divided into 31 areas and the localisation was reformulated into a soft-classification problem where each output corresponded to how much of the bleed was located in that area. The array containing these 31 values was normalised to show the predicted distribution. A separate branch in the network classified which domain the input data came from. While the results of the network was promising, the authors emphasised the limitations due to their data being measurements from only one phantom and notes that more extensive testing must be performed to give more trustworthy results.
In the spring of 2021, Ebba Ekblom and Rebecca Svensson wrote a master thesis on using a Generative Adversarial Network (GAN) to generate more data [18]. This was done in collaboration with Medfield Diagnostics AB, and used the same data used in this study. However, while their results were promising, further improvements of their work is ongoing and the choice was made to not include their generated data in this study.

2
Theory

This thesis uses dense neural networks as its base model. Built on top of this is a branch of transfer learning called domain adaptation, specifically adversarial domain adaptation for multiple source domains and a divergence based domain adaptation. The theory behind these concepts are introduced in this chapter.

2.1 Transfer learning

A standard neural network usually gets input from one specific source (called the source domain), and then performs a specific task. Sometimes however, we want to perform the same task as an existing network, but with input from a different source (called the target domain). Here is where transfer learning comes in handy. It enables the reuse and adjustment of the old network and data to combine it with the target domain [19]. This can save time, both by reusing a network, but also in data collection since less data is needed from the target domain.
In traditional transfer learning, the network is either kept completely, or some layers are removed or added [19]. Which method to use is largely dependent on what the specific layers are supposed to do. The first layers usually works as feature extractors, for example to find general shapes or contrasts in images, while the later layers find more details specific to the task at hand [19].

2.1.1 Domain Adaptation

Domain adaptation is a type of transfer learning that is used when both source and target domain are to perform the same task but the data comes from different distributions [20]. There are multiple types of domain adaptation, but they all have the same goal; to adjust the network in a way that removes the domain specific features of the source or sources. An adversarial based domain adaptation method can do this by confusing the network on which domain is which, while a divergence based domain adaptation method works by merging the different domains into a new domain with features from all sources as well as the target [**?**, 21].

2.1.1.1 Adversarial based Domain Adaptation

Domain adaptation can sometimes be based on adversarial training. A common way of implementing this is to use a Gradient Reversal Layer (GRL) [22]. This layer can be used in different ways, but most relevant in this thesis is the implementation of a separate domain classifying branch [22]. Somewhere in the existing network a

branching is implemented where the original task is performed in one branch and a domain classification is performed in the other, with the GRL as the first layer in this second branch. For multiple sources, it is found that using one branch per source generally improves the performance of the network [23]. The GRL uses an identity weight matrix during forward propagation, to not affect the input. During back propagation it reverses the gradient, making the training go backwards so that the network gets worse at distinguishing between the different domains [22]. The gradient can be scaled to determine the impact of the domain classification on the shared layers.

2.1.1.2 Divergence based Domain Adaptation

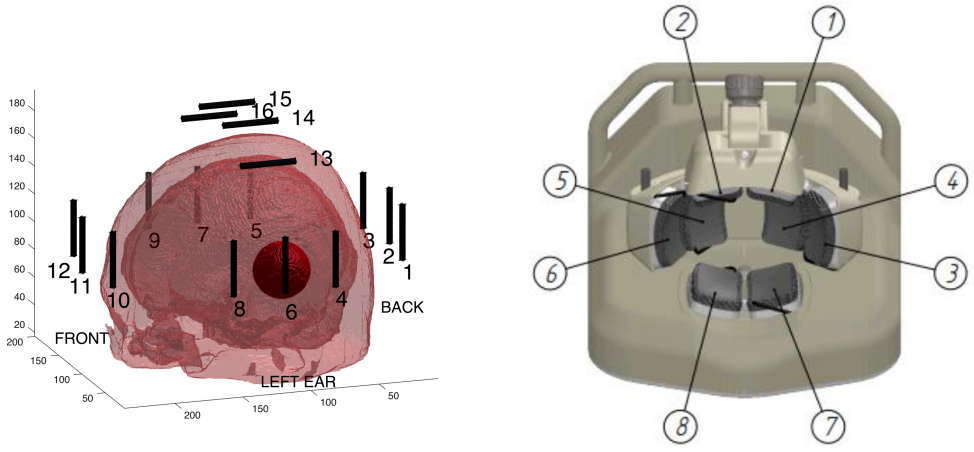
Another way of implementing domain adaptation is to use some divergence criteria to merge two domains into one [21]. This divergence can be computed after each layer, after the feature extractor, just before the output layer or any other combination [20, 21]. There is also many different divergence criteria which can be implemented in the model. The Wasserstein distance, KL divergence and correlation alignment are a few examples of such criteria [20, 21].

3
Methods

This section will cover the data sets used and the preprocessing done in this project. It will also describe the network structures, the training and the evaluation.

3.1 Data Sets

The data used in this study is divided into three sets, A, B and C, all of which was collected by Medfield Diagnostics AB prior to this study. The data consists of S-parameters from microwave antennas situated around the patients head as in figure 3.1. S-parameters, or scattering parameters, are a measurement of how much of the power in a signal entering a port leaves another port [24]. These can be used to describe how much signal is received in one antenna compared to what was transmitted in another. The S-parameters are computed once for each combination of pairs of antennas, with these pairs being called channels. The channels are symmetric, meaning the parameter for a signal from antenna 1 to antenna 2 is the same as that for antenna 2 to antenna 1. The number of unique channels are therefore computed by $N(N+1)/2$, where N is the number of antennas. There are 16 antennas for data set A, which equals 136 unique channels. For the other data sets, there are 8 antennas forming 36 unique channels.k



