Adapting 12-Lead ECG Deep Neural Networks For Screening Athlete Cardiac Health

Shaun Kyle[1]

[1]**UNSW Graduate School of Biomedical Engineering**

Running title: Adapting DNN models to athletic ECG

Supervisor: Dr Reza Argha

ABSTRACT

Athlete hearts undergo physiological adaptations in response to prolonged and intense exercise. An athlete's ECG recording presents differently to a patient from the general population, which leads to a higher likelihood of misdiagnosis from both human clinician and ECG classifier models. Due to a lack of available data, new athlete ECG classifiers based on deep neural networks (DNN) cannot be trained. The purpose of this study is to explore whether domain adaptation can be used to adapt existing DNN models trained on general population ECG to athletic ECG, while using only a minimal quantity of athletic ECG data.

The causes of this domain shift were found to be decomposed into two main components: concept shift and covariate shift. Domain adaptation methods for an existing DNN were devised based on source dataset reweighting and target dataset finetuning respectively. To evaluate the effectiveness of these methods, the original and adapted models were benchmarked across four datasets. Neither of the domain adaptation methods were found to address concept shift in the target domain. However, non-uniform distribution of patient demographics in the pretrained model's training data was found to negatively affect classification performance in both source and target domains, and could be alleviated with the adaptation methods used.

All code associated with analysis, training, and benchmarking can be found at:
`https://github.com/ShaunKyle/MisdiagnosisOfAthleteECG`

Keywords: Electrocardiogram, Deep Learning, Domain Shift, Classification, Training, Covariate Shift

CONTENTS

STATEMENT OF CONTRIBUTION

The datasets used in this study were collected by various research groups and organisations across Europe, Asia and North America. Their work is cited in this report.

The pre-trained ECG classifier model used in this study was originally developed by researchers associated with Seoul National University. Their work is cited in this report. The version of their code and model weights used in this project was forked and is available at:

`https://github.com/ShaunKyle/PhysioNet-Challange-2020`

All other work done in this study is my own. All code associated with analysis, training, and benchmarking can be found at:

`https://github.com/ShaunKyle/MisdiagnosisOfAthleteECG`

ACKNOWLEDGEMENTS

Thanks to my supervisor, Reza, for the helpful weekly discussions and giving me plenty of leeway to propose my own independent research directions.

1 INTRODUCTION

1.1 Challenges in screening athlete cardiac health

Athletes hearts undergo physiological adaptations (e.g. increased wall thickness and cardiac dimensions) due to prolonged and intense exercise for at least 4 to 8 hours per week [2]. This leads to a markedly different risk profile for cardiovascular diseases compared to the general population. A motivating example is the prevalence of Sudden Cardiac Death (SCD) as the leading cause of mortality for young athletes under 35 years of age, with an estimated incidence of around 0.6 to 3.6 on 100,000 persons per year. This has led many sporting organizations to consider implementing screening programs for athlete cardiac health. [1]

The first line of screening for cardiac health typically includes interpretation of a 12-lead resting electrocardiogram (ECG) recording for any observable abnormalities from heart electrical activity [11]. ECG findings arising from physiological adaptations of athletes' hearts in response to exercise can be mistaken for pathological changes due to cardiovascular disease. To address this issue, standard criteria for ECG interpretation have been developed and endorsed by many international medical societies [2]. A summary of these standards is shown in Figure 1.

However, there is still a shortage of physician expertise in sports cardiology, which places a limit on the effective reach of athlete cardiac screening programs.