(a) Antenna positions in data set A

(b) Antenna positions in data set B and C

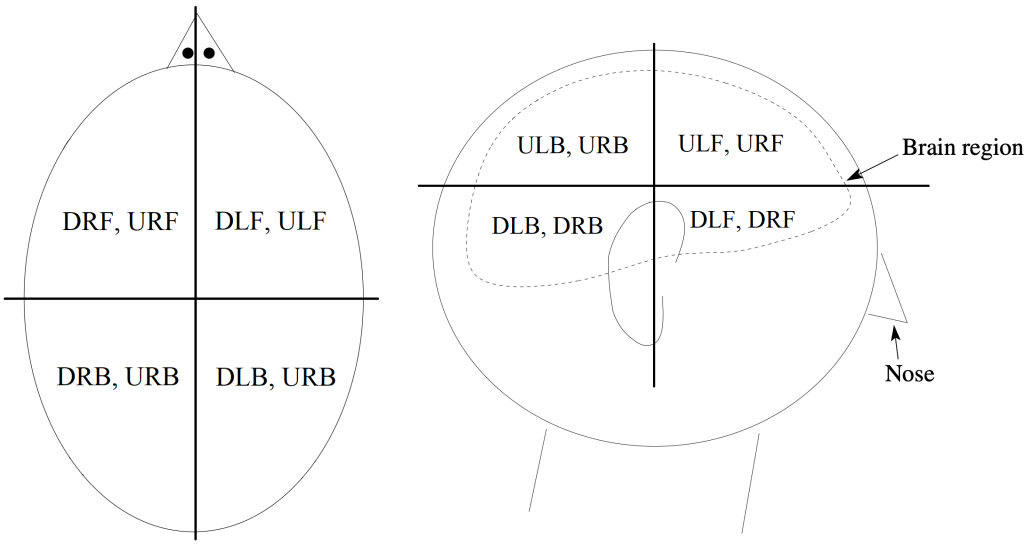
Figure 3.1: Positions of the antennas (Images by Medfield Diagnostics AB)

There are a couple main features that differs between the data sets. Data set A is

noise free, while data set B and C introduces noise. Data set A and B is uniform in volume, while data set C has some bias towards certain volumes. Data set B is uniform in positions, while data set A and C has some bias towards certain positions. The number of samples differs between the data sets, with A being the largest, and C the smallest. The following three sections will describe them more in depth.

3.1.1 Data set A

The first data set is collected through simulations of how microwave data would register on antennas situated around the patients head as shown in figure 3.1a. This data set has an equal numbers of samples representing healthy individuals as it has representing strokes, with a couple thousand samples for each of the two categories. The volumes of the stroke samples are uniformly distributed between 0 ml and 110 ml. The positions are uniformly distributed in the brain, but represented with a binary value for each of the three dimensions. These binary values divides the brain in eight parts by giving a 0 for left, 1 for right, and similar in the other two dimensions. This representation is how the data is stored, and is not done as a preprocessing step by the author. Due to the fact that the XY-plane crosses the Z-axis at half the brain height rather than adjusted to divide the brain volume in half, there are more samples in the 'down' category than in the 'up' category.



(a) Positions as seen from the right (XY-plane)

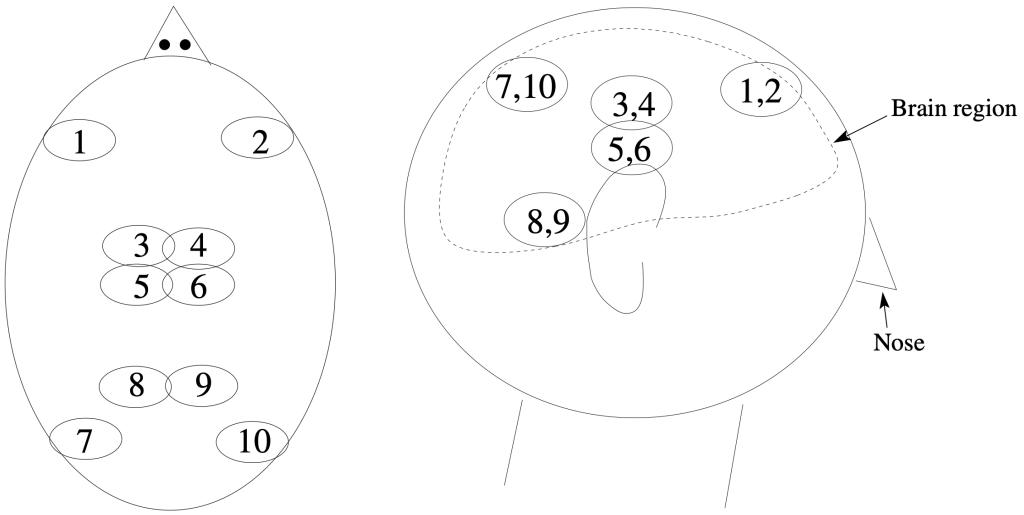
(b) Positions as seen from below (YZ-plane)

Figure 3.2: The positions in the simulation data (Image partly by Medfield Diagnostics AB, partly by author)

3.1.2 Data set B

Data set B is a bit different compared to the first data set. Firstly, the data was collected with three different Strokefinder MD100 instruments, which introduces noise.

The instruments are part of an old revision of the Strokefinder, which also introduces instrument specific variations that are less prominent in the current version of the product. Secondly, the volumes are discrete with values 0, 5, 10, 20, 40 or 60 ml rather than continuous. Thirdly, the positions are given as a number between 0 and 10, with 0 corresponding to healthy samples and the other 10 positions are shown in figure 3.3. The labelling of these were done by the company. Lastly, the number of samples is a couple thousand in total.



(a) As seen from below **(b)** As seen from the right

Figure 3.3: The positions in the phantom data (Image by Medfield Diagnostics AB)

3.1.3 Data set C

The last and smallest data set has similarities with both previous data sets. As in data set B, there exists some instrument noise. The noise variations in the data set mostly come from instruments in older revisions of the Strokefinder MD100. These noisier revisions were included to make more data samples available. There are instrument variations within each revision, but they are more prominent in earlier revisions and are not considered an issue for the current revision. The differences between revisions are mostly due to different materials being used for the instrument as well as an increased control over production in the newer revisions.
The volume and position representations are more similar to data set A. The volumes are once again represented as continuous numbers rather than categorical.The positions are either left or right if there exists a bleed and can be anywhere in that half of the head. Since there is also the possibility of the sample being healthy, a third category was introduced. An important note is that while the former two data sets uses the same samples for all tasks, this data set only has certain meta data for certain samples. To not restrict the number of samples even further, data with defined volumes may be included even if that particular sample is inconclusive

regarding position and vice versa. This leads to the data set having between 100 and 200 samples in total for each task.

3.2 Data Preprocessing

The original data sets contain S-parameters from a large range of frequencies, but to lower the amount of features sent to the network, only a fraction of these frequencies will be used. Another way the features was limited was by removing all reflection channels. These were chosen since the reflections are greatly affected by signals that never reaches the brain, and is instead reflected when hitting the scalp or similar.
Since the data is complex, the absolute value will be used instead. The input is also standardised by the following equation on all samples:

$$x_{scaled} = \frac{x - \bar{x}}{s}$$

with x representing one S-parameter, x_{scaled} being the scaled value of x, $\bar{x}$ is the mean of all x and s is the standard deviation of x over all samples.
In data set A, the samples were randomly split into a training, a validation and a test set. The split for the test set was the same for each run, leading to the network using the same samples for evaluation every run. The test set consists of 20% of the total number of samples, and the remaining 80% is split into training and validation with 5-fold cross validation as described in section 3.3.1.
In data set B, one instrument was chosen as a test instrument, meaning that only samples from this instrument was present in the test set. This was done to see whether of not the instrument variations affected the classifier.
In data set C, one instrument from each group was chosen for the test set. Similar to data set B, this was done to see if the instrument revision variations affected the classifier. The goal was to have around 20% of the samples in the test set. Since there were different numbers of samples from each instrument, which instrument from each revision was used in the test set was determined by the number of samples that specific instrument had compared to the total number of samples in that revision.

3.3 Training

The training is done for up to 5000 epochs. If overfitting occurs however, the network weights will be restored to the values they had at the best epoch and the training will end.

3.3.1 K-Fold Cross-Validation

When a data set is small, the performance of a neural network can depend largely on the samples it is trained on. This means that the way the data set is split into training, validation and test sets may affect the outcome. K-Fold Cross Validation is a way to account for this [25]. This is done by first splitting the data set into K groups. The network is then initialised, trained and evaluated K times, each time

with a different group as test set and the rest as training set. This means that each sample is in the test set exactly once and in the training set every other run. The implementation used in this study however, uses the K-fold cross validation to split into training and validation sets rather than training and test sets.

3.3.2 Network architecture

Three main types of network structures were used. One simple dense neural network (DNN) with a main structure as shown in figure 3.4 was used for all data sets.

Figure 3.4: Base structure for the dense neural network. A blue rectangle represent a single input layer. An orange rectangle consists of dense layers, together with any batch normalisation or dropout layers, which forms a feature extractor. The same is true for a violet rectangle, except it outputs a classification or regression result rather than a feature array.

For data set B and C, a Multi-Source Adversarial Domain Adaptation (MSADA) model was introduced with base structure as in figure 3.5. Another way of imple-

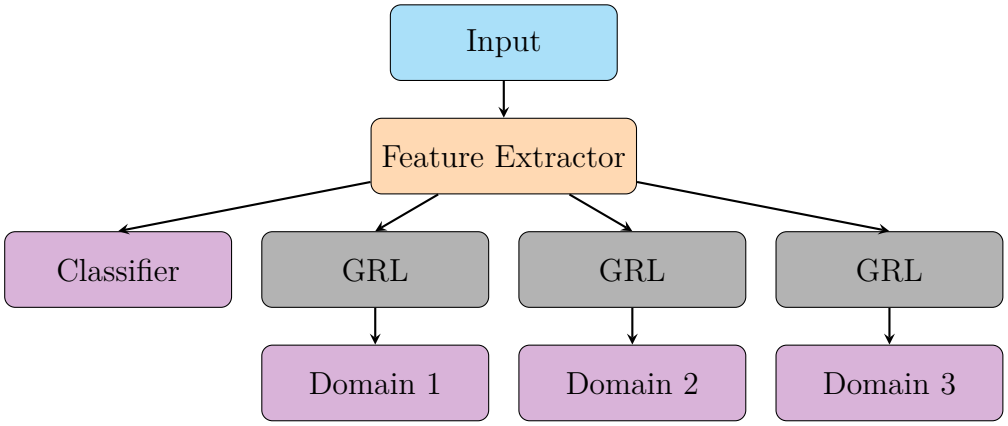


Figure 3.5: Base structure for the MSADA network. The same colour representations as in figure 3.4 were used, as well as introducing a grey rectangle to represent a gradient reversal layer.

menting adversarial domain adaptation is to simply have one task branch and one domain branch. The domain branch then classifies the domains as a regular categorical classifier. This was implemented and tested in this study as well as the structure shown in figure 3.5, but the former was concluded to be inferior and will therefore not be presented in this report.
A third network structure was introduced for data set C, which also used data from the other two data sets during training. The network is based on Divergence Based Domain Adaptation (DBDA) and its structure is shown in figure 3.6. The target domain is data set C, while source domain is either data set A or B for volume prediction or position classification respectively. Just as with MSADA, there are

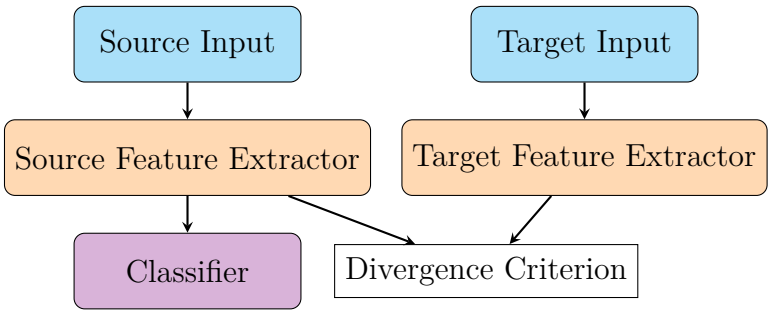


Figure 3.6: Base structure for the DBDA network. The same colour representations as in figure 3.4 were used, as well as introducing a white rectangle to represent a divergence loss.