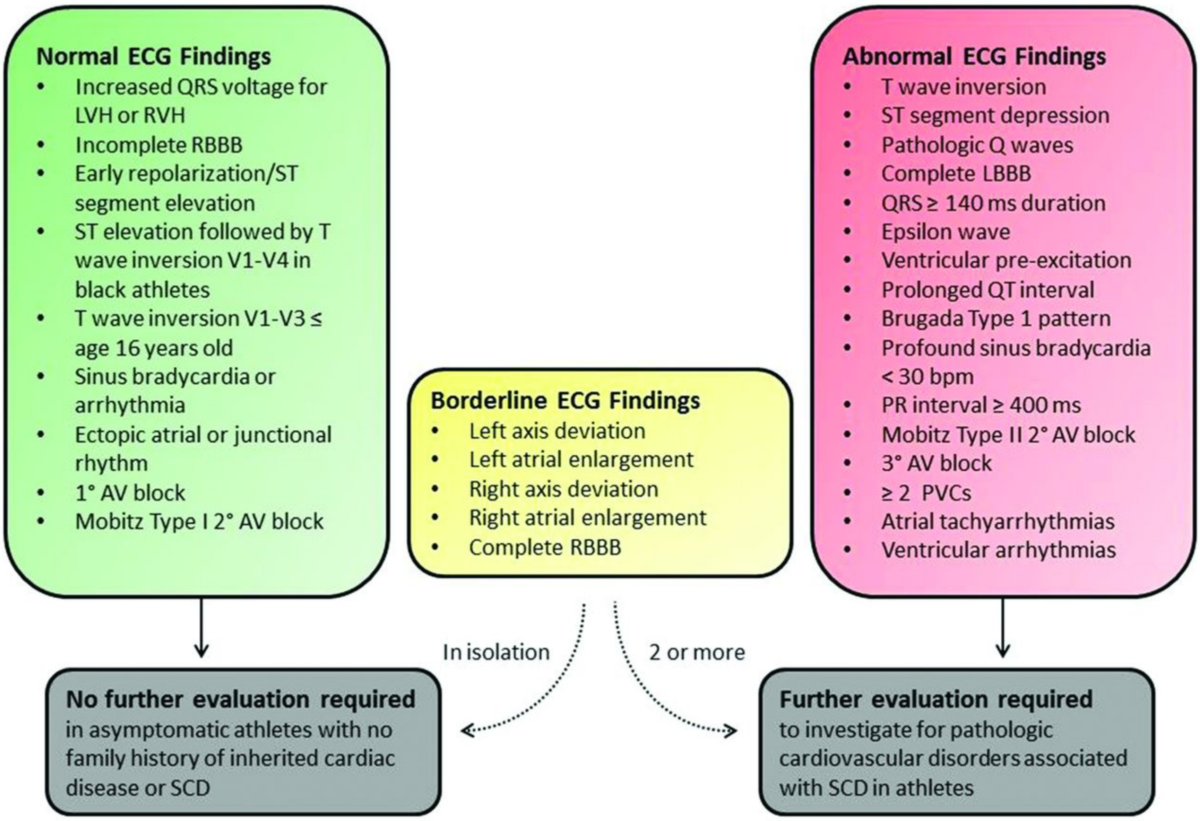
Normal ECG Findings
• Increased QRS voltage for LVH or RVH
• Incomplete RBBB
• Early repolarization/ST segment elevation
• ST elevation followed by T wave inversion V1-V4 in black athletes
• T wave inversion V1-V3 ≤ age 16 years old
• Sinus bradycardia or arrhythmia
• Ectopic atrial or junctional rhythm
• 1° AV block
• Mobitz Type I 2° AV block
Borderline ECG Findings
• Left axis deviation
• Left atrial enlargement
• Right axis deviation
• Right atrial enlargement
• Complete RBBB
Abnormal ECG Findings
• T wave inversion
• ST segment depression
• Pathologic Q waves
• Complete LBBB
• QRS ≥ 140 ms duration
• Epsilon wave
• Ventricular pre-excitation
• Prolonged QT interval
• Brugada Type 1 pattern
• Profound sinus bradycardia < 30 bpm
• PR interval ≥ 400 ms
• Mobitz Type II 2° AV block
• 3° AV block
• ≥ 2 PVCs
• Atrial tachyarrhythmias
• Ventricular arrhythmias
In isolation
2 or more
No further evaluation required in asymptomatic athletes with no family history of inherited cardiac disease or SCD
Further evaluation required to investigate for pathologic cardiovascular disorders associated with SCD in athletes

Figure 1. Overlap between pathological and physiological adaptations leads to borderline findings in athletes. **Normal** findings are unrelated to pathological changes that could suggest cardiovascular disease. **Abnormal** findings are unrelated to physiological changes from athlete training, but could suggest cardiovascular disease. **Borderline** findings could be caused by either pathological changes or athlete physiological adaptations. [2]

1.2 ECG classifiers based on Deep Neural Networks

Classifier models for automated ECG analysis are commonly used in conjunction with digital ECG machines to provide clinical decision support. In recent years, Machine

Learning (ML) and Deep Learning (DL) methods have been used to develop diagnostic models for a range of cardiovascular diseases in the general population. The reported accuracy of these models is usually quite high, typically above 90%. [11]

A key advantage of DL over traditional ML methods is that features can be learned implicitly, rather than being manually extracted by a human expert [3]. This allows deep learning models to be ”trained” using large datasets that would be impractical for a human to manually analyze, thereby giving the final model a far greater amount of experience to learn from. For example, supervised learning methods used to train ECG classifiers use large datasets consisting of raw ECG recordings as inputs and expert cardiologist ECG finding reports as labels. While human effort is required to label these datasets, the ECG features corresponding to particular diagnoses are learned implicitly by the model through training rather than human analysis.

An existing ECG classifier trained on a general population source domain is likely to have high misdiagnosis rates when used on an athletic population target domain for the reasons discussed in Section 1.1. This problem of a source domain differing from a target domain is referred to as a **domain shift** in the ML literature.

Developing a new athletic ECG classifier from scratch is not practical. The supervised DL methods used to develop Deep Neural Network (DNN) models for existing ECG classifiers require a high volume of training data, usually over 10,000 records. Under 200 records of athletes were found during a search of publicly available 12-lead resting ECG datasets, as shown in Table 1.