multiple versions of DBDA. Two different structural implementations was done for this study. The one showed in figure 3.6, as well as one with a shared feature extractor. Since they had a similar performance, only the results of the former version will be presented in this report. Three different divergence criteria was implemented: Wasserstein Difference, KL Divergence and Correlation Alignment. Only one of these per task will be presented in this report. The KL Divergence was chosen for volume prediction, while the correlation alignment is used for position. These were chosen since they showed the best average accuracy of the three criteria for each respective task.

For volume prediction in data set A and C, the output should be a real number above zero. For this purpose, the ReLU function was used as activation of the last layer of the volume branch of the network. To evaluate these networks, the mean squared error was used. In data set B however, where the volume is a categorical value, the activation function used was the SoftMax function. This normalises all categories of the input, meaning that all categories are given a percentage representing how likely each category is, compared to the others. Here, the categorical accuracy was used. It takes the index of the largest value of the output and counts how often that matches with the correct index.

All positions are given as categorical values, but differs in how many categories they are. Neither SoftMax nor the categorical accuracy functions depend on the number of categories and these were therefore used for all data sets.

In all data sets the diagnostic classification benchmark is a binary problem of healthy vs stroke. Therefore, the sigmoid function was used as activation function in the final layer of the classifier. The binary accuracy was then used for evaluation. It rounds the output from the network to the nearest integer (0 or 1, here corresponding to healthy or stroke respectively) and then count the percentage of samples where the predicted value was rounded to the correct answer.

In the MSADA networks, the branches representing instruments or groups were always binary. Just as with the diagnostic benchmark models, this lead to the sigmoid function and the binary accuracy metric being used.

3.4 Evaluation

To evaluate a network, the metrics mentioned earlier was used for the different tasks. The stability of the networks was evaluated with the F_1-score. It uses the harmonic mean of the precision and recall to define the accuracy of a classification [26]. The precision is defined as

$$\text{precision} = \frac{\text{true positives}}{\text{true positives+false positives}} \tag{3.1}$$

and the recall as

$$\text{recall} = \frac{\text{true positives}}{\text{true positives+false negatives}}. \tag{3.2}$$

When using the F_1-score on a multi-class problem such as for the volumes in data set B and all positions, the score is also averaged. For this thesis, where the number of samples in each category is not necessarily equal, a weighted average was used. It computes the scores for each label and computes the weighted average, taking the label imbalance into account [27].

4
Results

In this chapter, all results from the different tasks on the different data sets are presented. All models were run with 5-fold cross validation, but for each fold the network was initialised, trained and evaluated 300 times. This means each model was evaluated a total of 1500 times, which was done to give a good representation of the stability of the networks. The stability is represented as the mean and standard deviation (Std) of the accuracy, F1-score, recall and precision over all runs.

4.1 Results from data set A

For data set A, a simple model (with a structure as in figure 3.4) was used to examine whether the information of position and volume exists in the simulated S-parameters.Since the volumes are real valued, the different metrics were computed by grouping all volumes representing a stroke together, which formed binary diagnostic results. The results of these metrics are shown in table 4.1. Note that the results shown in this table is identical for each metric. This is not generally true, and was a product of rounding the results to two decimals.

Table 4.1: Scores of different models on data set A. See equations (3.1) and (3.2) for precision and recall computation. See section 3.3.2 and 3.4 for accuracy and F1-score computation.

Model	Accuracy		F1-score		Recall		Precision	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std
Volume	0.95	0.02	0.95	0.02	0.95	0.02	0.95	0.01
Position	0.88	0.01	0.88	0.01	0.88	0.01	0.88	0.01

4.1.1 Volume prediction

The results for the volume prediction in data set A is shown in table 4.1 and a randomly chosen sample is shown in figure 4.1. In this figure the blue dots represent the predictions compared to the true values, while the black line shows where a perfect prediction would preside. This sample has a Mean Squared Error (MSE) of around 17, and in general samples had a MSE of between 16 and 20, with some outliers in both directions.

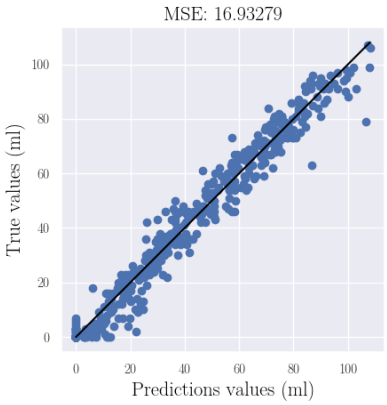


Figure 4.1: Volume prediction of data set A. Blue dots represent the coordinates as (x, y)=(predicted value, true value) for each input. The black line is the line x=y, which is the global optimum.

4.1.2 Position classification

The position classification results are shown in figure 4.2, where the labels are combinations of the letters describing each dimension: 'D' for 'Down', or 'U' or 'Up'. 'L' for 'Left', or 'R' for 'Right'. And 'B' for 'Back' or 'F' for 'Front'. The darker the colour in the figure, the more occurrences of that combination of predicted and true values was present. A perfect result would therefore be dark in the downward diagonal and light in the remaining squares. Every row is also normalised to sum to 1, meaning that 99% of the truly healthy samples were correctly predicted as healthy, while 0.75% were classified as the lower left back part of the brain.

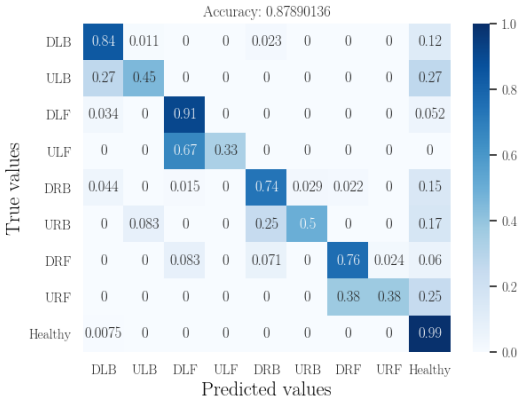


Figure 4.2: The position results of data set A. The darker the square, the higher the occurrence of that particular combination of predicted and true positions.

4.2 Results from data set B

For this data set, both a simple model with a structure as in figure 3.4, and a MSADA model with a structure as in figure 3.5 were implemented for each task.

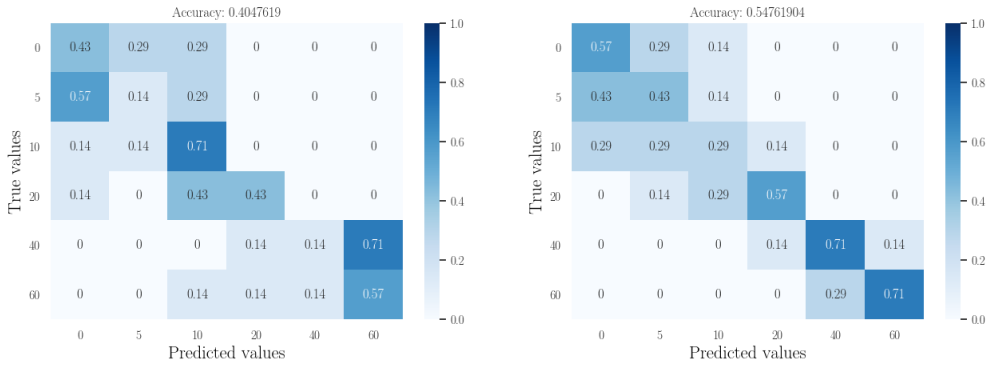
The second was introduced to deal with possible instrument specific variations which may cause the network to learn noise rather than information for the specific task. All gathered metrics for the models are shown in table 4.2.

Table 4.2: Scores of different models on data set B

Model	Accuracy		F1-score		Recall		Precision	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std
Volume	0.35	0.06	0.33	0.06	0.35	0.06	0.35	0.08
Volume (MSADA)	0.52	0.06	0.51	0.07	0.52	0.06	0.54	0.07
Position	0.46	0.05	0.47	0.05	0.46	0.05	0.57	0.06
Position (MSADA)	0.44	0.05	0.45	0.05	0.44	0.05	0.54	0.06

4.2.1 Volume prediction

In figure 4.3a the results from a simple dense network from a randomly selected run is shown. In an attempt to improve this by removing instrument specific noise, the MSADA network was implemented. Results from one run with this model are shown in figure 4.3b. The plot is computed the same as the plot for positions in data set A. The scores for these models are shown in table 4.2.

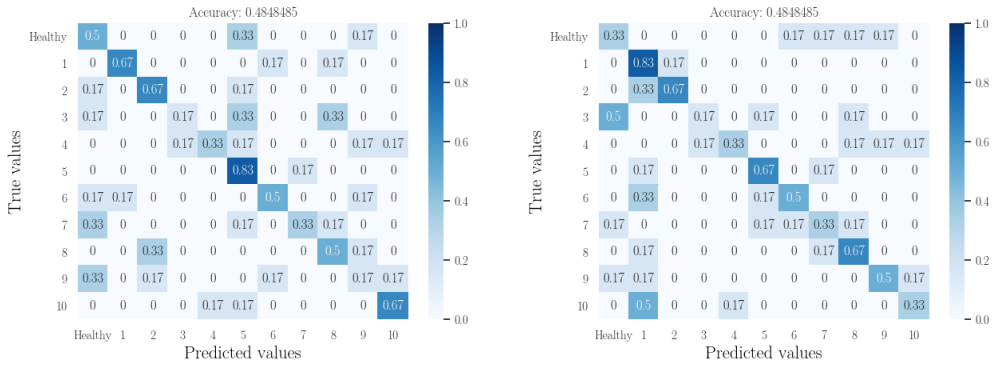


(a) DNN results **(b)** MSADA results

Figure 4.3: The volume results of data set B, with the accuracies of these specific runs.

4.2.2 Position classification

Just as with the volume prediction, both a simple network and a network with domain adaptation was implemented. The results of the two networks are shown in figure 4.4 with metrics shown in table 4.2. Important to note is that while a perfect result would mean that the downward diagonal was dark and the rest is light, a darker square close to the diagonal is not always better than one further away. Position 6, for example, is closer to position 3 than position 7 (see figure 3.3).



(a) DNN results **(b)** MSADA results

Figure 4.4: The position results of data set B

4.3 Results from data set C

Similarly to data set B, both a DNN and a MSADA model was implemented for all tasks for this data set. The difference is that the MSADA was used to confuse revisions of instruments, rather than specific instruments. A third network with structure as in 3.6 was also implemented for both tasks to attempt to account for the limitations of a small data set. The scores for all runs on this data set are shown in table 4.3.