Table 1. Size of publicly available 12-lead resting ECG recording datasets

Dataset	Domain	No. of records	Unique subjects
MIMIC-IV-ECG [4]	General	800,000	160,000
PTB-XL [18]	General	21,799	18,869
Georgia [5]	General	20,678	15,742
CPSC [7]	General	9,831	9,458
Norwegian-Endurance [15]	Athletic (endurance sports)	28	28
PF12RED [10]	Athletic (football)	163	54

1.3 Domain adaptation

Given that the available athletic ECG data is insufficient to train a new model using supervised DL methods, a transfer learning approach which can adapt a general population trained DNN may be appropriate. **Domain adaptation** describes a specific case of transfer learning where the task (e.g. ECG classification) remains unchanged, but the distribution of source and target domains differ [6].

A list of definitions and notation used throughout this report are given below:

- A **domain** is a combination of an input space X, output space Y, and an associated probability distribution P.
- In this project, inputs are feature vectors consisting of both a 12-lead resting ECG recording and associated patient demographic information (e.g. age and sex), and outputs are binary class labels for a list of ECG findings. A dataset can be considered a sampled subset of a particular domain.
- For a domain adaptation problem, X and Y stay constant between the source and target domains. Only the source and target distributions P^S and P^T may vary.
- $P(X)$ denotes a marginal distribution. $P(X,Y)$ denotes a joint distribution. $P(X|Y)$ denotes a joint distribution (probability of X given Y).

- A **domain shift** occurs when $P^S(X,Y) \neq P^T(X,Y)$.

The underlying cause of a domain shift can be traced to one or more individual **distribution shifts**. A joint probability distribution can be decomposed into such, as shown by Equation 1.

$$\begin{aligned} P(X,Y) &= P(X) \cdot P(Y|X) \\ &= P(Y) \cdot P(X|Y) \end{aligned} \tag{1}$$

There are three types of distribution shift which can be identified:

1. **Covariate shift** occurs when the input distribution $P(X)$ changes. For example, the distribution of patient demographics such as age or sex.
2. **Label shift** occurs when the output distribution $P(Y)$ changes. For example, the distribution of ECG diagnosis labels in a dataset.
3. **Concept shift** occurs when a conditional probability changes. For example, $P(Y|X)$ changing would mean that for the same input distribution $P(X)$, a different output distribution $P(Y)$ is expected.

The changed criteria for diagnosis of athletes is clearly an example of concept shift between a general population source domain and athletic target domain. There may also be marginal distribution shifts which contribute significantly to the domain shift.

1.4 Project outline

The research question being addressed in this report is:

Can an existing ECG classifier which has been trained on a general population source domain be adapted for an athletic population target domain?

To address this research question, the aims of this project are as follows:

1. Identify significant distribution shifts between the source and target domains.
2. Devise suitable domain adaptation methods that could be used to minimize the effect of these distribution shifts.
3. Apply these methods to an existing ECG classifier which was trained on a general population source domain.

Aim 3 will involve benchmarking the performance of domain adapted versions of the classifier on both general and athletic populations. This will allow two key hypotheses to be tested:

- Addressing a marginal distribution shift (identified through Aim 1) will slightly increase classifier performance on the target domain.
- Addressing a joint distribution shift (i.e. concept shift) will significantly increase classifier performance on the target domain.

Section 2 will cover methods used in this study, including the measurement of marginal distribution shifts and design of domain adaptation methods to be used on an existing ECG classifier model. Section 3 will present the results of both the distribution shift measurements and performance of the domain adapted classifier. Section 4 will discuss how the study results can be interpreted as well as comment on future research directions.

2 METHODS

2.1 Selection of existing ECG classifier

The diagnoses covered by this study, shown in Table 2, fall under two main categories: sinus rhythms (SB, NSR, ST, SA) and conduction block (IRBBB, RBBB). These ECG diagnosis labels were chosen to cover two ”normal” athlete ECG findings that are likely to be misdiagnosed as false-positive, SB and IRBBB, and other diagnostic labels that fall under the same category.

Table 2. ECG diagnosis labels used in this study

SNOMED-CT code	Diagnosis
426177001	Sinus bradycardia (SB)
426783006	Normal sinus rhythm (NSR)
427084000	Sinus tachycardia (ST)
427393009	Other sinus arrhythmia (SA)
713426002	Incomplete right bundle branch block (IRBBB)
713427006	Complete right bundle branch block (RBBB)

The criteria used to select the DNN used in this study were:

1. Model is a multi-class classifier which includes ECG findings listed in Table 2.
2. Model input is a 12-lead resting ECG recording.
3. Training datasets are open and available.
4. Training datasets contain patient demographic information (e.g. age, sex) and diagnosis labels which can be used for comparison of source/target domains.

The chosen model is an ensemble of ten wide residual networks (WRN). The final output of the model is obtained by averaging the scalar prediction scores of the networks, then computing a binary output using class-specific threshold values. The architecture of each network, shown in Figure 2, consists of three main sections: a convolutional input stem, a set of four residiual blocks with a widening factor of 2, and a linear output stem for the final classifier output. [9]

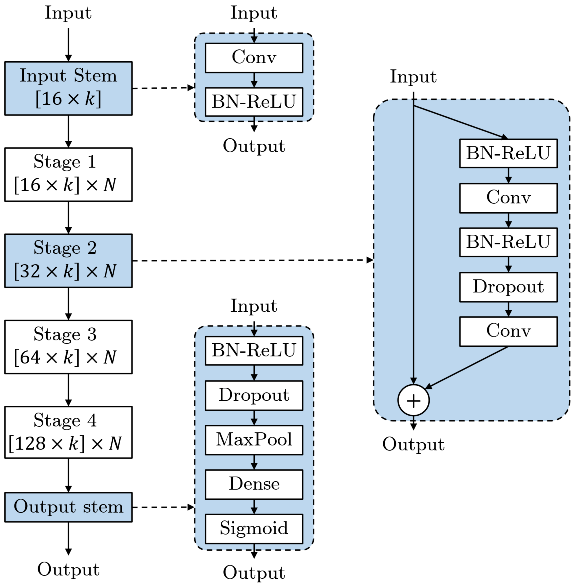

Figure 2. Wide residual network (WRN) architecture used by chosen ECG classifier [9]. Details of specific blocks shown in dashed lines. Widening factor $k = 2$ (i.e. double the number of convolution kernels).

The model training data comes from 4 of the 6 datasets provided for the 2020 George B. Moody PhysioNet Challenge [13], which are listed in Table 3. These datasets can be considered to be from the general population domain, with a slight demographic bias towards European patients. The 12-lead resting ECG recordings in the subset of four datasets used to train the model in [9] have an average recording length of 10 s, and were all resampled to a sampling rate of 500 Hz. Electrical noise sources were attenuated using 50 Hz and 60 Hz notch filters.

Table 3. Summary of 2020 PhysioNet Challenge training datasets [13]

Dataset	Records	Used for training [9]?	Clinical setting
Georgia	10,344	Yes	Emory University (Atlanta, Georgia, USA)
CPSC	6,877	Yes	11 hospitals across China
CPSC-Extra	3,453	Yes	Same as CPSC
PTB-XL	21,837	Yes	Collected by Schiller AG (European company)
PTB	516	No	University Clinic Benjamin Franklin (Berlin, Germany)
INCART	74	No	Institute of Cardiological Technics (St. Petersburg, Russia)

Each of the ten WRNs in the ensemble model was trained on a different set of training data using a 10-fold cross-validation method. Training was conducted using the PyTorch framework [12] on an NVIDIA V100 GPU. A summary of the training settings used is given in Table 4. There are 3,370,968 trainable parameters in each WRN, meaning the full model has over 30 million parameters.

Table 4. Summary of settings used for training model in [9]

Setting	Value
Epochs	100
Loss function	Binary cross-entropy
Optimizer	Adam
Learning rate	0.001
Dropout probability	0.3

2.2 Identifying significant distribution shifts

2.2.1 Target domain datasets

Two datasets belonging to the athletic target domain are listed in Table 5. The cohort for Norwegian-Endurance includes 28 elite athletes across endurance sports such as rowing, kayaking, and cycling [15]. The cohort for PF12RED includes pre-season and post-season ECG recordings for 54 professional football players [10]. Both datasets contain diagnosis labels for sinus rhythms and right bundle branch block.

Table 5. Summary of athletic (target) domain datasets

Dataset	Records	Female	Athletic setting
Norwegian-Endurance	28	32%	Endurance athletes in Norway
PF12RED	163	0%	Football team in La Liga, Spain

To minimize the effect of distribution shift due to sex demographics, the Norwegian-Endurance dataset will be used for any domain-adaptation methods requiring target domain data, while PF12RED will be reserved for evaluation purposes only.

Additionally, each ECG record in the Norwegian-Endurance dataset was interpreted by two different sources:

1. A cardiologist with specialization in athletes' hearts (assumed source-of-truth).
2. The Marquette 12SL algorithm on a MAC VUE 360 Electrocardiograph from GE Healthcare.

By comparing the ECG findings of the Marquette 12SL algorithm compared to the cardiologist, the misdiagnosis rates of the algorithm can be estimated, as shown in Table 6. The results presented in this table validate the ECG diagnosis labels chosen in Section 2.1.

Table 6. Misdiagnosis rates for Marquette 12SL algorithm on endurance athletes

Finding	False-positives (%)	False-negatives (%)
Sinus bradycardia	61	0
Other sinus arrhythmia	10	0
Right bundle branch block	2	50

2.2.2 Comparing marginal distribution shifts

In order to objectively compare distribution shifts between domains, a statistical measure of the "distance" between probability distributions is required. This section will introduce the concept of Kullback-Leibler Divergence based on writings from [8].

The Shannon information content of a single outcome x (measured in bits) is given by Equation 2. If the logarithm used is $\log_2$, the measurement is in bits. If the natural logarithm is used, the measurement is in nats.

$$h(x) = \log \frac{1}{P(x)} \tag{2}$$

For example, P could be the probability distribution of sinus rhythm findings in a domain, and $P(x)$ could be the probability of a particular finding such as a "normal sinus rhythm.