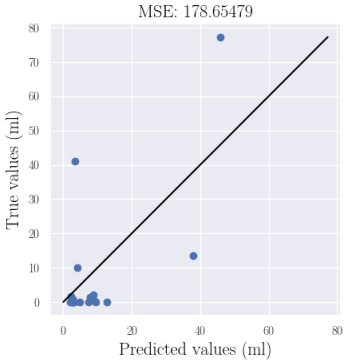
Table 4.3: Scores of different models on data set C

Model	Accuracy		F1-score		Recall		Precision	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std
Volume	0.60	0.12	0.52	0.19	0.61	0.12	0.56	0.26
Volume (MSADA)	0.60	0.12	0.52	0.19	0.60	0.12	0.56	0.26
Volume (DBDA)	0.53	0.12	0.43	0.16	0.53	0.12	0.42	0.22
Position	0.61	0.08	0.61	0.08	0.61	0.08	0.69	0.09
Position (MSADA)	0.60	0.08	0.60	0.08	060	0.08	0.67	0.08
Position (DBDA)	0.41	0.11	0.36	0.10	0.41	0.11	0.36	0.13

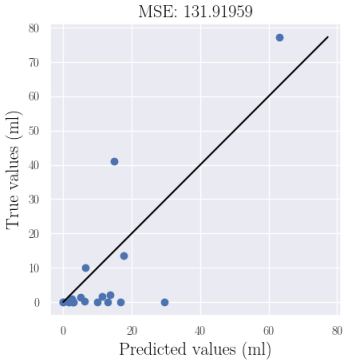
4.3.1 Volume prediction

The results of the volume prediction in data set C is shown in figure 4.5. The plots are made in the same way for both data set A and C, but due to the distribution of samples in this set, the full range of the volumes is not represented. The mean squared errors shown in these plots are for these specific runs, while the MSE could span between 100 and 400 for different runs. Excluding outliers in both directions. See table 4.3 for the accuracy metrics, computed as in data set A. Note that this means the accuracy and other metrics are only computed with the diagnostic classification in mind, meaning it grouped all volumes above 0 in one group and the rest

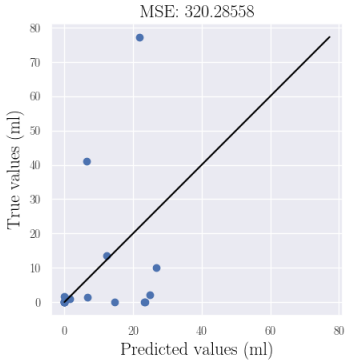
in a healthy group. The metrics are not representative of the mean squared error of the networks.



(a) DNN results



(b) MSADA results



(c) DBDA results (with KL Divergence as divergence loss)

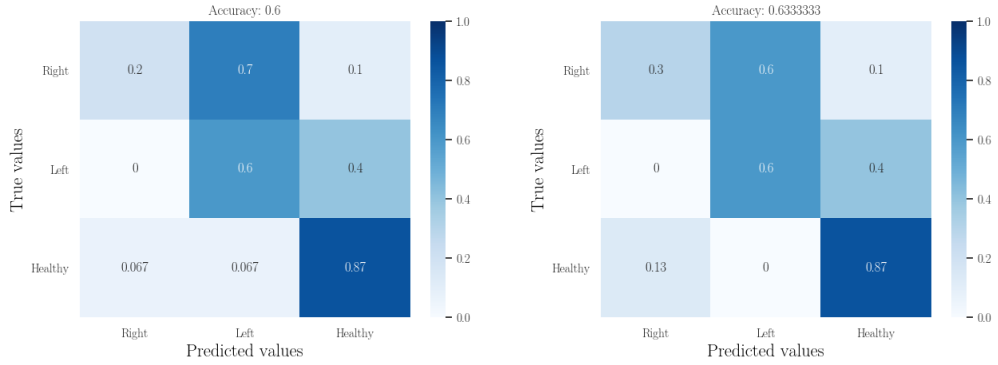
Figure 4.5: The volume results of data set C

4.3.2 Position classification

The results of a random run by each model is shown in figure 4.6. Similar to volume prediction for this data set, there are not many samples for each category, and each prediction affects the final score to a large degree.

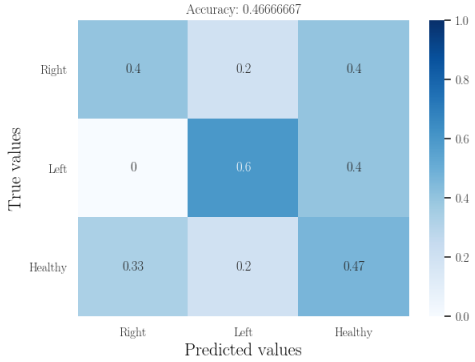
4.4 Diagnostic Accuracy

To be able to compare all results with a diagnostic classifier, the different positions and volumes that represented a stroke were grouped together to create a binary result for all data sets and networks. The results that the scores shown in table 4.4 are based on was collected in a similar way to previous scores, but with 200 runs rather than 1500. They are then compared to a diagnostic classifier with the same base structure. For example if a diagnostic classifier reaches an accuracy of 97% and a volume network only reaches 93% accuracy, the score would be $0.93/0.97 = 0.96$.



(a) DNN results

(b) MSADA results



(c) DBDA results (with correlation alignment as divergence loss)

Figure 4.6: The position results of data set C

Note that all results, including diagnostic accuracies, come from models created by the author and is not connected to the classifier used by the company.

Table 4.4: Diagnostic accuracies of different models on each data set compared to diagnostic classifier

Model	Data set A		Data set B		Data set C	
	Mean	Std	Mean	Std	Mean	Std
Volume	0.98	0.02	0.94	0.06	0.92	0.13
Volume (MSADA)	–	–	0.99	0.05	0.91	0.12
Volume (DBDA)	–	–	-	-	0.91	0.22
Position	0.97	0.008	0.96	0.06	1.20	0.08
Position (MSADA)	–	–	0.95	0.05	1.18	0.07
Position (DBDA)	–	–	-	-	0.84	0.13

5
Discussion

Here follows a discussion on each of the data sets as well as a comparison between them. Some ideas on how to proceed with the study, that was not done due to the time constraint of the thesis, are also discussed.

5.1 Data set A

5.1.1 Volume prediction

As seen in figure 4.1 the network is able to recognise some patterns in the data and learn some general behaviours of both larger and smaller bleeds. It does not appear to have a clear absolute difference in error for larger or smaller bleeds. However, estimating 4 ml too much on a 10 ml bleeding represents a 40% increase, while 4 ml too much on a 100 ml bleeding only a 4% increase. To improve this model, this could be taken into account during training by using a weighted error function or similar.