The entropy of an ensemble of outcomes X can be defined as the average Shannon information content of the outcomes:

$$H(X) = \sum_{x \in X} P(x) \log \frac{1}{P(x)} \tag{3}$$

For example, X could be the ensemble of all ECG findings related to sinus rhythm (bradycardia, normal, tachycardia, other arrhythmia). The higher the entropy, the more uncertain the outcomes sampled from a distribution.

The relative entropy between two distributions P and Q is known as the Kullback-Leibler Divergence (KL-Divergence). It can be defined using Equation 4, which is derived from the definition for entropy in Equation 3:

$$D_{\mathsf{KL}}(P||Q) = \sum_{x \in X} P(x) \log \frac{P(x)}{Q(x)} \tag{4}$$

If P and Q are continuous distributions (e.g. age of patients is a continuous variable), then the integral form of the equation is used [14]:

$$D_{\mathsf{KL}}(P||Q) = \int_{\infty}^{\infty} P(x) \log \frac{P(x)}{Q(x)} \tag{5}$$

KL-Divergence is not a perfect measure of difference between distributions. In general, $D_{\mathsf{KL}}(P||Q) \neq D_{\mathsf{KL}}(P||Q)$, so the measure is not strictly a ”distance”. However, it still suffices for the purpose of identifying distribution shifts between a source and target domain.

In this study, the *rel_entr* method (relative entropy) from the SciPy software library [17] is used to calculate KL-Divergences.

2.3 Domain adaptation methods

Based on the results of the marginal distribution shift analysis presented in Section 3.1 and the definition of concept shift presenting in Section 1.3, domain adaptation methods were devised to addressed two separate distribution shifts:

- Covariate shift due to differing age distribution between source and target datasets,
- Concept shift due to the changed criteria for diagnosing an athlete’s ECG [2].

As mentioned in Section 1, the quantity of target domain data available is not sufficient to train a new model from scratch using the training method proposed in [9]. A common paradigm in the transfer learning literature is **supervised finetuning**, where a model that has been pretrained on a source dataset has its weights adjusted through an iterative training training process which uses labels from target domain data to calculate training loss [16]. All domain adaptation methods devised in this study use a supervised finetuning process, detailed below:

1. Load the pretrained WRN weights.
2. For every record in the finetuning dataset: run a forward-pass through the WRN, calculate the binary cross-entropy (BCE) loss between the model predictions and finetuning dataset diagnostic labels, perform backpropogation on the loss function, perform gradient descent using a stochastic gradient optimizer to optimize the network weights.
3. Save the WRN weights and sum of BCE losses at the end of the training epoch.
4. Repeat from step 2 until the maximum number of training epochs has been reached.
5. Check which epoch had lowest BCE, use as the final weights for finetuned WRN.

To address different causes of domain shift, the finetuning dataset can be changed. A summary of the domain adaptation methods used in this study are listed in Table 7.

Covariate shift was addressed by training on a subset of patients under the age of 40 in the original training dataset (see Table 3). This could be more accurately termed as reweighting the source domain (shifting the age distribution younger) rather than finetuning, because no data from the target domain is used.

Concept shift was addressed by finetuning on the Norwegian-Endurance dataset. It was hypothesized that the ECG representation produced by the four stages of residual blocks of each WRN could be frozen, and that only the dense linear layer in the output stem which produces the classifier scores for each diagnosis class would need to be changed. This is tested by the *Finetune1* method.

Table 7. Summary of domain adaptation methods

Adaptation method	Max epochs	Learning rate	Finetune data	Trainable weights
Reweight	10	0.001	Source/training set, under age 40	Full model
Finetune1	100	0.0001	Norwegian-Endurance	Output stem
Finetune2	100	0.0001	Norwegian-Endurance	Full model

In addition to the three finetuned models in Table 7, another two models were produced. *Combine1* applies the *Finetune1* method to the weights of a *Reweight* model. *Combine2* applies the *Finetune2* method to the weights of a *Reweight* model. The goal of combining adaptation methods is to address both covariate and concept shift in the same model.

2.4 Benchmarking classifier performance

The classification performance of the original model from [9] and the adapted models described in Section 2.3 will be evaluated by running the model on ECG records from four evaluation datasets and comparing the predicted results of each model to the diagnosis labels from the datasets.

The datasets used to evaluate performance on a general population (source) domain are PTB and INCART, described in Table 3. Both datasets contain a mix of the sinus rhythm and bundle branch block diagnoses mentioned in Section 2.1.

The datasets used to evaluate performance on an athletic (target) domain are PF12RED and Norwegian-Endurance. While also being used for target finetuning, Norwegian-Endurance is still included as an evaluation dataset for the sake of evaluating both the *Original* and *Reweight* models. PF12RED data has not been used for any domain adaptation method, so will provide a fair comparison for all models.