5.1.2 Position prediction

What we can see from the plot in figure 4.2 is that the network can relatively easy learn if the bleed is situated in the front versus back or in the left versus right areas of the brain. It is a bit more difficult to discern whether the bleed is in the lower or upper half of the brain. This could be because of the fact that the number of samples in the upper part is smaller than in the lower part. Moving the point where the XY-plane crosses the Z-axis to create an evenly distributed training could potentially fix this problem.

5.2 Data set B

5.2.1 Volume prediction

Based on the scores in table 4.2, it is safe to say that the MSADA network improves the performance of volume prediction for this data set. Based on the accuracy score of the randomly chosen plotted heat maps in figure 4.3, it is somewhat representative of the performance of the model. The network show a tendency to being able to discern small, medium and large bleeds. It could then be the issue of there being smaller steps in between the different volumes that the network finds harder to

separate when instrument variations are present. If the noise variations from the different instruments is of considerable size compared to these differences in the microwave data, it is reasonable that the network have difficulties in separating them.
From a physics point of view, simply looking at the data should show a change in the S-parameters from a channel, if the antennas are close to the bleed. How large this difference is should then indicate how large the bleed is. If a network is trained on data with one noise level, it may learn to compensate for this. When testing on data with a different noise level, the difference in S-parameters should change with the noise. This could explain why removing instrument variations improves performance, since it trains the network on what is noise and what is an actual input difference.

5.2.2 Position prediction

In contrast to the volume prediction, the position classification does not gain from reducing the effect of the instrument variations. Instead, it seems to lose some relevant information according to the scores in table 4.2. Using the physics argument here could explain this as well. Since a bleed is shown through some change in the S-parameters, locating this change is what the network is aimed to do. How large the difference is does not matter, only where it is presented in which channel. Removing the noise variations may simply regulate how large the difference in S-parameter is, not where it is located.
From the plots in figure 4.4, is looks as though some positions are easier to classify than others. These are close to both temples (position 1 and 2) and just above the ears on both sides (position 5 and 6). For the MSADA network, position 8 and 9 are also showing promising performance, which represents bleeds just behind the ears on both sides. What is common to all these positions are that they are those closest to the antennas. The remaining four positions (3, 4, 7 and 10) are all positioned higher up in the brain, with a further distance to any antenna. This could mean that bleeds closer to antennas are easier to find. A possible solution to improve performance on positions 3, 4, 7, and 10 could therefore be to add an antenna for the top of the head on the instrument. Another possible solution, that does not require a change in instrument, could be to introduce classification weights in the model to prioritise these positions during training. It is also possible that these difficulties could be due to the high number of categories for such a small data set. Lowering the number of categories by grouping them together in some way, could possibly improve the accuracy, but would decrease the information received from the network.

5.3 Data Set C

5.3.1 Volume prediction

What is clear in the first two plots in figure 4.5 (4.5a, 4.5b) is the non-uniformity of the data set as well as the limitation in the amount of samples. Another thing that can be seen by comparing the DNN results and the MSADA results is that the

former is better in predicting the smaller bleeds, while the latter is better with larger bleeds. There is also a marginally higher average recall for the DNN model (as seen in table 4.3), which could also point to the DNN being better with smaller volumes. It is possible that most small bleeds are from one and the same revision. When removing that revisions specific variations, it is possible that all smaller bleeds are then translated into larger bleeds. Removing this bias could be done by introducing smaller bleeds in the other groups as well, leading to the end goal of a more uniform data set. The limitations in the number of samples could also lead to the network learning to classify positions or something completely different. For example, if all small bleeds are situated in the right side of the brain, and all the larger ones are on the left. Then the network could associate bleeds on the left as large, and when evaluated on a small bleed in the left half of the brain, predict a large bleed.

The introduction of the DBDA model was an attempt to deal with the limited available data, as well as its non-uniformity. As seen in figure 4.5c, table 4.3 and table 4.4 this was not successful. There are many reasons why this may have failed. One reason could be that the mapping of values are incorrectly done during training. Instead of mapping a small bleed to a small bleed, it may have been mapped to a medium or even a large. It is also possible that even when successfully mapping volumes together, they may be located at very different parts of the brain. This (as suggested by results in the position tasks) leads to noticeable differences in the input data and may therefore contribute to the incorrect mapping.

5.3.2 Position prediction

Just like in data set B, the position classification shows a decrease in performance when implementing domain adaptation, as seen in table 4.3. This could imply that the variations between different instrument revisions are not what causes confusion in the classifier. The limited data is also a problem when evaluating this model. Since the number of samples are so small, each prediction makes a large impact on the accuracy compared to if the data set was larger. The small data set could also lead to all samples occurring in the test set having similar volumes, which in turn could mean that when introduced to different volumes, it has a completely different performance score. This all leads to the authors hesitation in drawing any conclusions other than the need for more qualitative and quantitative data.

The DBDA model was implemented in an attempt to handle the lack of quantitative data. This was not successful however, and the results shown in figure 4.6c, table 4.3 and table 4.4 actually show a decrease in performance. As with volume prediction with this method, there is a possibility that the implementation is at fault concerning which samples from source and target domains are being mapped together during training. Since samples from data set C are only sorted into left ant right rather than something more specific, it is highly likely that a position in the front is compared to one in the back and vice versa.

5.4 Comparison between different sets

The differences in how the data is represented in the different sets limits the comparisons to some extent. For example since the positions are defined different in each data set, conclusions on whether some positions are more or less difficult to predict can not be generalised from one set to the other. The fact that the volumes in data set B is a classification task instead of a regression task in combination with the fact that the largest bleeds in this set is 60 ml, while both the other sets reaches above 100 ml is also a possible complication. The differences between the different data sets could be the reason as to why the DBDA network is not working. For example both the number of antennas as well as their positions differs between the two data sets used for volume prediction in the DBDA network.

5.5 Comparison with diagnostic benchmark

For almost all networks, the diagnostic classifier performs better.The only networks that outperforms their diagnostic equivalencies are the DNN and MSADA position classifiers in data set C. Since it is data set C, it is possible that the increase in performance is simply because the samples used in the position task are all large or for other reasons easy to classify. Given the nature of the scores from the position classifiers on the other two data sets, it is probable that this is the case. For an improved comparison, data set C needs an increase in samples that are more uniform in volume and position, such that the same samples is used for all tasks.

5.6 Future work

There are many ways to move forward with this study, and some have already been discussed. For most further examinations however, more qualitative and quantitative data would be needed. This is especially important for data set C, since this is the smallest set of the three as well as it being the only non-uniformly distributed set. One way of introducing more data could for example be to generate samples as suggested in [18], or try some other type of transfer learning.
It could be interesting to more thoroughly examine how a volume predictor or position network can compare to a diagnostic classifier in performance. To do this, a decision tree or a joined network of volume, position and diagnostic outputs could be examined.
There may be correlations in which samples are wrongly classified in the different networks, such that those missed by the volume network are also missed by the position classifier. This would be interesting to study, and examine whether there are certain combinations of volumes and positions that are more or less prone to misclassification.

6
Conclusion

In conclusion, the noise free microwave data in data set A contains enough information to not only detect a bleed, but also locate its position and volume to a satisfying degree. When introducing instrument variations in the other two sets, the instruments play some part in what is possible to separate and classify correctly. Using multi-source adversarial domain adaptation can increase performance on volume prediction in at least data set B. While instrument variations are part of what disturbs the performance of position classification, it is not enough to use domain adaptation on these variations to improve position classification.
The only networks that was able to outperform the diagnostic classifier was two of the position classifiers in data set C. This was however believed to be a consequence of the difference in which samples were used for the different tasks, rather than the position classifier being superior.
The largest problems for data set C seems to be the non-uniform distribution of the data samples together with the limited number of samples available. Using generated data, improving the current domain adaptation method or changing domain adaptation method is suggested as ways to handle this in future studies.

Bibliography

[1] M. D. AB, "Strokefinder MD100."

[2] R. C. M. et al., "Systematic outperformance of 112 dermatologists in multiclass skin cancer image classification by convolutional neural networks," *European Journal of Cancer*, vol. 119, pp. 57–65, 2019.

[3] Folkhälsomyndigheten, "Insjuknande i stroke," 2021.

[4] F. Hedlund, "Stroke – en kamp mot klockan," *Medicinsk Vetenskap*, vol. 1, 2013.

[5] M. von Euler, "Stroke - symptom och riskfaktorer," 2021.

[6] R. Kothari, W. Barsan, T. Brott, J. Broderick, and S. Ashbrock, "Frequency and accuracy of prehospital diagnosis of acute stroke," *Stroke*, vol. 26, no. 6, pp. 937–941, 1995.

[7] "Diagnosis stroke."

[8] "NIH strokeskala."

[9] B. Billot, D. Greve, K. V. Leemput, B. Fischl, J. E. Iglesias, and A. V. Dalca, "A learning strategy for contrast-agnostic mri segmentation," 2021.

[10] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention u-net: Learning where to look for the pancreas," 2018.

[11] A. Al-Saffar, A. Bialkowski, M. Baktashmotlagh, A. Trakic, L. Guo, and A. Abbosh, "Closing the gap of simulation to reality in electromagnetic imaging of brain strokes via deep neural networks," *IEEE Transactions on Computational Imaging*, vol. 7, pp. 13–21, 2021.

[12] J. Bassey, L. Qian, and X. Li, "A survey of complex-valued neural networks," 2021.

[13] E. Tensor, "In ambulance brain imaging for triage of stroke patients."

[14] E. Vision, "Technology overview."

[15] G. Zhu, A. Bialkowski, L. Guo, B. Mohammed, and A. Abbosh, "Stroke classification in simulated electromagnetic imaging using graph approaches," *IEEE Journal of Electromagnetics, RF and Microwaves in Medicine and Biology*, vol. 5, no. 1, pp. 46–53, 2021.

[16] A. S. M. Alqadami, A. Trakic, A. E. Stancombe, B. Mohammed, K. Bialkowski, and A. Abbosh, "Flexible electromagnetic cap for head imaging," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 14, no. 5, pp. 1097–1107, 2020.

[17] A. Brankovic, A. Zamani, A. Trakic, K. Bialkowski, B. Mohammed, D. Cook, J. Walsham, and A. M. Abbosh, "Unsupervised algorithm for brain anomalies

localization in electromagnetic imaging," *IEEE Transactions on Computational Imaging*, vol. 6, pp. 1595–1606, 2020.

[18] R. S. Ebba Ekblom, "Generative adversarial network for generation of atificial microwave data for stroke detection," 2021.

[19] J. Brownlee, "How to improve performance with transfer learning for deep learning neural networks," 2020.

[20] A. Farahani, S. Voghoei, K. Rasheed, and H. R. Arabnia, "A brief review of domain adaptation," 2020.

[21] B. Sun and K. Saenko, "Deep coral: Correlation alignment for deep domain adaptation," 2016.

[22] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," 2015.

[23] H. Zhao, S. Zhang, G. Wu, J. M. F. Moura, J. P. Costeira, and G. J. Gordon, "Adversarial multiple source domain adaptation," in *Advances in Neural Information Processing Systems* (S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, eds.), vol. 31, Curran Associates, Inc., 2018.

[24] D. M. Pozar, *Microwave engineering; 4th ed.* Hoboken, NJ: Wiley, 2005.

[25] J. Brownlee, "A gentle introduction to k-fold cross-validation," 2018.

[26] J. Korstanje, "The f1 score," 2021.

[27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

DEPARTMENT OF PHYSICS
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden
www.chalmers.se

CHALMERS
UNIVERSITY OF TECHNOLOGY