F1-score will be used as the primary metric to quantify the overall binary classification performance of the classifiers. The calculation for F1-score is defined in Equation 6, where TP, TN, FP, and FN stand for True Positive, True Negative, False Positive and False Negative respectively.

$$F_1 = \frac{2 \cdot \mathsf{TP}}{2 \cdot \mathsf{TP} + \mathsf{FP} + \mathsf{FN}} \tag{6}$$

False-positive rate (FPR), shown in Equation 7 will also be used as a metric to provide an indication of misdiagnosis on the athletic datasets. Based on the criteria for interpretation of athlete ECG shown in Figure 1, the most likely misdiagnosis for SB and IRBBB findings is a false-positive. If the false-positive rate does not decrease for domain adapted models on the athletic datasets, then further analysis should be done into the cause of misdiagnosis.

$$\mathsf{FPR} = \frac{\mathsf{FP}}{\mathsf{FP} + \mathsf{TN}} \tag{7}$$

3 RESULTS

3.1 Identifying significant distribution shifts

The effect of marginal distribution shifts were measured using the KL-Divergence method discussed in Section 2.2. The calculation results are summarized in Table 8. The source domain are the datasets used to pretrain the original ECG classifier model, defined in Table 3. The target domain distributions are derived from the Norwegian-Endurance dataset which is used for domain adaptation in this study, with the exception of age distribution.

Label shift was measured between the prevalence of diagnosis labels (defined in Table 2). Covariate shift was measured between two patient demographics: age and sex. Ideally, all shifts would have been measured with respect to the Norwegian-Endurance dataset being used for domain adaptation. Unfortunately, individual athlete ages were not provided for this dataset. The distribution of patient ages in the PF12RED dataset is assumed to be similar to that of Norway-Endurance.

Table 8. Calculation results for KL-Divergences between source and target domains

Source distribution P	Target distribution Q	$D_{\mathrm{KL}}(P\|\|Q)$
Diagnosis labels in training data	Diagnosis labels in Norwegian-Endurance	3.607 nats
Patient ages in training data	Athlete ages in PF12RED[1]	29.367 nats
Patient sex in training data	Athlete sex in Norwegian-Endurance	0.047 nats

It can be concluded from these results that the most significant marginal distribution shift is due to patient age demographics.

3.2 ECG classifier performance

The benchmark results for the original classifier and the adapted models from Table 7 are summarised in Tables 9 and 10. The PTB and INCART evaluation datasets both belong to a general population domain, while the PF12RED and Norwegian-Endurance datasets both belong to an athletic domain. F1-score is used as a measure of overall classifier performance. False positive rate is used to measure misdiagnosis of borderline athlete ECG findings.

On general population datasets, the adapted models significantly outperform the original model. The best performance was achieved by the *Reweight* model. Derivative models *Combine1* and *Combine2* did not improve on the performance of the *Reweight* model.

On athletic population datasets, the adapted models also outperform the original model. All adapted models achieved the exact same performance on the PF12RED benchmark, which is unusual when compared to the other three benchmarks. This could be due to characteristics of the ECG recordings in the dataset, rather than a reflection of the true performance of the models. For example, a consistent lead misplacement when taking ECG recordings.

An interesting result to note is the performance of the *Reweight* model on the Norwegian-Endurance benchmark. *Reweight* outperformed models finetuned on records from the Norwegian-Endurance dataset. This allows us to conclude that *Reweight* is the best performing domain adaptation method on the athletic domain.

[1] Ages of individual athletes were not provided in the Norwegian-Endurance dataset. The distribution is assumed to be similar between PF12RED and Norwegian-Endurance.

Table 9. Classifier performance on general population domain

Evaluation dataset	Model	F1-score	False positive rate
PTB	Original	0.783	0.0562
	Reweight	0.998	0.0004
	Finetune1	0.887	0.0248
	Finetune2	0.888	0.0252
	Combine1	0.998	0.0004
	Combine2	0.998	0.0004
INCART	Original	0.649	0.0703
	Reweight	0.824	0.0351
	Finetune1	0.716	0.0568
	Finetune2	0.689	0.0622
	Combine1	0.824	0.0351
	Combine2	0.824	0.0351

Table 10. Classifier performance on athletic domain

Evaluation dataset	Model	F1-score	False positive rate
PF12RED	Original	0.361	0.1408
	Reweight	0.376	0.1238
	Finetune1	0.376	0.1238
	Finetune2	0.376	0.1238
	Combine1	0.376	0.1238
	Combine2	0.376	0.1238
Norwegian-Endurance	Original	0.369	0.1610
	Reweight	0.576	0.0803
	Finetune1[2]	0.563	0.1090
	Finetune2[2]	0.563	0.1090
	Combine1[2]	0.576	0.0803
	Combine2[2]	0.576	0.0803

Summary of observations:

- *Reweight* is the best performing domain adaptation method.
- Domain adaptation methods for the athletic domain produced a significant improvement in the general population domain.
- The *Combine1* and *Combine2* models never exceed the performance of the *Reweight* model.
- The *Finetune1* model usually matches or slightly exceeds the performance of the *Finetune2* model.

[2] Norwegian-Endurance dataset samples present in training data for Finetune1 and Finetune2 methods

4 DISCUSSION

4.1 Comparison of domain adaptation methods

The *Combine1* and *Combine2* models never outperformed the *Reweight* model that they were based on. This means that finetuning on the Norwegian-Endurance dataset addressed the same distribution shift as the *Reweight* method (i.e. covariate shift of patient ages) rather than concept shift. The results for *Combine1* and *Combine2* can therefore be ignored for this discussion.

On the athletic (target) domain evaluations, the average improvement in F1-score was $+0.1$ for all three domain adaptation methods (*Reweight*, *Finetune1* and *Finetune2*). This suggests that the covariate (age) domain shift between source and target domains was addressed to a similar degree by all three methods.

On the general population (source) domain evaluation, *Reweight* exhibited a higher average improvement in F1-score of $+0.2$ compared to $+0.1$ for both athlete finetuning methods. This observation is surprising, because no significant domain shift could be measured between the training datasets and PTB/INCART evaluation datasets.

4.2 Mechanism for classifier performance improvement

One possible explanation for the improvement on general population evaluations is attenuation of a bias introduced in the training of the original model [9]. The original patient age distribution of the training data, as shown in Figure 3, shows a negative-skewed bell curve with a peak probability density seen between the 60 to 80 age range. This means younger patients below this age range are underrepresented in the training data. In an ideal scenario, the distribution should be completely flat (i.e. equal representation of all age groups).

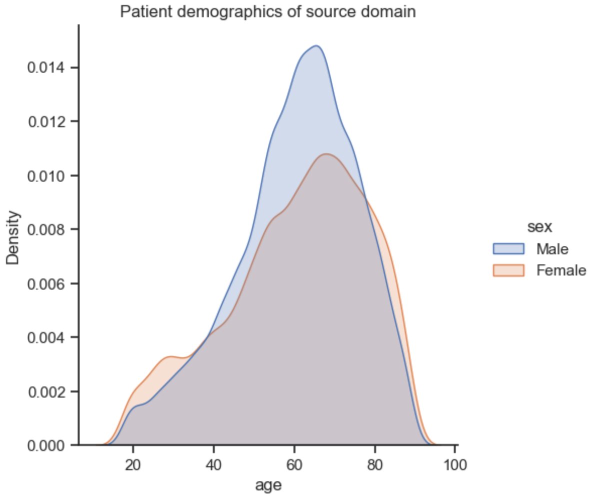

Figure 3. Visualization of age distribution across training datasets (Georgia, CPSC, CPSC-Extra, PTB-XL)

Both the source reweighting and target finetuning approaches effectively flatten the age distribution by presenting younger patients more often during the training/finetuning process. The higher degree of improvement for *Reweight* can be explained by the higher volume of training data available (5252 patients under age 40 for *Reweight* compared to 28 athletes with mean age 25 for target finetuning). The lack of additional improvement on the athletic domain evaluations is likely due to concept shift being the limiting factor for performance.

4.3 Limitations of study

The Norwegian-Endurance dataset was utilized for finetuning in the domain adaptation methods *Finetune1* and *Finetune2*. This calls into question the validity of the results in Table 10, where these target finetuned models are evaluated on the same Norwegian-Endurance dataset.

It can be argued that the observations made during this study are still valid. *Finetune1* and *Finetune2* did not outperform *Reweight* on the target domain evaluations, which would be the expected result for a validation dataset leaking into training data. The observation made by this study is that the target finetuning methods did not achieve their intended purpose, but rather improved the model using a similar mechanism to *Reweight*.

Before the results of this study can be presented to a wider audience, k-fold cross-validation should be introduced as a method for finetuning and evaluating the domain adapted models. This would involve splitting the target dataset into k equal folds/groups (e.g. 7 groups of 4 athletes each). One group can be set aside for validation, while the rest can be used for finetuning. To ensure that results are repeatable, the process can be repeated k times so that each fold is used for validation once.

4.4 Future directions

The main observation from this study is that the skewed patient age distribution used in the training data for the original model in [9] is not ideal, and that a flatter distribution would be better. One method of achieving this without needing to collect additional data or removing records from the dataset is to make use of data augmentation techniques such as synthetic ECG generation using Deep Generative Models trained on the source datasets or mathematical models of the cardiovascular system [19].

Concept shift was not successfully addressed by the target finetuning methods in this study. Methods for addressing concept shifts are still an active research topic in the field of domain adaptation. Additional domain adaptation methods that could be tried to extend this study include feature-based and inference-based approaches. Feature-based approaches aim to modify the feature space so that a classifier trained on the transformed source data can generalize well to the target data. Inference-based approaches integrate adaptation into the process of estimating model parameters. [6]

5 CONCLUSION

The main distribution shifts contributing to the domain shift between general population and athletic ECG were identified to be covariate shift (i.e. due to younger average age of athletes) and concept shift (i.e. due to change in criteria for ECG diagnosis of athletes). The domain adaptation methods devised for this study, source reweighting and target finetuning, successfully addressed covariate shift but failed to address concept shift. This placed a ceiling on performance improvement to $+0.1$ F1-score on the target athletic domain. An unexpected improvement in classifier performance on the general population (source) domain was discovered, and could be explained by the skewed distribution of patient ages present in the original training dataset.

REFERENCES

[1] Flavio D'Ascenzi et al. "Causes of sudden cardiac death in young athletes and non-athletes: systematic review and meta-analysis". In: *Trends in Cardiovascular Medicine* 32.5 (July 2022), pp. 299–308. ISSN: 1050-1738. DOI: 10.1016/j.tcm.2021.06.001.

[2] Jonathan A Drezner et al. "International criteria for electrocardiographic interpretation in athletes: Consensus statement". In: *British Journal of Sports Medicine* 51.9 (Mar. 2017), pp. 704–731. ISSN: 1473-0480. DOI: 10.1136/bjsports-2016-097331.

[3] Shenda Hong et al. "Opportunities and challenges of deep learning methods for electrocardiogram data: A systematic review". In: *Computers in Biology and Medicine* 122 (July 2020), p. 103801. ISSN: 0010-4825. DOI: 10.1016/j.compbiomed.2020.103801.

[4] Alistair E. W. Johnson et al. "MIMIC-IV, a freely accessible electronic health record dataset". In: *Scientific Data* 10.1 (Jan. 2023). ISSN: 2052-4463. DOI: 10.1038/s41597-022-01899-x.

[5] Kaggle. *Georgia 12-lead ECG Challenge (G12EC) Database*. 2024. URL: https://www.kaggle.com/datasets/bjoernjostein/georgia-12lead-ecg-challenge-database.

[6] Wouter M. Kouw and Marco Loog. "An introduction to domain adaptation and transfer learning". In: (Dec. 2018). DOI: 10.48550/ARXIV.1812.11806. arXiv: 1812.11806 [cs.LG].

[7] Feifei Liu et al. "An Open Access Database for Evaluating the Algorithms of Electrocardiogram Rhythm and Morphology Abnormality Detection". In: *Journal of Medical Imaging and Health Informatics* 8.7 (Sept. 2018), pp. 1368–1373. ISSN: 2156-7018. DOI: 10.1166/jmihi.2018.2442.

[8] David J. C. MacKay. *Information theory, inference, and learning algorithms*. 22nd printing. Cambridge [u.a.]: Cambridge University Press, 2019. 628 pp. ISBN: 9780521642989.

[9] Seonwoo Min et al. "Bag of Tricks for Electrocardiogram Classification with Deep Neural Networks". In: *2020 Computing in Cardiology Conference (CinC)*. CinC2020. Computing in Cardiology, Dec. 2020. DOI: 10.22489/cinc.2020.328.

[10] Adolfo Antonio Munoz-Macho, Manuel Jesus Dominguez-Morales, and Jose Luis Sevillano-Ramos. "An innovative 12-lead resting electrocardiogram dataset in professional football". In: *Data in Brief* 54 (June 2024), p. 110444. ISSN: 2352-3409. DOI: 10.1016/j.dib.2024.110444.

[11] Stefano Palermi et al. "Unlocking the potential of artificial intelligence in sports cardiology: does it have a role in evaluating athlete's heart?" In: *European Journal of Preventive Cardiology* 31.4 (Jan. 2024), pp. 470–482. ISSN: 2047-4881. DOI: 10.1093/eurjpc/zwae008.

[12] Adam Paszke et al. *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. 2019. DOI: 10.48550/ARXIV.1912.01703.

[13] Erick A Perez Alday et al. "Classification of 12-lead ECGs: the PhysioNet/Computing in Cardiology Challenge 2020". In: *Physiological Measurement* 41.12 (Dec. 2020), p. 124003. ISSN: 1361-6579. DOI: 10.1088/1361-6579/abc960.

[14] Fernando Perez-Cruz. "Kullback-Leibler divergence estimation of continuous distributions". In: *2008 IEEE International Symposium on Information Theory*. IEEE, July 2008, pp. 1666–1670. DOI: 10.1109/isit.2008.4595271.

[15] Bjørn-Jostein Singstad. *Norwegian Endurance Athlete ECG Database*. 2022. DOI: 10.13026/QPJF-GK87.

[16] Nima Tajbakhsh et al. "Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning?" In: (2017). DOI: 10.48550/ARXIV.1706.00712.

[17] Pauli Virtanen et al. "SciPy 1.0: fundamental algorithms for scientific computing in Python". In: *Nature Methods* 17.3 (Feb. 2020), pp. 261–272. ISSN: 1548-7105. DOI: 10.1038/s41592-019-0686-2.

[18] Patrick Wagner et al. "PTB-XL, a large publicly available electrocardiography dataset". In: *Scientific Data* 7.1 (May 2020). ISSN: 2052-4463. DOI: 10.1038/s41597-020-0495-6.

[19] Beatrice Zanchi et al. "Synthetic ECG signals generation: A scoping review". In: *Computers in Biology and Medicine* 184 (Jan. 2025), p. 109453. ISSN: 0010-4825. DOI: 10.1016/j.compbiomed.2024.109453.